On the Tightness and Computational Tractability of Higher-Dimensional Confidence Sequences

Fabian Denoodt[*]
Eindhoven University of Technology

Sibylle Hess
Eindhoven University of Technology

Joaquin Vanschoren
Eindhoven University of Technology

Christian A. Naesseth
University of Amsterdam

Abstract

Modern sequential monitoring problems often involve multiple metrics, where we monitor several data streams simultaneously and may act once the evidence is strong enough. Confidence sequences (CSs) are a natural tool for such continuous monitoring. However, for bounded *vector* means, existing multivariate CSs are either tight but computationally intractable, or fast to compute but conservative. To address this, we study three lifts of one-dimensional betting-based CSs to higher dimensions: a weighted Bonferroni region, an equivalent max-wealth form, and a portfolio region. The portfolio is typically much tighter, especially in higher dimensions, but its boundary and properties such as volume are not available in closed form. To make this tighter construction usable, we propose tractable outer approximations of the portfolio region that preserve statistical validity: a bounding box, an ℓ_p-ellipsoid, and their intersection. We prove set relations among all constructions and show empirically that these approximations (i) achieve regions close to the intractable portfolio, (ii) substantially outperform existing tractable multivariate CSs, and (iii) enable practical use cases such as multi-metric A/B testing and model comparison.

1 Introduction

Many modern decision problems are inherently sequential: we monitor the data as it arrives and want to make a decision as soon as the evidence is strong enough. In A/B testing [1, 2], for example, one may compare two versions of a website and stop collecting data once there is reliable evidence that one variant performs better. This kind of optional stopping requires statistical tools that remain valid regardless of when we stop. *Confidence sequences* (CSs) provide exactly this guarantee: a CS for a mean μ is a sequence of confidence sets that contains μ at *all* times t with a prescribed probability [3].

While much of the existing literature focuses on CSs for a *scalar* mean, many modern applications involve a *vector* mean $\boldsymbol{\mu} \in \mathbb{R}^D$. For instance, in website A/B testing, one may monitor several metrics simultaneously, such as browsing time, purchases, and page views. Similarly, when evaluating a new ML model's accuracy against an already deployed one, one may also monitor subgroup accuracies to detect possible subgroup degradations. To be useful in such settings, a multivariate CS should satisfy two requirements: it should be *tight*, meaning that its sets shrink rapidly as more data are observed (enabling early decisions), and *computationally tractable*, allowing for real-time monitoring in moderate to high dimensions.

[*]Research conducted during a research visit at the University of Amsterdam and completed at Eindhoven University of Technology.

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

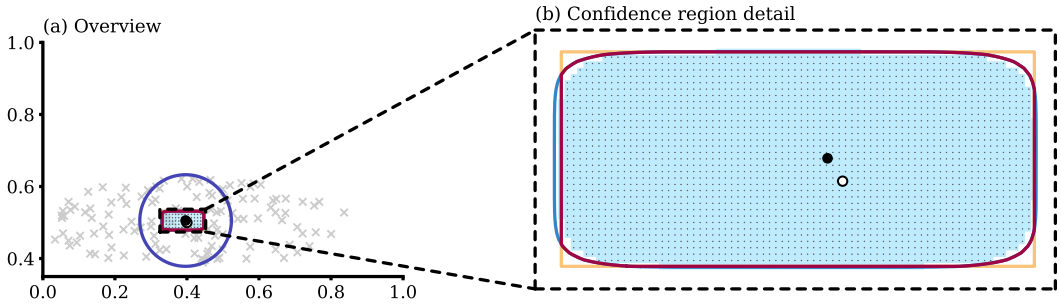


Figure 1: Intractable CS $\mathcal{C}^{\text{port}}$ (▪) and its computable superset approximations: the bounding box $\mathcal{C}^{\text{bbox}}$ (—), the ℓ_p-ellipsoid $\mathcal{C}^{\ell_p}$ (—), and their intersection $\mathcal{C}^{\text{bbox}\cap\ell_p}$ (—). The figure also shows observations (×), the population mean (○), the empirical mean (●), and a previously strong baseline [4] (—).

Achieving both requirements simultaneously is challenging. If one focuses on tightness, the multivariate CSs of Ryu and Wornell [3] provide a strong benchmark: they can be tight, but lack a closed-form solution for the region they define. They also have a computational cost of $O(t^D)$ for verifying whether a vector belongs to the CS at time t [4], limiting their practical use. On the other hand, closed-form regions such as confidence spheres [5, 4] are easy to compute, but they expand and shrink in all directions together, which is wasteful when some coordinates have low variance and others do not, as shown by the baseline in Fig. 1. Meanwhile, the ellipsoids by Whitehouse et al. [6] can adapt to the per-axis variance but target more general vector processes, whereas we specialize to bounded observations and exploit this structure for smaller regions.

Motivated by recent progress on *one-dimensional* CSs [7, 8], we lift the betting-based CS of Waudby-Smith and Ramdas [8] to bounded *vectors*. A simple approach is to build a 1D CS for each coordinate and adjust the significance level using a multiple-testing correction, resulting in a weighted Bonferroni hyperrectangle. This construction is easy to compute, but becomes increasingly conservative as D grows. We therefore also consider a tighter multivariate construction that aggregates the coordinate-wise wealth processes of [8] through a weighted average, which we call a portfolio. The portfolio CS can be substantially tighter, but is less practical since the boundary is not described explicitly, and properties such as volume or maximum width are not easily available in closed form.

This paper addresses that gap. **(i)** We treat the portfolio as a tight but expensive target and derive three *outer* approximations that scale linearly with the dimension: a bounding box, an ℓ_p-ellipsoid, and their intersection. **(ii)** We prove set relations among all constructions (see Fig. 2) and provide the necessary theory to ensure that our approximations remain statistically valid, even *after implementation*. **(iii)** Experiments show that our constructions **consistently outperform existing tractable CSs**, resulting in substantially smaller regions, particularly as D grows. Two case studies further illustrate their practical value by showing that they can detect statistical significance with substantially fewer observations[2].

Roadmap. Building on the univariate CSs described in § 2, § 3 formalizes three multivariate lifts (a weighted Bonferroni rectangle, an equivalent max-form, and a portfolio construction), proves their relations, and identifies the portfolio as the tightest but computationally intractable variant. §§ 4 and 5 then develop the three tractable outer approximations to this portfolio and show how to compute them in practice so their regions remain statistically valid.

2 Background: One-Dimensional Confidence Sequences by Betting

Preliminaries. We observe random variables $y_1, y_2, \cdots \in [0,1]$ with $\mathbb{E}[y_t|y_{1:t-1}] = \mu$ for all $t \geq 1$, where $\mu \in [0,1]$. This setup also includes the classical i.i.d. setting where $\mathbb{E}[y_t|y_{1:t-1}] = \mathbb{E}[y_t] = \mu$. A CS for the mean μ is a sequence of sets $\mathcal{C}_1, \mathcal{C}_2, \ldots$ (denoted compactly as $(\mathcal{C}_t)_{t\geq 1}$) satisfying

$$\mathbb{P}(\forall t \geq 1 : \ \mu \in \mathcal{C}_t) \geq 1 - \alpha,$$

[2]Code and a practical tutorial on using our CSs to monitor ML models are available here.

for a prescribed error level α [3]. We can also denote it as a $(1-\alpha)$-CS, to emphasize α. It can be constructed through a technique that relies on Ville's [1939] inequality for supermartingales. Let $\mathcal{F}_{t-1} := \sigma(Y_{1:t-1})$ denote the *sigma-algebra generated by the past observations* $y_1, \ldots, y_{t-1}$, representing all information available up to time $t-1$. A nonnegative process $(M_t)_{t=0}^{\infty}$ is a *supermartingale* if

$$\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] \leq M_{t-1} \tag{1}$$

for every $t \geq 1$. If $M_0 = 1$, it is a *test* supermartingale (or test martingale when equality holds in Eq. (1)). Ville's inequality states that for any test (super)martingale,

$$\mathbb{P}\big(\exists t : M_t \geq 1/\alpha\big) \;\leq\; \alpha.$$

Hence, if for candidate $m \in [0,1]$ we define $M_t(m)$ such that $(M_t(\mu))_{t\geq 0}$ is a test (super)martingale and $M_t(m)$ is $\mathcal{F}_t$-measurable, then the sublevel set

$$\mathcal{C}_t \;=\; \{\, m \in [0,1] \mid M_t(m) \leq 1/\alpha \} \tag{2}$$

gives a $(1-\alpha)$-CS [3, 8]. Ideally, $\mathcal{C}_t$ should shrink rapidly over time. Concretely, this means that for any candidate value $m \neq \mu$, the process $M_t(m)$ should cross the rejection threshold $1/\alpha$ rapidly, so that such values are excluded from $\mathcal{C}_t$ as early as possible.

Hedged Capital Process. Waudby-Smith and Ramdas [8] define $M_t(m)$ that is a test martingale for $m = \mu$ through the introduction of the *Hedged Capital Process* [8], or *wealth* process. It is based on two functions, respectively non-increasing and non-decreasing in $m \in [0,1]$:

$$K_t^+(m) \;=\; \prod_{i=1}^{t}\big(1 + \lambda_i^+(m)\,(Y_i - m)\big), \quad K_t^-(m) \;=\; \prod_{i=1}^{t}\big(1 - \lambda_i^-(m)\,(Y_i - m)\big),$$

with $K_0^+(m) = K_0^-(m) = 1$, where $\lambda_i^+(m) \in [0, 1/m)$ and $\lambda_i^-(m) \in [0, 1/(1-m))$ are data-dependent "stakes". Concrete definitions are given in App. G.1.2. The Hedged Capital Process is defined as the weighted maximum between the two functions,

$$K_t(m) \;=\; \max\{\theta K_t^+(m),\, (1-\theta) K_t^-(m)\}, \tag{3}$$

for a weighting parameter $\theta \in [0,1]$.

Remark. Waudby-Smith and Ramdas initially introduce this wealth process as the convex-combination

$$\dot{K}_t(m) \;=\; \theta K_t^+(m) + (1-\theta) K_t^-(m), \tag{4}$$

and show that $\dot{K}_t(m)$ is a test martingale when $m = \mu$. Applying Ville's inequality to $\dot{K}_t(m)$ therefore forms a valid CS via Eq. (2). Since $K_t(m) \leq \dot{K}_t(m)$ for all t and m, the sublevel set obtained by $K_t(m)$ defines a conservative outer CS and is thus also valid. The preference for the less powerful max-form in Eq. (3) is motivated by its properties; $K_t(m)$ is nonnegative and quasiconvex, so the sublevel set in Eq. (2) is an interval, computable by root finding.

3 From One-Dimensional Confidence Intervals to Higher-Dimensional Sets

Having described univariate CSs in § 2, we now turn to the multivariate setting. Denote the vector-valued observations $\mathbf{y}_t \in [0,1]^D$ with conditional mean $\mathbb{E}[\mathbf{y}_t \mid \mathbf{y}_{1:t-1}] = \boldsymbol{\mu} \in [0,1]^D$. The goal is to construct wealth processes $K_1(\mathbf{m}), K_2(\mathbf{m}), \ldots$ such that

$$\mathcal{C}_t \;=\; \{\, \mathbf{m} \in [0,1]^D \mid K_t(\mathbf{m}) \leq 1/\alpha \} \tag{5}$$

forms a valid $(1-\alpha)$-CS. The next subsections formalize three such multivariate lifts: (i) a Bonferroni rectangle $\mathcal{C}_t^{\text{bonf}}$, (ii) an equivalent max-form $\mathcal{C}_t^{\max}$, and a (iii) portfolio construction $\mathcal{C}_t^{\text{port}}$, and compare their relations. This comparison motivates $\mathcal{C}_t^{\text{port}}$ as the tightest construction in this family, but also computationally intractable. Wealth processes for scalar m are denoted by lowercase $k_t(\cdot)$, and those for vector $\mathbf{m}$ by uppercase $K_t(\cdot)$.

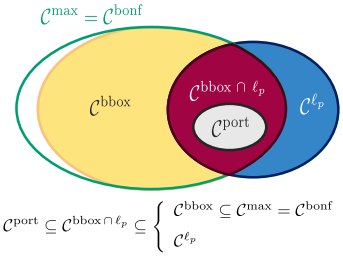


Figure 2: Venn diagram depicting the relations among the various sets.

3.1 Bonferroni Hyperrectangle

Suppose we have D wealth processes, one for each coordinate of $\mathbf{m} = (m_1, \ldots, m_D)$, denoted by $k_t^1(m_1), \ldots, k_t^D(m_D)$. A straightforward multivariate lift is obtained via a weighted Bonferroni correction [10, 11].

Given a weight vector $\mathbf{w} \in \Delta^D$ on the probability simplex (i.e., $\sum_{d=1}^D w_d = 1$ and $w_d > 0$), we assign an error level αw_d to the d-th coordinate. This forms the Bonferroni confidence set

$$\mathcal{C}_t^{\text{bonf}} := \left\{ \mathbf{m} \in [0,1]^D \;\middle|\; k_t^d(m_d) \leq \tfrac{1}{\alpha w_d},\ \forall d \in [D] \right\}.$$

When all weights are equal, i.e., $w_d = 1/D$, this reduces to the classical Bonferroni correction [12, 13]. In § 3.2 we indirectly show that $\mathcal{C}_t^{\text{bonf}}$ does indeed form a valid (1-α)-CS through a proof on its equivalent max-form construction.

3.2 Maximum of Wealth Processes

The weighted Bonferroni correction enables the construction of a multivariate CS, but does not make the associated wealth process $K_t(\cdot)$ over vectors $\mathbf{m}$ explicit. Here, we introduce an equivalent formulation in terms of a single wealth process. Consider the max-wealth process

$$K_t^{\max}(\mathbf{m}) = \max_{d \in [D]} w_d k_t^d(m_d),$$

with $\mathbf{w} \in \Delta^D$, and define the corresponding confidence set

$$\mathcal{C}_t^{\max} := \left\{ \mathbf{m} \in [0,1]^D \mid K_t^{\max}(\mathbf{m}) \leq \frac{1}{\alpha} \right\}.$$

Proposition 3.1. *The set $\mathcal{C}_t^{\max}$ is a valid $(1-\alpha)$-CS.*

The proof follows by showing that $K_t^{\max}(\boldsymbol{\mu})$ is upper bounded by a test martingale and applying Ville's inequality; full details are provided in App. A. The equivalence $\mathcal{C}_t^{\text{bonf}} = \mathcal{C}_t^{\max}$ is shown in Thm. B.1 in the appendix. This equivalence also (re)confirms the validity of $\mathcal{C}_t^{\text{bonf}}$ as a CS.

The main advantage of (weighted) Bonferroni rectangles (and thus also their equivalent $\mathcal{C}_t^{\max}$ form) is their simplicity. They are easy to compute, and the confidence set defines a D-dimensional hyperrectangle. However, eliminating candidates $\mathbf{m}$ from this set is achieved by considering only a single coordinate's wealth function $k_t^d(m_d)$ at a time. The dependence between coordinates is not exploited, typically making them overly conservative as D grows.

3.3 Portfolio Process

Inspired by the convex-combination baseline in [3], a stronger alternative that combines the wealths across coordinates is to consider a weighted sum of wealth processes (i.e., a *portfolio process*):

$$K_t^{\text{port}}(\mathbf{m}) = \sum_{d=1}^{D} w_d k_t^d(m_d), \tag{6}$$

with $\mathbf{w} \in \Delta^D$. It is straightforward to verify, in an analogous fashion to Prop. 3.1, that the set

$$\mathcal{C}_t^{\text{port}} := \left\{ \mathbf{m} \in [0,1]^D \mid K_t^{\text{port}}(\mathbf{m}) \leq \frac{1}{\alpha} \right\} \tag{7}$$

is a $(1-\alpha)$-CS. The fact that $\mathcal{C}_t^{\text{port}} \subset \mathcal{C}_t^{\text{bonf}}$ is a known result [14]. For a proof, we refer to Thm. B.1 in the appendix.

While $\mathcal{C}_t^{\text{port}}$ can be substantially smaller than $\mathcal{C}_t^{\text{bonf}}$, its practical use raises a new challenge: how can we represent this confidence set in a form that allows efficient computation of quantities such as volume or maximum width? Although each coordinate-wise wealth function $k_t^d(\cdot)$ is quasiconvex, the weighted sum $K_t^{\text{port}}(\cdot)$ does not preserve this property in general, and the resulting sublevel set is not necessarily convex anymore.

For practical applications, we thus require tractable approximations. A naive approach would be to approximate $\mathcal{C}_t^{\mathrm{port}}$ by uniformly discretizing $[0,1]^D$ and evaluating $K_t^{\mathrm{port}}(\mathbf{m})$ at each grid point. However, this quickly becomes computationally infeasible: a discretization with N points per coordinate requires N^D evaluations of $K_t^{\mathrm{port}}(\cdot)$, which scales exponentially with the dimension.

In the following sections, we therefore develop efficient *outer* approximations to $\mathcal{C}_t^{\mathrm{port}}$ that have explicit representations, yet remain computationally efficient.

4 Approximating the Portfolio Process

We introduce three *outer* approximations, i.e., supersets, of the region $\mathcal{C}^{\mathrm{port}}$ that allow for easy computation of volume and maximum width: the bounding box $\mathcal{C}^{\mathrm{bbox}}$, the ℓ_p-ellipsoid $\mathcal{C}^{\ell_p}$, and their intersection $\mathcal{C}^{\mathrm{bbox}\,\cap\,\ell_p}$. Importantly, since these approximations contain $\mathcal{C}^{\mathrm{port}}$, they still form valid CSs. We omit the t-indexing for simplicity.

4.1 Bounding Box

We define the axis-aligned bounding box

$$\mathcal{C}^{\mathrm{bbox}} = \left\{\mathbf{x} \in [0,1]^D \mid x_d \in [l_d, u_d],\ \ \forall d \in [D]\right\}, \tag{8}$$

$$[l_d, u_d] = \left\{ x_d \in [0,1] \mid w_d k^d(x_d) \leq \frac{1}{\alpha} - \sum_{j \neq d} w_j k_\star^j \right\}, \tag{9}$$

with $k_\star^j$ the global minimum of $k^j(\cdot)$ (derivation in App. E). It satisfies $\mathcal{C}^{\mathrm{port}} \subseteq \mathcal{C}^{\mathrm{bbox}} \subseteq \mathcal{C}^{\max} = \mathcal{C}^{\mathrm{bonf}}$. Intuitively, $\mathcal{C}^{\mathrm{bbox}}$ defines each coordinate's interval assuming the *minimum possible contribution* of all other coordinates (their global minima), and thus, still contains the full portfolio set $\mathcal{C}^{\mathrm{port}}$. At the same time, these coordinate-wise constraints remain tighter than the max/Bonferroni threshold, which gives $\mathcal{C}^{\mathrm{bbox}} \subseteq \mathcal{C}^{\max} = \mathcal{C}^{\mathrm{bonf}}$. Formal proofs are given in App. B. A discussion on how to compute the bounding box in practice, together with a principled method to estimate the global minimum $k_\star^j$, follows in § 5.

4.2 ℓ_p-Ellipsoid

The $\mathcal{C}^{\mathrm{bbox}}$ in Eq. (8) is simple and efficiently computable, but can become a loose approximation when $\mathcal{C}^{\mathrm{port}}$ does not have an approximately rectangular shape. An alternative is to outer-approximate $\mathcal{C}^{\mathrm{port}}$ through an ℓ_p-hyper-ellipsoid $\mathcal{C}^{\ell_p}$. Concretely, we replace each function $k^d(x)$ with a simpler counterpart $\tilde{k}^d(x)$, resulting in the sublevel set

$$\mathcal{C}^{\ell_p} := \{\mathbf{x} \mid \sum_{d=1}^{D} w_d \tilde{k}^d(x_d) \leq \frac{1}{\alpha}\}, \quad \tilde{k}^d(x) := \frac{|x - a_d|^p}{b_d^p} - c_d. \tag{10}$$

with parameters $a_d, b_d, c_d \in \mathbb{R}$, $b_d > 0$, and a shared degree $p \in \mathbb{N}_{>0}$. The practical benefit of this construction is that the sublevel set $\mathcal{C}^{\ell_p}$ can be expressed as an axis-aligned ℓ_p-hyper-ellipsoid:

$$\sum_{d=1}^{D} \frac{|x_d - a_d|^p}{(b'_d)^p} \ \leq\ 1, \quad b'_d := \frac{b_d \sqrt[p]{1/\alpha + \sum_{j=1}^{D} w_j\, c_j}}{\sqrt[p]{w_d}},$$

and thus its volume can be computed in closed form [15, 16]: $V(\mathcal{C}^{\ell_p}) = 2^D \left(\prod_{d=1}^{D} b'_d\right) \frac{\Gamma\left(1+\frac{1}{p}\right)^D}{\Gamma\left(1+\frac{D}{p}\right)}$. Here, $\Gamma(\cdot)$ denotes the Gamma function. Note that for $p = 2$, we get the conventional D-dimensional ellipsoid (see App. F.1 for details on the derivation and the generalization to its matrix form).

To ensure that $\mathcal{C}^{\mathrm{port}} \subseteq \mathcal{C}^{\ell_p}$, it suffices that $\tilde{k}^d(x) \leq k^d(x)$ for all $d \in [D]$ and all x (proven in Thm. B.1). The domain constraint can be made more specific: it is not necessary to enforce this inequality over the entire domain $x \in [0,1]$. Instead, it suffices to satisfy the constraint on a restricted interval $[x^d_{\text{start}}, x^d_{\text{end}}] \subset [0,1]$, where $(x^d_{\text{start}}, x^d_{\text{end}})$ denote the boundary points of the interval

$\{x \mid k^d(x) \leq \frac{1}{\alpha w_d}\}$. As shown in the proof, outside this interval, the constraint is irrelevant for membership in $\mathcal{C}^{\text{port}}$, which allows us to restrict attention to a smaller domain when fitting $\tilde{k}^d$. How to obtain these parameters such that $\tilde{k}^d(x) \leq k^d(x)$ is described in § 5.3.

4.3 Intersection: Bounding Box $\cap$ ℓ_p-Ellipsoid

Finally, we also consider the intersection $\mathcal{C}^{\text{bbox} \cap \ell_p} := \mathcal{C}^{\text{bbox}} \cap \mathcal{C}^{\ell_p}$. Since $\mathcal{C}^{\text{port}} \subseteq \mathcal{C}^{\text{bbox}}$ and $\mathcal{C}^{\text{port}} \subseteq \mathcal{C}^{\ell_p}$, it immediately follows that $\mathcal{C}^{\text{port}} \subseteq \mathcal{C}^{\text{bbox} \cap \ell_p}$. This completes the set relations shown in Fig. 2, with the details and formal proofs described in App. B.

Among the proposed approximations, $\mathcal{C}^{\text{bbox} \cap \ell_p}$ is also our smallest superset. However, contrary to $\mathcal{C}^{\text{bbox}}$ and $\mathcal{C}^{\ell_p}$, to our knowledge, there is no longer an analytical solution for the volume. As a result, volume estimation must rely on hit-or-miss Monte Carlo methods. A practical advantage is that the bounding rectangle from $\mathcal{C}^{\text{bbox}}$ can serve as the sampling region, instead of the naive $[0,1]^D$ hypercube. Since $\mathcal{C}^{\text{bbox}}$ is typically much closer to $\mathcal{C}^{\text{bbox} \cap \ell_p}$, a larger fraction of sampled points fall inside the target region, reducing the number of samples needed for a given precision.

Moreover, when membership evaluation in $\mathcal{C}^{\text{port}}$ is computationally expensive (e.g., due to costly evaluations of $k^d(\cdot)$), one can first construct $\mathcal{C}^{\text{bbox} \cap \ell_p}$ using only a limited number of such evaluations, and then perform all subsequent membership tests directly on $\mathcal{C}^{\text{bbox} \cap \ell_p}$, which is computationally inexpensive.

Certain other interesting properties of $\mathcal{C}^{\text{bbox} \cap \ell_p}$ are also easy to bound. For instance, for its maximum width, we have: $\text{width}(\mathcal{C}^{\text{bbox} \cap \ell_p}) \leq \min\{\text{width}(\mathcal{C}^{\text{bbox}}), \text{width}(\mathcal{C}^{\ell_p})\}$.

5 Computing the Regions in Practice

Having defined the outer approximations $\mathcal{C}^{\text{bbox}}$ and $\mathcal{C}^{\ell_p}$, we now turn to their implementation, with emphasis on preserving statistical validity: the numerical procedures should only enlarge the proposed regions, so the resulting computed sets remain valid CSs.

5.1 A Lower Bound on the Global Minimum of $k(\cdot)$

Several steps in constructing $\mathcal{C}^{\text{bbox}}$ and $\mathcal{C}^{\ell_p}$ require a *conservative* estimate of the global minimum

$$k_\star := \min_{x \in [0,1]} k(x) \quad \text{with} \quad k(x) = \max\{\theta k^+(x), (1-\theta)k^-(x)\}$$

the one-dimensional function defined in Eq. (3). Rather than estimating $k_\star$ via dense grid search, we exploit the monotonic structure of k^+ and k^- to obtain a lower-bounded estimate through a small number of root-finding calls. Concretely, we show that (i) minimizers always exist, (ii) they are characterized by the roots of $h(x) = (1-\theta)k^-(x) - \theta k^+(x)$, and (iii) based on this result derive Alg. 1 which returns explicit upper and lower bounds on $k_\star$ (together with minimizer estimate $\hat{x}$). Thus, the resulting procedure is theory-driven and ensures that $\underline{k}^\star \leq k_\star$, with the gap controlled by a preset tolerance. See App. D for details.

5.2 Computing the Bounding Box

Following § 4.1, $\mathcal{C}^{\text{bbox}}$ is given by $\{\mathbf{x} \mid w_d k^d(x_d) \leq \frac{1}{\alpha} - \sum_{j \neq d} w_j k_\star^j, \quad \forall d \in [D]\}$. Since the true minima $k_\star^j$ are unknown, we use lower bounds $\underline{k_\star^j}$ from ESTIMATE-$k_\star$ (Alg. 1), which also returns minimizer estimates $\hat{x}_j$. As $\underline{k_\star^j} \leq k_\star^j$, this enlarges the box. Given $\hat{x}_d$, a conservative interval $[l_d, u_d]$ is obtained by solving $w_d k^d(x_d) - (\frac{1}{\alpha} - \sum_{j \neq d} w_j \underline{k_\star^j}) - \epsilon = 0$ on $[0, \hat{x}_d]$ and $[\hat{x}_d, 1]$, with constant $\epsilon > 0$ ensuring enclosure in case of flat regions. Applying this for all $d \in [D]$ yields $\mathcal{C}^{\text{bbox}}$, which remains a valid outer approximation of $\mathcal{C}^{\text{port}}$ after implementation.

Computational Complexity. Using bisection for root finding, the global minimum of each $k^d(\cdot)$ is first approximated using one bisection call per dimension. Next, the endpoints l_d and u_d are computed via two additional bisection calls per dimension. In total, this results in $3D$ bisections. Let C_{eval} denote the cost of evaluating $k^d(\cdot)$. Since bisection converges in $O(\log(1/\varepsilon))$ iterations, the total

time complexity is $O(D\, C_{\text{eval}}\, \log(1/\varepsilon))$, which is equivalent to the complexity of the Bonferroni construction $\mathcal{C}^{\text{bonf}}$.

5.3 Computing the ℓ_p-Ellipsoid

Obtaining an efficient implementation of the ℓ_p-ellipsoid while preserving validity requires substantially more technical care. Recall that $\mathcal{C}^{\ell_p} := \{\mathbf{x} \mid \sum_{d=1}^{D} w_d\, \tilde{k}^d(x_d) \le \frac{1}{\alpha}\}$, where each $\tilde{k}^d(x) := \frac{|x-a_d|^p}{b_d^p} - c_d$ is constructed to satisfy $\tilde{k}^d(x) \le k^d(x)$ for all $x \in [x^d_{\text{start}}, x^d_{\text{end}}]$. We describe two variants: (i) an approximate construction that enforces this condition only on a grid, and (ii) a conservative modification $\mathcal{C}^{\ell_p\text{safe}}$ that guarantees $\mathcal{C}^{\text{port}} \subseteq \mathcal{C}^{\ell_p\text{safe}}$.

Overview. For a fixed degree p, we fit parameters $\{(a_d, b_d, c_d)\}_{d=1}^{D}$ independently per dimension. We then select p by a small grid search over candidates $\mathcal{P}$ (e.g., $\{6, 7, \ldots, 16\}$), choosing the minimum-volume ellipsoid:

$$p^\star = \arg\min_{p\in\mathcal{P}} V\left(\mathcal{C}^{\ell_p}(p)\right).$$

Per-Dimension Fit via Two Touch Points. Rather than approximating $k^d(\cdot)$ globally, we enforce a good fit only in the most relevant regions. We select a target level $\gamma_d > k^d_\star$ and fit $\tilde{k}^d(\cdot)$ so that it (approximately) touches $k^d(\cdot)$ at two points. Outside this region, some slack is allowed (Fig. 3). Formally, given two touch points $x^{\text{touch}}_{L,d}$ and $x^{\text{touch}}_{R,d}$, we enforce

$$\tilde{k}^d(x^{\text{touch}}_{L,d}) = \tilde{k}^d(x^{\text{touch}}_{R,d}) = \gamma_d, \tag{11}$$

$$\tilde{k}^d(x_i) \le k^d(x_i) \;\text{ for all } x_i \in \{x_i\}_{i=1}^{N}, \tag{12}$$

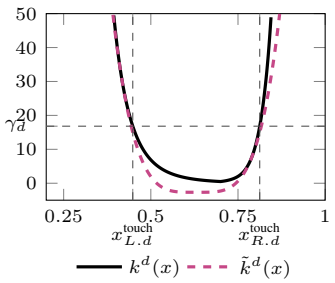


Figure 3: Example where $\tilde{k}^d(x)$ lower-bounds $k^d(x)$, touching γ_d at two points; tight near γ_d and looser farther away.

where $\{x_i\}_{i=1}^{N}$ is a fine grid on the restricted domain $[x^d_{\text{start}}, x^d_{\text{end}}]$. Importantly, the fit is largely analytic: a_d is fixed directly by the two touch-point constraints, and c_d is obtained in closed form once b_d is chosen. Numerically, the procedure reduces to a single pass over $\{x_i\}_{i=1}^{N}$ to obtain feasibility bounds for b^p_d. Full derivations, the choice of γ_d, and fallback procedures (including adaptive touch-point widening and domain extension) are given in App. F.2.

Discretization and Safe Option. Constraint (12) enforces $\tilde{k}^d(x_i) \le k^d(x_i)$ only on a grid $\{x_i\}_{i=1}^{N}$. In App. F.6, we derive an upper bound on the worst-case violation between the grid points:

$$\max_{x\in[x^d_{\text{start}},x^d_{\text{end}}]} \left(\tilde{k}^d(x) - k^d(x)\right) \;\le\; \max_i \Big(\max\{\tilde{k}^d(x_i), \tilde{k}^d(x_{i+1})\} - L^d_i\Big),$$

where L^d_i is a computable lower bound on $k^d(\cdot)$ over the interval $[x_i, x_{i+1}]$ (defined in Thm. F.2). Based on this bound, we introduce the "safe" construction $\mathcal{C}^{\ell_p\text{safe}}$ that modifies the grid points so that $\tilde{k}^d(x) \le k^d(x)$ holds **for all** $x \in [x^d_{\text{start}}, x^d_{\text{end}}]$ (Prop. F.1). This safe mode can then be used when a true superset guarantee is required. Tab. 1 summarizes all outer approximations and whether they remain valid after implementation.

Computational Complexity. Let N denote the grid size per dimension and let $N_p = |\mathcal{P}|$ denote the number of p-candidates. The ellipsoid construction has complexity $O(D\, N_p\, N\, (\log(L/\text{tol}) + C_{\text{eval}}))$, where C_{eval} is the cost of evaluating $k^d(\cdot)$ and the logarithmic term comes from the bisection-based fallback procedures. Algorithms and a detailed breakdown are provided in App. F.3.

6 Empirical Results

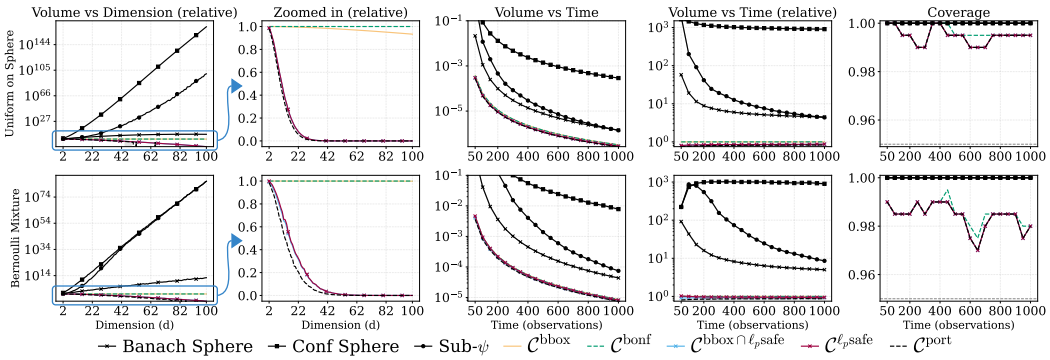


Figure 4: **Comparison on synthetic data**. The proposed constructions consistently achieve smaller volumes while maintaining the target coverage. "Relative" plots report $V(\text{method})/V(\mathcal{C}^{\text{bonf}})$.

We evaluate the tightness and efficiency of the proposed CSs, compare them against recent alternatives, and demonstrate their practical use in two case studies. We show that the portfolio is the tightest but intractable, while our approximations recover much of this gap while remaining efficient.

Table 1: Overview of our outer approximations and whether they form a valid CS after implementation. $\sim$ indicates an approximation without a formal CS guarantee.

Approximation	Valid CS
$\mathcal{C}^{\text{bbox}}$	✓
$\mathcal{C}^{\ell_p}$	$\sim$
$\mathcal{C}^{\ell_p\text{safe}}$	✓
$\mathcal{C}^{\text{bbox}\cap\ell_p}$	$\sim$
$\mathcal{C}^{\text{bbox}\cap\ell_p\text{safe}}$	✓

Baselines. We compare with four recent CSs: *Banach Sphere* [4], *Conf Sphere* [5], and *Sub-ψ ellipsoids* [6], all of which have analytical solutions and are inexpensive to compute. We refer to smaller-scale results for the bounded-vector reduction *Bounded UP* [3] in App. G, since its full comparisons are infeasible due to lack of a closed-form solution and membership checking costs $O(t^{D+1})$.

Synthetic Experiments Setup. We use two synthetic multivariate distributions: **(i)** i.i.d. samples from the unit ball (favors sphere baselines), and **(ii)** a Bernoulli-mixture construction in $[0,1]^D$ (favors our constructions and Bounded UP). Full details are in App. G.

Tightness and Coverage. Fig. 4 compares the various CSs, and demonstrates how across all settings, our proposed regions achieve substantially smaller volumes than the sphere-based baselines (columns 1, 3, and 4). This also results in less conservative CSs with a per-time-step coverage closer to the $1-\alpha$ target (column 5).

Approximation Quality. We observe from Fig. 4, column 2, that the gap between $\mathcal{C}^{\text{bonf}}$ and $\mathcal{C}^{\text{port}}$ rapidly grows with the dimension. For example, at $D=10$, the volume of $\mathcal{C}^{\text{port}}$ is up to 50% smaller than that of $\mathcal{C}^{\text{bonf}}$. Both the ℓ_p-ellipsoid $\mathcal{C}^{\ell_p\text{safe}}$ and the intersection $\mathcal{C}^{\text{bbox}\cap\ell_p\text{safe}}$ approximate $\mathcal{C}^{\text{port}}$ well, showing a clear improvement over the Bonferroni construction, yet at substantially lower computational cost than $\mathcal{C}^{\text{port}}$. Finally, the bounding-box construction $\mathcal{C}^{\text{bbox}}$ produces slightly smaller regions than $\mathcal{C}^{\text{bonf}}$ at the exact same computational complexity.

Computational Cost Tab. 2 reports the number of $k_t^d(\cdot)$ evaluations required to obtain volumes at $t=100$. Our approximations are orders of magnitude cheaper than direct $\mathcal{C}^{\text{port}}$ gridding (N^D), with $N=2001$. For $\mathcal{C}^{\text{bbox}\cap\ell_p\text{safe}}$, we count only the construction-time $k_t^d(\cdot)$ calls; although its volume has no closed-form expression, once constructed, it can be estimated using inexpensive membership tests on the bounding box and the ℓ_p-ellipsoid, rather than direct evaluations of $k_t^d(\cdot)$. The external baselines are not included since they do not define $k_t^d(\cdot)$. We instead report wall-clock evaluation times in Tab. 4.

Paired A/B-Testing Experiment of Simulated Web Metrics. We compare two versions of a website across ten simulated metrics, with the goal of determining whether one version shows a statistically significant difference for a particular metric.

Table 2: Number of $k_t^d(\cdot)$ calls to compute the volume at $t = 100$ (sorted from most to least efficient). The outer approximations are orders of magnitude more efficient than $\mathcal{C}^{\text{port}}$'s prohibitive N^D cost.

D	$\mathcal{C}^{\text{bbox}}$	$\mathcal{C}^{\text{bonf}}$	$\mathcal{C}^{\ell_p\text{safe}}$	$\mathcal{C}^{\text{bbox}} \cap \mathcal{C}^{\ell_p\text{safe}}$	$\mathcal{C}^{\text{port}}$
2	5.01×10^{2}	5.05×10^{2}	1.60×10^{5}	1.61×10^{5}	8.00×10^{6}
3	7.78×10^{2}	7.78×10^{2}	2.40×10^{5}	2.41×10^{5}	2.40×10^{10}
4	1.02×10^{3}	1.02×10^{3}	3.20×10^{5}	3.21×10^{5}	6.40×10^{13}
5	1.30×10^{3}	1.30×10^{3}	4.00×10^{5}	4.01×10^{5}	1.61×10^{17}
6	1.54×10^{3}	1.55×10^{3}	4.80×10^{5}	4.82×10^{5}	3.85×10^{20}
7	1.85×10^{3}	1.85×10^{3}	5.61×10^{5}	5.63×10^{5}	8.96×10^{23}
8	2.11×10^{3}	2.12×10^{3}	6.41×10^{5}	6.43×10^{5}	2.06×10^{27}
9	2.40×10^{3}	2.40×10^{3}	7.22×10^{5}	7.24×10^{5}	4.63×10^{30}
10	2.68×10^{3}	2.68×10^{3}	8.03×10^{5}	8.06×10^{5}	1.03×10^{34}
Complexity	$O(D\,C_{\text{eval}}\,\log(1/\varepsilon))$	$O(D\,C_{\text{eval}}\,\log(1/\varepsilon))$	$O\big(D\,N_p\,N\,(\log(L/\text{tol}) + C_{\text{eval}})\big)$	$O\Big(\begin{smallmatrix}D\,C_{\text{eval}}\,\log(1/\varepsilon)+\\ D\,N_p\,N\big(\log(L/\text{tol})+C_{\text{eval}}\big)\end{smallmatrix}\Big)$	$O\big(DC_{\text{eval}}N^D\big)$

This is achieved by simulating batches of web-shop sessions for variants A and B, with each session described by ten metrics, including time-on-site, clicks, scroll depth, purchase behavior, and two noise variables. The simulation captures several aspects of real-world settings: the number of observations assigned to each variant may differ across batches, and correlations between metrics are induced through a shared latent variable. To simulate a treatment effect, variant B is optionally improved on a subset of metrics by slightly increasing this latent variable. Fig. 17 visualizes the resulting per-session metric distributions for a representative batch.

The metrics are fed to the CS as follows: let $\mathbf{y}_t^{(A)}$ and $\mathbf{y}_t^{(B)}$ denote the metric vectors for variants A and B in batch t, respectively, and define $\boldsymbol{\Delta}_t := \mathbf{y}_t^{(B)} - \mathbf{y}_t^{(A)}$. A CS then gives a region $\mathcal{C}_t$ for $\mathbb{E}[\boldsymbol{\Delta}_t]$ at every time t. If the interval for a particular coordinate lies entirely above or below 0, we can conclude that the variants differ significantly on that metric. On the other hand, if the region becomes sufficiently small without excluding 0 (measured through volume), we may conclude that there is no practically meaningful difference and stop collecting data. See App. G for details.

Tab. 3 shows that our methods achieve **much earlier stopping times** than existing tractable CSs when an effect is present, and in the no-effect setting, the ℓ_p-ellipsoid-based variants stop the fastest.

Multi-metric Model Comparison Experiment. To illustrate the value on a real-world dataset, we compare a challenger and baseline classifier on the Adult [17] dataset across six metrics: overall accuracy, Brier score, and four subgroup accuracies. The goal is to certify an improvement in overall accuracy while simultaneously monitoring subgroup degradations with statistical guarantees.

Table 3: **A/B stopping outcomes.** We stop for *significance* once a coordinate interval of $\boldsymbol{\Delta}_t$ excludes 0. Otherwise, we stop and report *unsubstantial* once volume has shrunken to a preset threshold. If at the end of the 10,000 batches no decision was made, we report *no decision*.

	With effect		No effect	
Method	Stop$_t$	Decision	Stop$_t$	Decision
ConfSphere	10,000	No decision	10,000	No decision
BanachSphere	5,018	Significant	8,052	Unsubstantial
Sub-ψ	5,198	Significant	6,650	Unsubstantial
$\mathcal{C}^{\text{bonf}}$	178	Significant	399	Unsubstantial
$\mathcal{C}^{\text{bbox}}$	178	Significant	398	Unsubstantial
$\mathcal{C}^{\ell_p\text{safe}}$	178	Significant	356	Unsubstantial
$\mathcal{C}^{\text{bbox}} \cap \mathcal{C}^{\ell_p\text{safe}}$	**178**	Significant	**356**	Unsubstantial

We let model A denote the baseline and model B the challenger. Both classifiers use features such as gender, education, and age to predict whether a person's income exceeds \$50,000. The setup is similar to the A/B experiment: each $\boldsymbol{\Delta}_t$ is defined as the difference between two metric vectors, with $\mathbf{y}_t^{(A)}$ representing the metrics of model A (fully defined in Eq. (33)) and $\mathbf{y}_t^{(B)}$ those of model B. Within these vectors, overall accuracy is represented by the indicator $\mathbf{1}\{f_A(\mathbf{x}_t) = f_t\}$ where f_t is the true label, while subgroup performance is represented through joint indicators, e.g., $\mathbf{1}\{f_A(\mathbf{x}_t) = f_t \text{ and } \mathbf{x}_t \text{ is woman}\}$. The main technical challenge is handling conditional accuracies, such as $\Pr(f_B(\mathbf{x}_t) = \text{correct} \mid \mathbf{x}_t \text{ is woman})$, while preserving anytime validity. We achieve this by expressing each conditional accuracy as a ratio of two means: the corresponding joint accuracy and an additional coordinate representing the subgroup proportion, e.g., $\Pr(\mathbf{x}_t \text{ is woman})$. Technical details are provided in App. G.6.

Our CSs detect an improvement in overall accuracy after about 356 batches, whereas existing tractable multivariate CS baselines remain inconclusive even after observing the entire test dataset of 509 batches. The intervals in Fig. 5 also flag possible subgroup degradation. While the new model shows a statistically significant improvement in overall accuracy, its accuracy on women shows a plausible performance decrease of up to 4%. This suggests a concrete next step, such as collecting more representative data or improving performance on that subgroup before deployment.

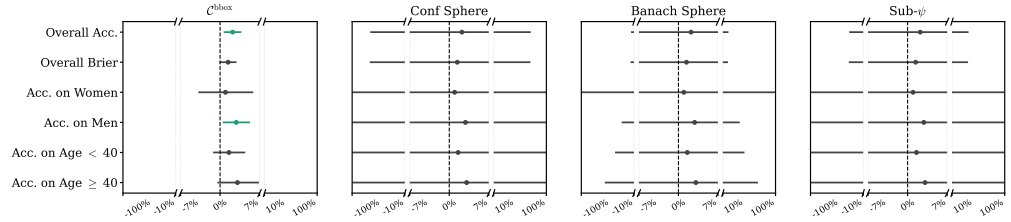


Figure 5: **Model-gain intervals on six Adult metrics** after 509 batches. Only $\mathcal{C}^{\text{bbox}}$ certifies overall-accuracy improvement (0.6%-3.7%), reaching significance at 356 batches; baselines remain inconclusive. The other hedged-capital variants behave similarly to $\mathcal{C}^{\text{bbox}}$ (Fig. 18).

7 Conclusion

We study three approaches for lifting one-dimensional confidence sequences (CSs), defined as sub-level sets of a wealth process, to higher-dimensional means. These include a weighted Bonferroni construction $\mathcal{C}^{\text{bonf}}$, an equivalent maximum-wealth formulation $\mathcal{C}^{\max}$, and a portfolio-based construction $\mathcal{C}^{\text{port}}$, which leads to the tightest CS but is computationally intractable. For $\mathcal{C}^{\text{port}}$, we propose three *outer* approximations: a bounding box $\mathcal{C}^{\text{bbox}}$, an ℓ_p-ellipsoid $\mathcal{C}^{\ell_p}$, and their intersection.

We theoretically prove set relations among all constructions and obtain several interesting insights: we show that $\mathcal{C}^{\max}$ is equivalent to $\mathcal{C}^{\text{bonf}}$, and that the region of $\mathcal{C}^{\text{bbox}}$ is always contained in the corresponding Bonferroni region. At exactly the same computational cost, one can thus replace $\mathcal{C}^{\text{bonf}}$ by $\mathcal{C}^{\text{bbox}}$ and obtain an at least as tight valid CS. We also find that $\mathcal{C}^{\ell_p}$ tracks $\mathcal{C}^{\text{port}}$ closely in practice while allowing for a closed-form expression for its volume. Importantly, our proposed approximations produce regions that maintain the statistical validity of the CS, even after implementation.

Overall, the proposed CSs and their outer approximations lead to substantially smaller regions than existing tractable alternatives while remaining practical to compute in higher dimensions. The practical value is further highlighted in both the A/B testing and model-comparison case studies, which lead to detecting significance after substantially fewer observations, allowing practitioners to make reliable decisions earlier and reduce data-collection costs.

Limitations and Future Work While our outer approximations define valid CSs, the ℓ_p-ellipsoid relies on per-dimension discretization, which can be costly for dense grids. Also, the bound used for its safe construction is very general, requiring only that the underlying function is quasiconvex, continuous, and globally lower bounded by a scalar ϕ. However, this generality can make the bound conservative, requiring a dense grid to compensate. Moreover, our smallest outer approximation, given by the intersection $\mathcal{C}^{\text{bbox}} \cap \mathcal{C}^{\ell_p}$, no longer has a closed-form volume expression and must therefore be estimated using Monte Carlo sampling. This estimate becomes less accurate as the dimension increases, requiring more samples to maintain the same level of accuracy. Several directions for future work remain. First, one could study whether other wealth processes (e.g., Ryu and Wornell [3]) can be integrated into the proposed framework. Second, the computational efficiency of our procedures could be further improved through more advanced root-finding or adaptive discretization strategies. Third, one could investigate whether dependencies between coordinates can be exploited to construct tighter CSs. Finally, since CSs remain valid under intersections over time, $\mathcal{C}_t := \cap_{i \leq t} \mathcal{C}_i$, it would be interesting to incorporate this structure into the approximations as well.

Acknowledgments and Disclosure of Funding

We thank Dario Coscia and Rajeev Verma for the discussions that led to this work. We also thank José Oramas, Roel Hulsman, Alexander Timans, and Mátyás Schubert for their valuable input and discussions during the early stages of the project. This project was supported by the University of Antwerp, the University of Amsterdam, Eindhoven University of Technology, the Research Foundation - Flanders (FWO) under grant agreement V429625N, and the MOSAIC project under Grant Agreement No. 101194414.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

[1] Akash Maharaj, Ritwik Sinha, David Arbour, Ian Waudby-Smith, Simon Z Liu, Moumita Sinha, Raghavendra Addanki, Aaditya Ramdas, Manas Garg, and Viswanathan Swaminathan. Anytime-valid confidence sequences in an enterprise a/b testing platform. In *Companion Proceedings of the ACM Web Conference 2023*, pages 396–400, 2023.

[2] Ramesh Johari, Leo Pekelis, and David J Walsh. Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*, 2015.

[3] Jongha Jon Ryu and Gregory W. Wornell. Gambling-based confidence sequences for bounded random vectors. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=mu7Er7f9NQ.

[4] Diego Martinez-Taboada and Aaditya Ramdas. Empirical bernstein in smooth banach spaces. *arXiv preprint arXiv:2409.06060*, 2024.

[5] Ben Chugg, Hongjian Wang, and Aaditya Ramdas. Time-uniform confidence spheres for means of random vectors. *arXiv preprint arXiv:2311.08168*, 2023.

[6] Justin Whitehouse, Zhiwei Steven Wu, and Aaditya Ramdas. Time-uniform self-normalized concentration for vector-valued processes. *arXiv preprint arXiv:2310.09100*, 2023.

[7] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

[8] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

[9] Jean Ville. *Étude critique de la notion de collectif*. 1939. URL http://eudml.org/doc/192893.

[10] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4615733.

[11] Guolian Kang, Keying Ye, Nianjun Liu, David B Allison, and Guimin Gao. Weighted multiple hypothesis testing procedures. *Statistical applications in genetics and molecular biology*, 8(1): 23, 2009.

[12] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. doi: 10.1080/01621459.1961.10482090. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1961.10482090.

[13] Matthew A Napierala. What is the bonferroni correction? *Aaos now*, pages 40–41, 2012.

[14] Brian Cho, Kyra Gan, and Nathan Kallus. Peeking with PEAK: sequential, nonparametric composite hypothesis tests for means of multiple data streams. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=hcASxFvmZ5.

[15] Shahnawaz Ahmed and Elias G Saleeby. On volumes of hyper-ellipsoids. *Mathematics Magazine*, 91(1):43–50, 2018.

[16] Xianfu Wang. Volumes of generalized unit balls. *Mathematics Magazine*, 78(5):390–395, 2005. ISSN 0025570X, 19300980. URL http://www.jstor.org/stable/30044198.

[17] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[18] Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux—effortless bayesian deep learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 20089–20103, 2021.

[19] Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13132–13143, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/118921efba23fc329e6560b27861f0c2-Abstract.html`.

[20] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org, 2016. URL `http://proceedings.mlr.press/v48/gal16.html`.

[21] Fabian Denoodt and José Oramas. Efficient post-hoc uncertainty calibration via variance-based smoothing. *arXiv preprint arXiv:2503.15583*, 2025.

[22] Andreas Krause and Jonas Hübotter. Probabilistic artificial intelligence. *arXiv preprint arXiv:2502.05244*, 2025.

Appendix

This appendix is divided into three parts. Part I presents the core validity and set relation proofs as well as several auxiliary theoretical results. Part II focuses on the proposed outer approximations: it first derives the bounding-box construction, and then develops the ℓ_p-ellipsoid construction, together with its certified between-grid control and safe variant. Part III provides experimental setup details, together with ablations and additional case studies.

Part I: Theory and Proofs **14**

A Proof of Prop. 3.1 . . . 14
B Set Relations . . . 14
C Useful Lemmas . . . 16
C.1 Quasiconvexity of $\tilde{k}$ . . . 16
C.2 Continuity of $k(x)$ . . . 16
D Bounding the Global Minimum of $k(x)$ . . . 17

Part II: Outer Approximations and Practical Details **18**

E Bounding box . . . 18
F ℓ_p-Ellipsoid . . . 20
F.1 ℓ_p-Ellipsoid Derivation . . . 20
F.2 Fitting $\tilde{k}(x)$ from Two Touch Points . . . 21
F.3 Algorithms . . . 23
F.4 Analysis of $R(x)$ when dividing by zero . . . 25
F.5 Side Note - Function Approximation with Two Distinct Touch Levels . . . 27
F.6 A bound on the Between-Grid Violation of $\tilde{k}$ . . . 29

Part III: Experiments **32**

G Experimental Setup & Additional Experiments . . . 32
G.1 Setup . . . 32
G.2 Small-Scale Experiments . . . 34
G.3 $\mathcal{C}^{\ell_p}$ vs $\mathcal{C}^{\ell_p\text{safe}}$ . . . 34
G.4 Ablation Studies . . . 34
G.5 A/B-Testing Experiment: Synthetic Web Metrics . . . 40
G.6 Model Comparison on Adult . . . 42

Part IV: Varia **45**

H Additional Related Work . . . 45
I Broader Impacts . . . 45
J Declaration of LLM Usage . . . 46

Part I: Theory and Proofs

A Proof of Prop. 3.1

Proposition A.1. *The set $\mathcal{C}_t^{\max}$ is a valid $(1-\alpha)$-CS.*

Proof. First, note that for any $\mathbf{m} \in [0,1]^D$,

$$K_t^{\max}(\mathbf{m}) = \max_d \, w_d k_t^d(m_d) \;\le\; \sum_{d=1}^D w_d k_t^d(m_d)$$
$$\le\; \sum_{d=1}^D w_d \dot{k}_t^d(m_d), \tag{13}$$

where the inequalities use the non-negativity of w_d, $k_t^d(\cdot)$, and $\dot{k}_t^d(\cdot)$ with $\dot{k}(\cdot)$ defined in Eq. (4).

Since each $\dot{k}_t^d(m_d)$ is a test martingale (for $m_d = \mu_d$), their weighted sum is also a test martingale:

$$\mathbb{E}\left[\sum_{d=1}^D w_d \dot{k}_t^d(m_d) \,\middle|\, \mathcal{F}_{t-1}\right] = \sum_{d=1}^D w_d \mathbb{E}\left[\dot{k}_t^d(m_d) \,\middle|\, \mathcal{F}_{t-1}\right]$$
$$= \sum_{d=1}^D w_d \dot{k}_{t-1}^d(m_d).$$

By Ville's inequality (2), the sublevel set

$$\left\{\mathbf{m} : \sum_{d=1}^D w_d \dot{k}_t^d(m_d) \le \tfrac{1}{\alpha}\right\}$$

is therefore a valid $(1-\alpha)$-CS. Finally, since $K_t^{\max}(\mathbf{m})$ is upper bounded by a test martingale, its sublevel set

$$\left\{\mathbf{m} : K_t^{\max}(\mathbf{m}) \le \tfrac{1}{\alpha}\right\}$$

is a conservative outer CS and hence also valid.

□

B Set Relations

Theorem B.1. *Assume $\alpha \in (0,1]$, $D \in \mathbb{N}_{>0}$, and $\mathbf{w} \in \Delta^D$ fixed with $w_d > 0$ for all d, then the following relations hold:*

$$\mathcal{C}^{\text{port}} \subseteq \mathcal{C}^{\textit{bbox} \cap \ell_p} \subseteq \mathcal{C}^{\textit{bbox}} \subseteq \mathcal{C}^{\max} = \mathcal{C}^{\textit{bonf}},$$
$$\mathcal{C}^{\textit{bbox} \cap \ell_p} \subseteq \mathcal{C}^{\ell_p}.$$

Proof. We show the relations:

1. $\mathcal{C}^{\max} = \mathcal{C}^{\text{bonf}}$.

   Denote

   $$\mathcal{C}^{\max} = \left\{\mathbf{x} \,\middle|\, \max_{d\in[D]} \, w_d \, k^d(x_d) \le \frac{1}{\alpha}\right\} \quad \text{and} \quad \mathcal{C}^{\text{bonf}} = \left\{\mathbf{x} \,\middle|\, \forall d \in [D] : k^d(x_d) \le \frac{1}{\alpha w_d}\right\}.$$

   We have, for any $\mathbf{x} \in [0,1]^D$,

   $$\mathbf{x} \in \mathcal{C}^{\max} \iff \max_{d\in[D]} w_d k^d(x_d) \le \tfrac{1}{\alpha} \iff \forall d \in [D] :\; w_d k^d(x_d) \le \tfrac{1}{\alpha} \tag{14}$$
   $$\iff \forall d \in [D] :\; k^d(x_d) \le \tfrac{1}{\alpha w_d} \iff \mathbf{x} \in \mathcal{C}^{\text{bonf}}, \tag{15}$$

   which shows that $\mathcal{C}^{\max} = \mathcal{C}^{\text{bonf}}$.

2. $\mathcal{C}^{\text{bbox}} \subseteq \mathcal{C}^{\max}$.

From the definition in Eq. (8),

$$[l_d, u_d] = \left\{ x_d \in [0,1] \mid w_d k^d(x_d) \leq \frac{1}{\alpha} - \sum_{j \neq d} w_j k_\star^j \right\}.$$

Hence

$$\mathcal{C}^{\text{bbox}} = \left\{ \mathbf{x} \in [0,1]^D \,\middle|\, \forall d: \; w_d k^d(x_d) \leq \frac{1}{\alpha} - \sum_{j \neq d} w_j k_\star^j \right\}.$$

Let $\mathbf{x} \in \mathcal{C}^{\text{bbox}}$. Then for each d,

$$w_d k^d(x_d) \leq \tfrac{1}{\alpha} - \sum_{j \neq d} w_j k_\star^j \leq \tfrac{1}{\alpha}.$$

Since this holds for every d,

$$\max_d \big(w_d k^d(x_d) \big) \leq \tfrac{1}{\alpha}.$$

Hence $\mathbf{x} \in \mathcal{C}^{\max}$.

Note that the reverse direction is not necessarily true, as shown by the example in Fig. 6. For an equality between the two sets to hold, $k^d(\cdot)$ should have a minimum = 0 ($k_\star^d = 0$) for all d:

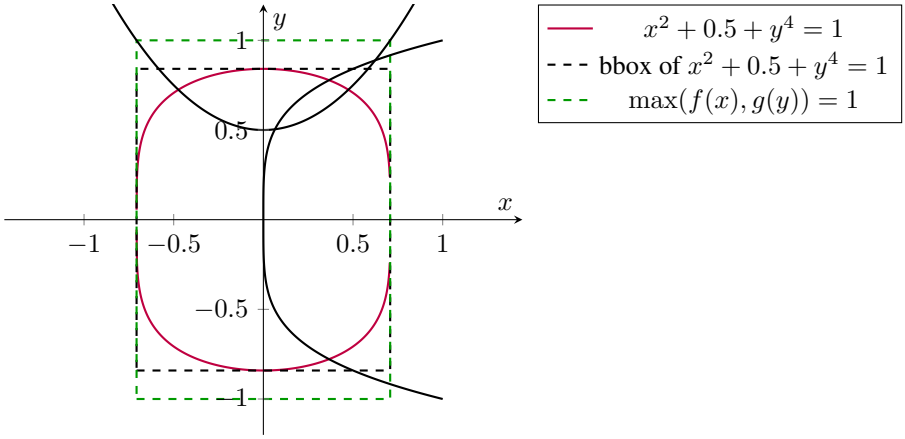


Figure 6: Base functions: $f(x) = x^2 + 0.5$, $g(y) = y^4$. The figure depicts the set $f(x) + g(y) = 1$, its bounding-box and $\max\left(f(x),\, g(y)\right) = 1$ superset. We see that the bounding box and max are not necessarily equivalent.

3. $\mathcal{C}^{\text{port}} \subseteq \mathcal{C}^{\text{bbox}}$.

Let $\mathbf{x} \in \mathcal{C}^{\text{port}}$ so that

$$\sum_{d=1}^{D} w_d k^d(x_d) \leq \frac{1}{\alpha} \iff$$

$$\sum_{d \neq j}^{D} w_d k^d(x_d) + w_j k^j(x_j) \leq \frac{1}{\alpha} \implies$$

$$w_j k^j(x_j) \leq \frac{1}{\alpha} - \sum_{d \neq j} w_d k^d(x_d) \leq \tfrac{1}{\alpha} - \sum_{d \neq j} w_d k_\star^d.$$

Hence, we have obtained that $\forall j : w_j k^j(x_j) \leq \frac{1}{\alpha} - \sum_{d \neq j} w_d k_\star^d$, i.e., $x_j \in [l_j, u_j]$ by Eq. (8), such that $\mathbf{x} \in \mathcal{C}^{\text{bbox}}$.

4. $\mathcal{C}^{\text{port}} \subseteq \mathcal{C}^{\ell_p}$.

   Define $\mathcal{C}^{\ell_p} = \left\{\mathbf{x} | \sum_d^D w_d \tilde{k}^d(x_d) \leq \frac{1}{\alpha}\right\}$, with $\tilde{k}^d(x) \leq k^d(x)$ for all $x \in [0,1]$.

   Let $\mathbf{x} \in \mathcal{C}^{\text{port}}$ such that

$$\sum_d^D w_d k^d(x_d) \leq \frac{1}{\alpha} \Rightarrow \sum_d^D w_d \tilde{k}^d(x_d) \leq \frac{1}{\alpha},$$

   since $\tilde{k}^d(x) \leq k^d(x)$, so $\mathbf{x} \in \mathcal{C}^{\ell_p}$.

   Note that when we only have that $\tilde{k}^d(x) \leq k^d(x)$ for all $x \in [x^d_{\text{start}}, x^d_{\text{end}}]$, with $(x^d_{\text{start}}, x^d_{\text{end}})$ defined as the end-point solutions of the interval: $\{x_d | k^d(x_d) \leq \frac{1}{\alpha w_d}\}$, the relation still holds. Indeed, since for $\mathbf{x} \in \mathcal{C}^{\text{port}}$ we have that

$$\sum_d w_d k^d(x_d) \leq 1/\alpha \Rightarrow w_d k^d(x_d) \leq 1/\alpha \Rightarrow k^d(x_d) \leq \frac{1}{\alpha w_d}.$$

   I.e., for $\mathbf{x} \in \mathcal{C}^{\text{port}}$, all of its coordinates must satisfy $k^d(x_d) \leq \frac{1}{\alpha w_d}$ (and hence we have that $x_d \in [x^d_{\text{start}}, x^d_{\text{end}}]$).

5. $\mathcal{C}^{\text{bbox} \cap \ell_p} \subseteq \mathcal{C}^{\text{bbox}}$ and $\mathcal{C}^{\text{bbox} \cap \ell_p} \subseteq \mathcal{C}^{\ell_p}$.

   This is evident from its definition:

$$\mathcal{C}^{\text{bbox} \cap \ell_p} = \mathcal{C}^{\text{bbox}} \cap \mathcal{C}^{\ell_p}.$$

6. $\mathcal{C}^{\text{port}} \subseteq \mathcal{C}^{\text{bbox} \cap \ell_p}$.

   Since $\mathcal{C}^{\text{port}} \subseteq \mathcal{C}^{\text{bbox}}$ and $\mathcal{C}^{\text{port}} \subseteq \mathcal{C}^{\ell_p}$, it follows that

$$\mathcal{C}^{\text{port}} \subseteq \mathcal{C}^{\text{bbox}} \cap \mathcal{C}^{\ell_p} = \mathcal{C}^{\text{bbox} \cap \ell_p}.$$

□

C Useful Lemmas

C.1 Quasiconvexity of $\tilde{k}$

Lemma C.1. *$\tilde{k}(x) = \frac{|x-a|^p}{b^p} - c$ is quasiconvex on $\mathbb{R}$ for any $p \geq 1$, $b > 0$.*

Proof. For $\gamma \in \mathbb{R}$ the sublevel set is

$$\{x : \tilde{k}(x) \leq \gamma\} = \{x : |x-a|^p \leq |b|^p(\gamma + c)\}.$$

If $\gamma + c < 0$ it is empty; if $\gamma + c = 0$ it is $\{a\}$; if $\gamma + c > 0$ it is the interval

$$[a - |b|(\gamma + c)^{1/p},\ a + |b|(\gamma + c)^{1/p}].$$

Hence, every sublevel set is convex and $\tilde{k}$ is quasiconvex □

C.2 Continuity of $k(x)$

Lemma C.2. *Assume the constants $\theta \in [0,1]$, $c \in \mathbb{R}_{>0}$ and the series of fixed constants $\tilde{\lambda}_i \in (0, \infty)$ and $x_i \in [0,1]$ for $i \leq t$. Define, for $m \in [0,1]$,*

$$\lambda_i^+(m) := \min\left(\tilde{\lambda}_i\,;\, \frac{c}{m}\right), \qquad \lambda_i^-(m) := \min\left(\tilde{\lambda}_i\,;\, \frac{c}{1-m}\right),$$

with the boundary conventions $\lambda_i^+(0) := \tilde{\lambda}_i$ and $\lambda_i^-(1) := \tilde{\lambda}_i$. Then the following three functions are continuous on $[0,1]$:

$$k_t^+(m) := \prod_{i=1}^t \big(1 + \lambda_i^+(m)(x_i - m)\big), \qquad k_t^-(m) := \prod_{i=1}^t \big(1 - \lambda_i^-(m)(x_i - m)\big),$$

and

$$k_t(m) := \max\big(\theta k_t^+(m),\ (1-\theta) k_t^-(m)\big).$$

Proof. 1. For each i, the function $m \mapsto c/m$ is continuous on $(0, 1]$ and diverges to $+\infty$ as $m \to 0^+$. The function $\lambda_i^+(m) = \min(\tilde{\lambda}_i\,; c/m)$ is therefore continuous on $(0, 1]$ and has the limit $\tilde{\lambda}_i$ as $m \to 0^+$. So by defining $\lambda_i^+(0) := \tilde{\lambda}_i$ we obtain a function continuous on the whole $[0, 1]$. The same argument (with $c/1 - m$ and the convention at $m = 1$) shows $\lambda_i^-(\cdot)$ is continuous on $[0, 1]$.

2. For each i the function $m \mapsto x_i - m$ is continuous and $\lambda_i^+(m)$ is continuous, hence their product $\lambda_i^+(m)(x_i - m)$ is continuous. Adding the constant 1 gives the continuous factor
$$f_i^+(m) := 1 + \lambda_i^+(m)(x_i - m).$$
Analogously, define $f_i^-(m) := 1 - \lambda_i^-(m)(x_i - m)$; it is continuous on $[0, 1]$.

3. Each $k_t^+(m)$ is a finite product of continuous factors $f_i^+(m)$, therefore continuous on $[0, 1]$. The same holds for $k_t^-(m)$.

4. Finally, the function $m \mapsto \max(\theta k_t^+(m),\, (1 - \theta)k_t^-(m))$ is the pointwise maximum of two continuous functions and is therefore continuous. This proves the claim.

□

D Bounding the Global Minimum of $k(x)$

The construction of $\mathcal{C}^{\text{bbox}}$ in § 4.1 requires knowledge of the global minimum of each function $k(x) = \max\{\theta k^+(x),\, (1 - \theta)k^-(x)\}$. In this section, we introduce an algorithm to approximate this minimum and provide explicit guarantees on its correctness. Specifically, given an estimate $\tilde{k}_\star$, the algorithm produces bounds $[\underline{k}^\star, \overline{k}_\star]$ that are guaranteed to contain the true minimum $k_\star$. Note that we assume continuity of k^+ and k^- which follows from Lem. C.2.

This discussion is also used later to certify the discretization error in the ℓ_p-ellipsoid construction.

Lemma D.1. *Let $[a, b] \subseteq [x_{start}, x_{end}]$ with $a < b$, and define*
$$k(x) = \max\left(\theta k^+(x),\, (1 - \theta)k^-(x)\right),$$
where k^+ and k^- are continuous on $[x_{start}, x_{end}]$, with k^+ non-increasing and k^- non-decreasing, and $\theta \in [0, 1]$. Let $x^\star$ be a global minimizer of k on $[a, b]$. Then

- *if θk^+ and $(1 - \theta)k^-$ meet inside $[a, b]$ (i.e., there exists $x \in [a, b]$ with $\theta k^+(x) = (1 - \theta)k^-(x)$), then*
$$\underbrace{\max\left(\theta k^+(b)\,, (1 - \theta)k^-(a)\right)}_{\text{lower bound } \underline{k}_\star} \le k(x^\star) \le \underbrace{\min\left(\theta k^+(a), (1 - \theta)k^-(b)\right)}_{\text{upper bound } \overline{k_\star}},$$

- *if θk^+ and $(1 - \theta)k^-$ do not meet on $[a, b]$, then the global minimum occurs at an endpoint and*
$$k(x^\star) = \min\left(k(a), k(b)\right).$$

Proof. Define $h(x) := (1 - \theta)k^-(x) - \theta k^+(x)$. Because k^- is non-decreasing and k^+ is non-increasing, the function $-k^+$ is non-decreasing, so h is non-decreasing on $[a, b]$. We distinguish three cases, based on the sign of $h(x)$ on $[a, b]$:

1. $h(x) > 0$ for all $x \in [a, b]$,
2. $h(x) < 0$ for all $x \in [a, b]$,
3. $h(x) = 0$ for some $x \in [a, b]$.

Note that in case 3, because of the continuity of h a root does indeed exist. Because of its monotonicity, this is either a single point or an interval.

We treat the cases.

1. Case A - No intersection on $[a, b]$ (cases 1 and 2).
   If $h > 0$ on $[a, b]$ then $(1-\theta)k^-(x) > \theta k^+(x)$ for every $x \in [a, b]$, so $k(x) = (1-\theta)k^-(x)$ on $[a, b]$. Since k^- is non-decreasing, k is non-decreasing and thus attains its minimum at $x = a$. Hence $k(x^\star) = k(a) = \min(k(a), k(b))$. Analogously, if $h < 0$ on $[a, b]$ then $k(x) = \theta k^+(x)$ on $[a, b]$. Since θk^+ is non-increasing, k is non-increasing and attains its minimum at $x = b$. Hence $k(x^\star) = k(b) = \min(k(a), k(b))$. This proves the endpoint statement.

2. Case B - Let $h(x) = 0$ have a solution (case 3).
   We show the bound in two steps: 1) we show that the roots of $h(x) = (1-\theta)k^-(x) - \theta k^+(x)$ correspond to the global minima of $k(x)$. 2) From this intermediate result, we can derive a lower and upper bound.

   Let $x_0 \in [a, b]$ s.t. $h(x_0) = 0$, we thus have that $(1-\theta)k^-(x_0) = \theta k^+(x_0) = k(x_0)$. But since $k(x) = \max(\theta k^+(x), (1-\theta)k^-(x))$, we have:

   - for $x \leq x_0 : k(x) = \theta k^+(x)$ (since for $x \leq x_0, h(x) \leq 0 \Rightarrow (1-\theta)k^-(x) \leq \theta k^+(x)$.)
   - for $x \geq x_0 : k(x) = (1-\theta)k^-(x)$.

   But if $k(x)$ is non-increasing to the left of x_0, and non-decreasing to the right of x_0, surely, x_0 must be a global minimizer of $k(x)$.

   Now the bounds:

   - Upper bound.
     Since we just derived that the roots of $h(x)$ are global minimizers for $k(x)$, pick an $x^\star$ from this zero set. It then follows that $k(x^\star) = (1-\theta)k^-(x^\star) = \theta k^+(x^\star)$. Now, from the non-decreasing property of $(1-\theta)k^-$, we get $(1-\theta)k^-(x^\star) \leq (1-\theta)k^-(b)$, and analogously for k, we get $\theta k^+(x^\star) \leq \theta k^+(a)$. Combining these two inequalities, together with the equality at $k(x^\star)$, gives us the upper bound:

     $$k(x^\star) \leq \min\left((1-\theta)k^-(b), \theta k^+(a)\right).$$

     Note that for any global optimizer $x_2^\star \in [a, b]$ lying outside the zero set of $h(x)$, the bound still applies, since by definition of a global optimizer we have $k(x^\star) = k(x_2^\star)$.
   - Lower bound.
     Because k^- is non-decreasing and $a \leq x^\star$, we have $(1-\theta)k^-(a) \leq (1-\theta)k^-(x^\star)$. Because k^+ is non-increasing and $x^\star \leq b$ we have $\theta k^+(b) \leq k(x^\star)$. Finally, taking the maximum gives us our bound:

     $$\max\left((1-\theta)k^-(a), \theta k^+(b)\right) \leq \max\left((1-\theta)k^-(x^\star), \theta k^+(x^\star)\right) = k(x^\star).$$

□

The theorem implies two practical consequences. First, a conservative estimate $\underline{k}^\star$ of $k(x^\star)$ can be computed via a root-finding procedure applied to $h(x)$; since h is non-decreasing, its roots form an interval. Second, the approximation error on this minimum can be explicitly controlled using the derived bounds. Algorithm ESTIMATE-$k_\star$ (Alg. 1) describes this procedure in detail.

Part II: Outer Approximations and Practical Details

E Bounding box

We derive the construction of the axis-aligned bounding box $\mathcal{C}^{\text{bbox}}$ of $\mathcal{C}^{\text{port}}$. This is achieved by projecting $\mathcal{C}^{\text{port}}$ onto each coordinate i and deriving its interval $[l_i, u_i]$. To compute it, write

$$\sum_{d=1}^{D} w_d k^d(x_d) \leq \frac{1}{\alpha} \iff w_i k^i(x_i) + \sum_{d \neq i} w_d k^d(x_d) \leq \frac{1}{\alpha},$$

Algorithm 1 ESTIMATE-$k_\star$

Input: functions $k^+(\cdot), k^-(\cdot)$ on $[x_{\text{start}}, x_{\text{end}}]$; $\theta \in [0,1]$; tolerances ϵ_x, ϵ_y
Output: $(\underline{k_\star}, \overline{k_\star}, \hat{x})$
$a \leftarrow x_{\text{start}}, \quad b \leftarrow x_{\text{end}}$
$h(x) \leftarrow (1-\theta)k^-(x) - \theta k^+(x)$
if $h(a) > 0$ **or** $h(b) < 0$ **then**
  $k(x) = \max\{\theta k^+(x), (1-\theta)k^-(x)\}$
  $x_\star \leftarrow \arg\min_{x \in \{a,b\}} k(x)$
  $k_\star \leftarrow k(x_\star)$
  return $(k_\star, k_\star, x_\star)$
end if
$\underline{k_\star} \leftarrow -\infty, \quad \overline{k_\star} \leftarrow \infty$
while $(b - a > \epsilon_x)$ **or** $(\overline{k_\star} - \underline{k_\star} > \epsilon_y)$ **do**
  $\underline{k_\star} \leftarrow \max\big((1-\theta)k^-(a),\, \theta k^+(b)\big), \quad \overline{k_\star} \leftarrow \min\big((1-\theta)k^-(b),\, \theta k^+(a)\big)$
  $m \leftarrow \frac{1}{2}(a+b)$
  if $h(m) < 0$ **then**
    $a \leftarrow m$
  else if $h(m) > 0$ **then**
    $b \leftarrow m$
  else
    $x_\star \leftarrow m, \quad k_\star \leftarrow k(m)$
    return $(k_\star, k_\star, x_\star)$
  end if
end while
$\underline{k_\star} \leftarrow \max\big((1-\theta)k^-(a),\, \theta k^+(b)\big), \quad \overline{k_\star} \leftarrow \min\big((1-\theta)k^-(b),\, \theta k^+(a)\big)$
$\hat{x} \leftarrow \frac{1}{2}(a+b)$
return $(\underline{k_\star}, \overline{k_\star}, \hat{x})$

such that isolating $k^i(x_i)$ leads to

$$w_i k^i(x_i) \leq \frac{1}{\alpha} - \sum_{d \neq i} w_d k^d(x_d).$$

Remark that maximizing the right-hand side gives the largest admissible level for $k^i(\cdot)$, which is achieved when $\sum_{d \neq i} w_d k^d(x_d)$ is minimized. Denote the global minimum for each $k^d(\cdot)$ as $k_\star^d \in \mathbb{R}_{\geq 0}$, then the largest level set is achieved by the constant: $\frac{1}{\alpha} - \sum_{d \neq i} w_d k_\star^d$. Now, because $k^i(\cdot)$ is quasiconvex, its sublevel set is a (possibly degenerate) interval, and the endpoints l_i, u_i are therefore obtained by the outer two solutions of

$$w_i k^i(x_i) = \frac{1}{\alpha} - \sum_{d \neq i} w_d k_\star^d. \tag{16}$$

Under the quasiconvexity alone, $k^i(\cdot)$ could in principle contain flat segments, such that more than two x_i exist satisfying Eq. (16). As such, a safe choice is to solve for the equation:

$$w_i k^i(x_i) = \frac{1}{\alpha} - \sum_{d \neq i} w_d k_\star^d + \epsilon, \tag{17}$$

where ϵ is a small positive constant (e.g., 1×10^{-12}) ensuring that any two solutions found are guaranteed to contain the interval of Eq. (16).

In practice, the two solutions to Eq. (17) can be computed numerically, for example via two bisection searches applied to

$$w_i k^i(x_i) - \left(\frac{1}{\alpha} - \sum_{d \neq i} w_d k_\star^d \right) - \epsilon = 0,$$

performed on intervals on either side of a minimizer of $k^i(\cdot)$. In our implementation, we use the minimizer $x_i^\star$ (corresponding to the global minimum $k_\star^i$) as the splitting point. All global minima are obtained using Alg. 1 and selecting the theoretically "worst case" minimum. Repeating this construction for all coordinates results in the bounding box

$$\mathcal{C}^{\text{bbox}} = \prod_{d=1}^{D} [l_d, u_d], \qquad \mathcal{C}^{\text{port}} \subseteq \mathcal{C}^{\text{bbox}}.$$

F ℓ_p-Ellipsoid

F.1 ℓ_p-Ellipsoid Derivation

We derive the ℓ_p-ellipsoid representation for the set $\mathcal{C}^{\ell_p}$.

$$\sum_{d=1}^{D} w_d \left(\frac{|x_d - a_d|^p}{b_d^p} - c_d \right) \;\leq\; \frac{1}{\alpha}. \tag{18}$$

Rearranging gives

$$\sum_{d=1}^{D} w_d \, \frac{|x_d - a_d|^p}{b_d^p} \;\leq\; \underbrace{\frac{1}{\alpha} + \sum_{d=1}^{D} w_d \, c_d}_{=:\, R}, \tag{19}$$

where $R > 0$ is required for the set to be nonempty. Noting that

$$w_d \, \frac{|x_d - a_d|^p}{b_d^p} = \frac{|x_d - a_d|^p}{(b_d / \sqrt[p]{w_d})^p}, \tag{20}$$

we can rewrite the inequality as

$$\sum_{d=1}^{D} \frac{|x_d - a_d|^p}{(b_d / \sqrt[p]{w_d})^p} \;\leq\; R. \tag{21}$$

Dividing through R results in

$$\sum_{d=1}^{D} \frac{|x_d - a_d|^p}{\left(b_d \sqrt[p]{R} / \sqrt[p]{w_d} \right)^p} \;\leq\; 1. \tag{22}$$

We therefore identify the semi-axis lengths

$$b'_d := \frac{b_d \sqrt[p]{R}}{\sqrt[p]{w_d}}, \quad R = \frac{1}{\alpha} + \sum_{j=1}^{D} w_j \, c_j, \tag{23}$$

and the ℓ_p-ellipsoid can be written as

$$\sum_{d=1}^{D} \frac{|x_d - a_d|^p}{(b'_d)^p} \;\leq\; 1. \tag{24}$$

The result is an axis-aligned ℓ_p-ellipsoid, and can also be written in matrix form as follows:

$$\|B^{-1}(\mathbf{x} - \mathbf{a})\|_p \leq 1,$$

with

$$B := \operatorname{diag}\big(b'_1, \ldots, b'_D\big) \tag{25}$$

with $\mathbf{a} \in \mathbb{R}^D$, consisting of the elements $a_1 \ldots a_D$.

F.2 Fitting $\tilde{k}(x)$ from Two Touch Points

Given a threshold $\gamma > k_\star$, we are given two distinct points $x_L^{\text{touch}}, x_R^{\text{touch}}$ that lie on the left and right of the global optimizer $x^\star$ so that:

$$k(x_L^{\text{touch}}) = k(x_R^{\text{touch}}) = \gamma + \epsilon,$$

for a small positive constant ϵ. The motivation for this small constant is discussed in App. F.4, but it generally also helps with the feasibility of $\tilde{k}$.

For a fixed integer $p \geq 1$, we seek parameters (a, b, c) such that

$$\tilde{k}(x_L^{\text{touch}}) = \tilde{k}(x_R^{\text{touch}}) = \gamma \qquad \text{and} \qquad \forall x \in [x_{\text{start}}, x_{\text{end}}] : \ \tilde{k}(x) \leq k(x),$$

where

$$\tilde{k}(x) = \frac{|x-a|^p}{b^p} - c$$

with $b > 0$. Note that once a, b, and c are obtained, p can be calculated by considering a series of candidates over a small fixed grid of integers (e.g., $\{6 \dots 16\}$) and choosing the p that minimizes the volume of the respective ℓ_p-ellipsoid.

Deriving (a, b c)

1. **Finding a.** From $\tilde{k}(x_L^{\text{touch}}) = \tilde{k}(x_R^{\text{touch}}) = \gamma$ we have

   $$\frac{|x_L^{\text{touch}} - a|^p}{b^p} - c = \gamma, \qquad \frac{|x_R^{\text{touch}} - a|^p}{b^p} - c = \gamma.$$

   Subtracting the two equations results in

   $$|x_L^{\text{touch}} - a|^p = |x_R^{\text{touch}} - a|^p \Rightarrow |x_L^{\text{touch}} - a| = |x_R^{\text{touch}} - a|.$$

2. And since, $x_L^{\text{touch}} \neq x_R^{\text{touch}}$ we obtain

   $$a = \frac{x_L^{\text{touch}} + x_R^{\text{touch}}}{2},$$

   i.e., a is the midpoint of x_L^{touch} and x_R^{touch}.

3. **Finding c (in terms of b).** Plugging one touching point (say x_L^{touch}) into $\tilde{k}(x_L^{\text{touch}}) = \gamma$ gives

   $$\frac{|x_L^{\text{touch}} - a|^p}{b^p} - c = \gamma \quad \Rightarrow \quad c = \frac{|x_L^{\text{touch}} - a|^p}{b^p} - \gamma. \tag{26}$$

   Thus once b is chosen, c is also determined.

4. **Finding b.** Substituting c from (26) into $\tilde{k}(x)$ gives

   $$\tilde{k}(x) = \frac{|x-a|^p}{b^p} - \frac{|x_L^{\text{touch}} - a|^p}{b^p} + \gamma = \frac{|x-a|^p - |x_L^{\text{touch}} - a|^p}{b^p} + \gamma.$$

   The requirement $\tilde{k}(x) \leq k(x)$ for all $x \in [x_{\text{start}}, x_{\text{end}}]$ is therefore

   $$\frac{|x-a|^p - |x_L^{\text{touch}} - a|^p}{b^p} + \gamma \leq k(x) \quad \Longleftrightarrow \quad |x-a|^p - |x_L^{\text{touch}} - a|^p \leq (k(x) - \gamma)\, b^p.$$

   Define the function

   $$D(x) := k(x) - \gamma.$$

   We must satisfy, for every x in the domain, the inequality

   $$|x-a|^p - |x_L^{\text{touch}} - a|^p \leq D(x)\, b^p.$$

   Now, consider three separate cases depending on the sign of $D(x)$:

- If $D(x) = 0$ (i.e. $k(x) = \gamma$), the constraint reduces to

$$|x-a|^p - |x_L^{\text{touch}} - a|^p \le 0 \quad \Longleftrightarrow \quad |x-a| \le |x_L^{\text{touch}} - a|.$$

  Thus, every point x where $k(x) = \gamma$ must lie within the closed interval around a of radius $|x_L^{\text{touch}} - a|$. If there exists an x with $k(x) = \gamma$ but $|x-a| > |x_L^{\text{touch}} - a|$, no $\tilde{k}$ of the chosen form can simultaneously touch at $x_L^{\text{touch}}, x_R^{\text{touch}}$ and remain below k at that x.
- If $D(x) > 0$ then dividing by $D(x)$ (positive) gives the lower bound

$$\frac{|x-a|^p - |x_L^{\text{touch}} - a|^p}{D(x)} \le b^p.$$

- If $D(x) < 0$ then dividing by $D(x)$ (negative) reverses the inequality, resulting in the upper bound

$$\frac{|x-a|^p - |x_L^{\text{touch}} - a|^p}{D(x)} \ge b^p.$$

For easier notation, we define this fraction as $R(x) := \frac{|x-a|^p - |x_L^{\text{touch}} - a|^p}{D(x)}$ (and denote it as $\underline{R}$ when it defines a lower bound and $\overline{R}$ otherwise). A feasible b must therefore satisfy

$$b^p \ge \sup_{x:\, D(x)>0} \underline{R}(x) \qquad \text{and} \qquad b^p \le \inf_{x:\, D(x)<0} \overline{R}(x), \tag{27}$$

while every x with $D(x) = 0$ must satisfy $|x-a| \le |x_L^{\text{touch}} - a|$. If $\sup_{x:\, D(x)>0} \underline{R}(x) > \inf_{x:\, D(x)<0} \overline{R}(x)$, then the feasible interval for b^p is empty and no $\tilde{k}$ exists that meets the inequality on the whole interval (under the given p).

Practical remark. Numerically, one evaluates the functions on a fine grid x_i with $i = 1, \dots, n$, covering $[x_{\text{start}}, x_{\text{end}}]$. For example choose $x_i = x_{\text{start}} + (i-1)\,h, \qquad h = \frac{x_{\text{end}} - x_{\text{start}}}{n-1}$. Define

$$\underline{b^p} = \max_{i \in \{1 \dots n \,|\, D(x_i) > 0\}} \underline{R}(x_i), \qquad \overline{b^p} = \min_{i \in \{1 \dots n \,|\, D(x_i) < 0\}} \overline{R}(x_i).$$

When the interval $[\max\{0, \underline{b^p}\}, \overline{b^p}] \neq \emptyset$ and all i with $|D(x_i)| \le \epsilon$ satisfy $|x_i - a| \le |x_L^{\text{touch}} - a|$, a solution exists, and a simple choice for b^p is to take the midpoint:

$$b^p \leftarrow \frac{\underline{b^p} + \overline{b^p}}{2}, \qquad b \leftarrow \left(b^p\right)^{1/p}.$$

Finally obtain c from (26).

Choosing γ The parameter γ controls which areas of $k(x)$ receive the tightest fit. We generally found that a simple rule, which depends on the dimension D and the original sublevel-set's threshold $1/\alpha$, seemed to work well in practice. We suggest setting it to: $\gamma_d = \frac{\sqrt[p]{w_d}}{\alpha}$ as a solid default choice. When all $w_1 \dots w_D$ are equal, this corresponds to setting each $\gamma_d = \frac{1}{\alpha \sqrt[p]{D}}$.

Adaptive widening when the touch constraints are infeasible When the touch-point equalities at $(x_L^{\text{touch}}, x_R^{\text{touch}})$ do not have a solution, we relax the tightness by widening the touch points:

$$x_L^{\text{touch}} \leftarrow \max(x_L^{\text{touch}} - \delta,\ x_{\text{start}}), \quad x_R^{\text{touch}} \leftarrow \min(x_R^{\text{touch}} + \delta,\ x_{\text{end}}), \tag{28}$$

and re-run the feasibility test. Since sufficiently large δ eventually makes the constraints feasible in practice, we find the smallest feasible δ by bisection on $\delta \in [0,\ x_{\text{end}} - x_{\text{start}}]$. Alg. 3 implements this search. The algorithm repeatedly splits the interval into two halves, searching for the transition point from no solution to a solution. This procedure is repeated until the resulting interval has reduced to a predefined tolerance. The parameter γ remains fixed.

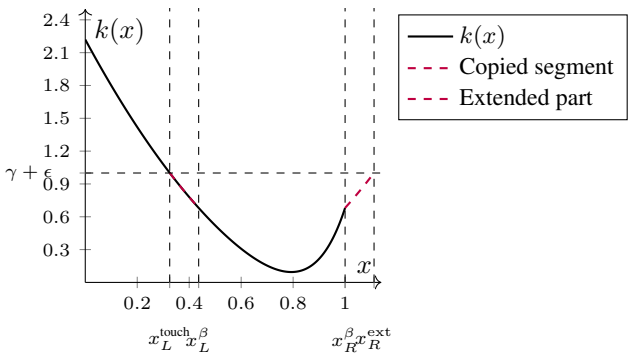


Figure 7: Right part extension.

Domain extension to guarantee two touch points For a continuous quasiconvex k on $[0,1]$, the equation $k(x) = \gamma + \epsilon$ may have only a solution on one side of $x^\star$. As such, the procedure from App. F.2 will fail as either x_L^{touch} or x_R^{touch} does not exist. Here, we construct an extended function k_{ext} by mirroring a segment of k, creating a second touch point at the level $\gamma + \epsilon$ while preserving quasiconvexity and ensuring $k_{\text{ext}}(x) = k(x)$ for all $x \in [0,1]$ (so all set-containment guarantees on the original domain are unchanged).

- **Right extension (missing right touch point).** Assume k crosses $\gamma + \epsilon$ on the left of $x^\star$ but not on the right. I.e., there exists $x_L^{\text{touch}} \in [0, x^\star]$ (so that $k(x_L^{\text{touch}}) = \gamma + \epsilon$) but no $x_L^{\text{touch}} \in [x^\star, 1]$. We additionally assume that k is not monotone on $[0,1]$; the monotone case is discussed separately below.

  Let $x_{\beta,R} := 1$ and define $\beta := k(x_{\beta,R}) < \gamma + \epsilon$. Since k is continuous and quasiconvex but not monotone, there exists $x_{\beta,L} \in (x_L^{\text{touch}}, x^\star)$ with $k(x_{\beta,L}) = \beta$.

  Define the extended endpoint $x_R^{\text{ext}} := x_{\beta,R} + (x_{\beta,L} - x_L^{\text{touch}})$ and

$$k_{\text{ext}}(x) := \begin{cases} k(x), & x \le x_{\beta,R}, \\ k(-x + x_{\beta,R} + x_{\beta,L}), & x > x_{\beta,R}. \end{cases}$$

  Then $k_{\text{ext}}(x_R^{\text{ext}}) = \gamma + \epsilon$ and k_{ext} is quasiconvex. The process is visualized in Fig. 7.

- **Left extension (missing left touch point).** If instead there exists x_R^{touch} but no x_L^{touch}, the above procedure is applied symmetrically by reflecting a segment of k from the right side onto the left. The construction is entirely analogous.
- **Edge case:** Finally, we consider the case where $k(x) = \max\{\theta k^+(x), (1-\theta)k^-(x)\}$ is non-decreasing or non-increasing over its entire domain (with k^+ non-increasing and k^- non-decreasing on $[0,1]$). To check whether these two functions ever intersect, define $h(x) = (1-\theta)k^-(x) - \theta k^+(x)$, which is non-decreasing. Hence, the curves do not intersect on $[0,1]$ iff either $h(0) > 0$ (so $(1-\theta)k^-(x) > \theta k^+(x)$ for all x) or $h(1) < 0$ (so $(1-\theta)k^-(x) < \theta k^+(x)$ for all x). In this "no-intersection" case, k is monotone on $[0,1]$. We then extend the domain by mirroring: if k is non-decreasing, mirror $[0,1]$ to the left to obtain $[-1,1]$; otherwise mirror it to the right to obtain $[0,2]$.

F.3 Algorithms

The algorithm to construct the l_p-ellipsoid is presented in FINDELLIPSOID (Alg. 2). It iterates over a candidate set $\mathcal{P}$, and for each $p \in \mathcal{P}$ computes an l_p-ellipsoid by fitting D functions $\tilde{k}^d(\cdot)$ through FITPARAMETERS-$\tilde{k}$ (Alg. 3). This algorithm performs a bisection search to find two valid touch points, and in turn, uses the FITPARAMSATTOUCH (Alg. 4) to fit the actual $\tilde{k}(x)$ given the required touching points.

The FINDINTERVAL call (used by FINDELLIPSOID), given a threshold γ, returns the two distinct points (a and b) where k approximately touches γ, i.e., $k(a) \approx \gamma$. This can be achieved through two root-finding calls (e.g., bisection), on $f(x) = k(x) - \gamma$, one call on the interval $[a, \hat{x})$ and the second

Algorithm 2 FINDELLIPSOID (grid-search over p)

Input: Per-dimension functions $\{k^d(\cdot)\}_{d=1}^D$, weights $\mathbf{w} \in \Delta^D$, error level α, candidate set $\mathcal{P}$, grid size n, ϵ, max_iters, tol
Output: Best parameters $\{(a_d, b_d, c_d)\}_{d=1}^D$ and best $p \in \mathcal{P}$
best_params $\leftarrow \varnothing$, best_p $\leftarrow \varnothing$, best_vol $\leftarrow +\infty$
for each $p \in \mathcal{P}$ **do**
  current_params $\leftarrow$ empty list
  feasible $\leftarrow$ TRUE
  for $d = 1$ **to** D **do**
    $\gamma_d \leftarrow \frac{\sqrt[p]{w_d}}{\alpha}$
    $(k^d, x^d_{\text{start}}, x^d_{\text{end}}) \leftarrow \text{EXTEND}(k^d, \gamma_d + \epsilon)$
    $(x^d_{\text{start}}, x^d_{\text{end}}) \leftarrow \text{FINDINTERVAL}(k^d, \text{target} = \frac{1}{\alpha w_d} + \epsilon;\ x^d_{\text{start}}, x^d_{\text{end}})$
    $S^d \leftarrow \{(x_j, k^d(x_j))\}_{j=1}^n$ {x_j uniform on $[x^d_{\text{start}}, x^d_{\text{end}}]$}
    $\gamma_d = \min\{\gamma_d, k^d(x^d_{\text{start}}), k^d(x^d_{\text{end}})\}$
    $(x^{\text{touch}}_{L,d}, x^{\text{touch}}_{R,d}) \leftarrow \text{FINDINTERVAL}(k^d, \gamma_d + \epsilon;\ x^d_{\text{start}}, x^d_{\text{end}})$ {ϵ comes from App. F.4}
    params$_d \leftarrow$ FITPARAMETERS-$\tilde{k}(S^d, x^{\text{touch}}_{L,d}, x^{\text{touch}}_{R,d}, [x^d_{\text{start}}, x^d_{\text{end}}], p, \gamma_d, \epsilon, \text{max_iters}, \text{tol})$
    if params$_d$ is NO SOLUTION **then**
      feasible $\leftarrow$ FALSE
      break
    end if
    append params$_d$ to current_params
  end for
  if feasible = FALSE **then**
    continue
  end if
  vol $\leftarrow \text{ELLIPSOIDVOLUME}(\{(a_d, b_d, c_d)\}_{d=1}^D,\ \mathbf{w},\ \alpha,\ p)$
  if vol $<$ best_vol **then**
    best_vol $\leftarrow$ vol
    best_params $\leftarrow$ current_params
    best_p $\leftarrow p$
  end if
end for
return (best_params, best_p)

on $(\hat{x}, b]$. Here, $\hat{x}$ corresponds to an estimated global minimizer of $k(x)$, obtained from ESTIMATE-$k_\star$ (Alg. 1).

Computational complexity. FITPARAMSATTOUCH: Given grid evaluations S^d and a pair of touch points $(x^{\text{touch}}_{L,d}, x^{\text{touch}}_{R,d})$, the algorithm performs a single pass over the grid and thus runs in $O(N)$ time.

FITPARAMETERS-$\tilde{k}$: repeatedly calls FITPARAMSATTOUCH while performing a bisection search over the touch-distance parameter δ on an interval of length $L := \delta_{\text{high}} - \delta_{\text{low}} \le x_{\text{end}} - x_{\text{start}}$. With termination tolerance `tol`, the number of bisection iterations is $T_{\text{bisect}} = O\big(\log \frac{L}{\texttt{tol}}\big)$. Since $L \le 1$ when $k(x)$ was not extended (since its domain is $[0, 1]$), or $L \le 2$ when k-extension is enabled (as in the worst case, the entire domain is copied). Consequently, the dominant cost of fitting a single 1D $\tilde{k}^d$ is

$$O\left(N \log \frac{L}{\texttt{tol}}\right) = O\left(N \log \frac{1}{\texttt{tol}}\right)$$

operations on precomputed samples S.

FINDELLIPSOID: finally, FINDELLIPSOID repeats this procedure for each coordinate $d \in \{1, \dots, D\}$ and each candidate p in a set $\mathcal{P}$ of size N_p. Additionally, for each d and p the grid samples S^d are computed, with each sample's cost denoted as C_{eval} (i.e., the cost of a single evaluation of $k^d(\cdot)$). Note that S^d depends on EXTEND which requires $\gamma_d = \frac{\sqrt[p]{w_d}}{\alpha}$ (Alg. 2). This is why S^d is recomputed for the various p, but there are certainly more efficient alternatives which we do not cover here.

Algorithm 3 FITPARAMETERS-$\tilde{k}$ (binary search over touch-distance)

Input: Samples $S = \{(x_i, k(x_i))\}_{i=1}^n$, initial touch points x_L^{touch}, x_R^{touch}, domain $[x_{\text{start}}, x_{\text{end}}]$, p, γ, ϵ, max_iters, tol
Output: (a, b, c) or NO SOLUTION
params $\leftarrow$ FITPARAMSATTOUCH$(S, x_L^{\text{touch}}, x_R^{\text{touch}}, p, \gamma, \epsilon)$
if params exist **then**
  return params
end if
$\delta_{\text{low}} \leftarrow 0$, $\delta_{\text{high}} \leftarrow (x_{\text{end}} - x_{\text{start}})$
best_params $\leftarrow \varnothing$
for $t = 1$ **to** max_iters **do**
  $\delta_{\text{mid}} \leftarrow 0.5(\delta_{\text{low}} + \delta_{\text{high}})$
  $x_L^{\text{test}} \leftarrow \max(x_L^{\text{touch}} - \delta_{\text{mid}},\ x_{\text{start}})$
  $x_R^{\text{test}} \leftarrow \min(x_R^{\text{touch}} + \delta_{\text{mid}},\ x_{\text{end}})$
  params $\leftarrow$ FITPARAMSATTOUCH$(S, x_L^{\text{test}}, x_R^{\text{test}}, p, \gamma, \epsilon)$
  if params exist **then**
    $\delta_{\text{high}} \leftarrow \delta_{\text{mid}}$
    best_params $\leftarrow$ params
  else
    $\delta_{\text{low}} \leftarrow \delta_{\text{mid}}$
  end if
  if $\delta_{\text{high}} - \delta_{\text{low}} \leq$ tol **and** best_params $\neq \varnothing$ **then**
    return best_params
  end if
end for
return NO SOLUTION

Returning to FINDELLIPSOID's complexity analysis, we obtain a total of

$$O\left(D\, N_p\, N(\log \frac{L}{\texttt{tol}} + C_{\text{eval}})\right)$$

where we have ignored the comparatively small overhead of FINDINTERVAL and EXTEND.

F.4 Analysis of $R(x)$ when dividing by zero

The construction in App. F.2 obtains the parameters (a, b, c) such that $k(x_i) \geq \tilde{k}(x_i)$ holds for all points x_i. Recall that the feasibility condition for b^p is expressed in terms of

$$R(x) := \frac{|x - a|^p - |x_L^{\text{touch}} - a|^p}{D(x)}, \qquad D(x) := k(x) - \gamma.$$

Points with $D(x) > 0$ impose a lower bound on b^p, while points with $D(x) < 0$ impose an upper bound. When $D(x) = 0$, the constraint reduces to

$$|x - a| \leq |x_L^{\text{touch}} - a|.$$

In practice, testing the condition $D(x) = 0$ exactly is numerically unstable. Therefore, the implementation (see Alg. 4) treats all points with $|D(x)| < \epsilon$ as belonging to the $D(x) = 0$ case. Here, we analyze whether this relaxed handling could eliminate critical constraints on b^p and thereby violate the condition $k(x_i) \geq \tilde{k}(x_i)$.

F.4.1 Analytical analysis

Setup. Let $x^\star$ satisfy $D(x^\star) = 0$, i.e., $k(x^\star) = \gamma$ (not to be confused with the global minimizer of $k(x)$). Following the choice for the touch points in App. F.2, we assume that the two touch points $x_L^{\text{touch}}, x_R^{\text{touch}}$ are chosen such that

$$k(x_L^{\text{touch}}) > \gamma, \quad \text{and} \quad k(x_R^{\text{touch}}) > \gamma,$$

such that by the quasiconvexity of k, we obtain the strict inequalities $x_L^{\text{touch}} < x_L^\star$ and $x_R^\star < x_R^{\text{touch}}$. In total, we have:

$$x_{\text{start}} \leq x_L^{\text{touch}} < x_L^\star < a < x_R^\star < x_R^{\text{touch}} \leq x_{\text{end}}. \tag{29}$$

Algorithm 4 FITPARAMSATTOUCH

Input: Samples $S = \{(x_i, k(x_i))\}_{i=1}^n$, x_L^{touch}, x_R^{touch}, p, γ, ϵ
Output: (a, b, c) or NO SOLUTION
$a \leftarrow (x_L^{\text{touch}} + x_R^{\text{touch}})/2$ {Midpoint of touch points}
$\underline{b^p} \leftarrow 0$, $\overline{b^p} \leftarrow +\infty$
for $i = 1$ **to** n **do**
  $N_i \leftarrow |x_i - a|^p - |x_L^{\text{touch}} - a|^p$
  $D_i \leftarrow k(x_i) - \gamma$
  if $|D_i| < \epsilon$ **and** $N_i > 0$ **then**
    return NO SOLUTION
  end if
  if $D_i > 0$ **then**
    $\underline{R_i} \leftarrow N_i / D_i$ {Lower bound}
    $\underline{b^p} \leftarrow \max(\underline{b^p}, \underline{R_i})$
  else
    $\overline{R_i} \leftarrow N_i / D_i$ {Upper bound}
    $\overline{b^p} \leftarrow \min(\overline{b^p}, \overline{R_i})$
  end if
end for
if $\underline{b^p} > \overline{b^p}$ **then**
  return NO SOLUTION
end if
if $\overline{b^p} = \infty$ **then**
  $b^p \leftarrow \underline{b^p}$
else
  $b^p \leftarrow (\underline{b^p} + \overline{b^p})/2$ {Average bounds}
end if
$c \leftarrow |x_L^{\text{touch}} - a|^p / b^p - \gamma$ {Compute parameter c}
return (a, b, c)

Behavior of $R(x)$ near $D(x) = 0$. We focus on the right-side root $x_R^\star$ and analyze the limiting behavior of $R(x)$ as $x \to x_R^\star$. Since $D(x_R^\star) = 0$, $R(x)$ is undefined at this point, so we consider its one-sided limits.

The numerator

$$|x - a|^p - |x_L^{\text{touch}} - a|^p \tag{30}$$

evaluated at $x = x_R^\star$ is strictly negative and non-zero, because $|x_R^\star - a| < |x_L^{\text{touch}} - a|$ by construction (follows from Eq. (29) and the fact that $|x_L^{\text{touch}} - a| = |x_R^{\text{touch}} - a|$). We denote Eq. (30) by $-c$ for some $c > 0$. As $x \to {x_R^\star}^-$ we have $D(x) < 0$, while as $x \to {x_R^\star}^+$ we have $D(x) > 0$. Therefore,

$$\lim_{x \to {x_R^\star}^-} \overline{R}(x) = \lim_{x \to {x_R^\star}^-} \frac{-c}{D(x)} = +\infty, \qquad \lim_{x \to {x_R^\star}^+} \underline{R}(x) = \lim_{x \to {x_R^\star}^+} \frac{-c}{D(x)} = -\infty.$$

Implication for feasibility bounds. Recall that points with $D(x) < 0$ contribute upper bounds on b^p via $\inf_{D(x)<0} \overline{R}(x)$, while points with $D(x) > 0$ contribute lower bounds via $\sup_{D(x)>0} \underline{R}(x)$. The divergent limits above, therefore, do not lead to finite or active constraints:

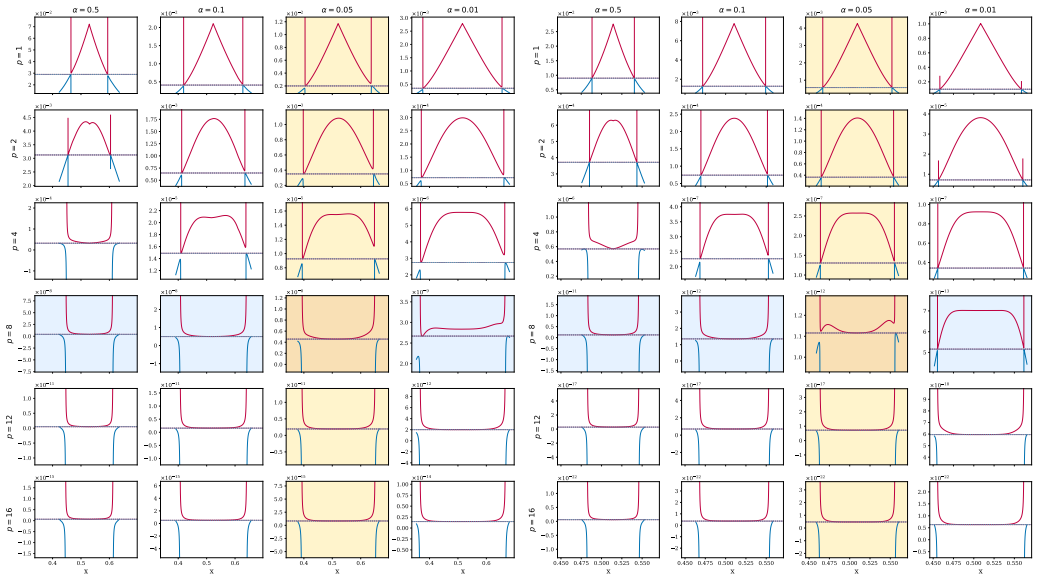
- As $x \to {x_R^\star}^-$, $\overline{R}(x) \to +\infty$, which cannot tighten the upper bound on b^p.
- As $x \to {x_R^\star}^+$, $\underline{R}(x) \to -\infty$, which cannot tighten the lower bound on b^p.

Consequently, for sufficiently small ϵ, treating points with $|D(x)| < \epsilon$ as belonging to the $D(x) = 0$ case does not eliminate any finite, critical bounds on b^p. We also validate this empirically below.

F.4.2 Empirical analysis

Fig. 8 depicts the upper (red) and lower bound (blue) for legal b^p candidates. Based on the equation:

$$R(x) = \frac{|x - a|^p - |x_L^{\text{touch}} - a|^p}{D(x)}$$



(a) 50 samples (b) 500 samples

Figure 8: Plots of $R(x) = \frac{|x-a|^p - |x_1-a|^p}{D(x)}$ for various α and p combinations. The blue part of $R(x)$ corresponds to the lower bound of b^p (i.e., $D(x) > 0$), and the red part to the upper bound (i.e., $D(x) < 0$).

from App. F.2. We find that when $D(x) = k(x) - \gamma$ approaches zero, $R(x)$ diverges either to $\pm\infty$. For the upper bound $R(x)$ approaches $+\infty$, resulting in a meaningless bound on b^p, and similarly, the lower bound is $-\infty$ near $D(x) = 0$. Hence, at the limit, no crucial upper/lower bound values are being ignored by treating $|D(x)| \leq \delta$ as case 1 in Item 4 (from the three items). Interestingly, as p increases, the key boundary points move away from the points where $D(x) = 0$. We also observe that the sub-regions containing these crucial boundary points appear to be somewhat convex, suggesting that future algorithms could exploit this structure to derive upper and lower bounds for b more efficiently than the uniform grid evaluation used here.

Details to reproduce the figure: Underlying data distribution: $y \sim \mathcal{U}(0.5, 0.8)$, where $k(x)$ is based on either 50 or 500 observations. The grid size N is set to 1,000,001 pts.

F.5 Side Note - Function Approximation with Two Distinct Touch Levels

So far, we assumed that the level γ is touched on both sides of the global minimizer of k, but this does not always hold in practice, motivating the extended construction k_{ext} introduced in App. F.2.

An alternative approach is to directly work with a more general formulation in which the two sides are allowed to correspond to different levels. Specifically, we consider two points x_L^{touch} and x_R^{touch} satisfying

$$k(x_L^{\text{touch}}) = \alpha, \qquad k(x_R^{\text{touch}}) = \beta,$$

with α and β not necessarily equal, and construct an approximation $\tilde{k}$ that touches k at these locations.

We derive this construction for the quadratic case

$$\tilde{k}(x) = \frac{(x-a)^2}{b^2} - c,$$

corresponding to $p = 2$. For $p > 2$, the resulting parameter equations become substantially more involved, and we are not aware of closed-form solutions. Below, we present the full derivation for $p = 2$ and highlight the technical difficulties that appear for $p > 2$.

Assume that $x_L^{\text{touch}}, x_R^{\text{touch}}$ and the corresponding levels α, β are given. Our goal is to determine parameters (a, b, c) such that

$$\tilde{k}(x_L^{\text{touch}}) = \alpha, \qquad \tilde{k}(x_R^{\text{touch}}) = \beta,$$

while additionally ensuring that

$$\forall x \in [x_{\text{start}}, x_{\text{end}}] : \quad \tilde{k}(x) \le k(x).$$

We first solve for a and c in terms of b, and subsequently derive the constraints that b must satisfy.

1. **Solving for a (as a function of b).** Setting $\tilde{k}\left(x_L^{\text{touch}}\right) = \alpha$ and $\tilde{k}\left(x_R^{\text{touch}}\right) = \beta$ and subtracting gives

$$\frac{\left(x_R^{\text{touch}} - a\right)^2}{b^2} - \beta = \frac{\left(x_L^{\text{touch}} - a\right)^2}{b^2} - \alpha \iff \frac{\left(x_R^{\text{touch}} - a\right)^2 - \left(x_L^{\text{touch}} - a\right)^2}{b^2} = \beta - \alpha.$$

   Using $v^2 - w^2 = (v - w)(v + w)$ the numerator becomes

$$\left(x_R^{\text{touch}} - a\right)^2 - \left(x_L^{\text{touch}} - a\right)^2 = \left(x_R^{\text{touch}} - x_L^{\text{touch}}\right)\left(x_L^{\text{touch}} + x_R^{\text{touch}} - 2a\right).$$

   Hence

$$\left(x_R^{\text{touch}} - x_L^{\text{touch}}\right)\left(x_L^{\text{touch}} + x_R^{\text{touch}} - 2a\right) = (\beta - \alpha)b^2,$$

   so

$$a = \frac{x_L^{\text{touch}} + x_R^{\text{touch}}}{2} - \frac{(\beta - \alpha)b^2}{2(x_R^{\text{touch}} - x_L^{\text{touch}})}. \tag{31}$$

   For compactness define the constants

$$A_0 := \frac{x_L^{\text{touch}} + x_R^{\text{touch}}}{2}, \quad K := \frac{\beta - \alpha}{2\left(x_R^{\text{touch}} - x_L^{\text{touch}}\right)},$$

   so that

$$a = A_0 - Kb^2.$$

   (When $\alpha = \beta$ we obtain $a = A_0$, the same midpoint as we obtained before.)

2. **Solving for c (as function of a, b).** From $\tilde{k}\left(x_L^{\text{touch}}\right) = \alpha$ we get

$$\frac{\left(x_L^{\text{touch}} - a\right)^2}{b^2} - c = \alpha \quad \Longrightarrow \quad c = \frac{\left(x_L^{\text{touch}} - a\right)^2}{b^2} - \alpha.$$

   Thus, once b is known, we can compute a and then c.

3. **Constraint on b from $\tilde{k}(x) \le k(x)$.** Write $\tilde{k}(x)$ using the expression for c :

$$\tilde{k}(x) = \frac{(x - a)^2}{b^2} - \left(\frac{(x_L^{\text{touch}} - a)^2}{b^2} - \alpha\right) = \frac{(x - a)^2 - (x_L^{\text{touch}} - a)^2}{b^2} + \alpha.$$

   Denote

$$N(x) := (x - a)^2 - \left(x_L^{\text{touch}} - a\right)^2.$$

   Factor $N(x)$ as

$$N(x) = \left(x - x_L^{\text{touch}}\right)\left(x + x_L^{\text{touch}} - 2a\right).$$

   Substitute $a = A_0 - Kb^2$. Since $2A_0 = x_L^{\text{touch}} + x_R^{\text{touch}}$ we have

$$x + x_L^{\text{touch}} - 2a = x + x_L^{\text{touch}} - 2A_0 + 2Kb^2 = x - x_R^{\text{touch}} + 2Kb^2$$

   and therefore

$$N(x) = \left(x - x_L^{\text{touch}}\right)\left(x - x_R^{\text{touch}}\right) + 2Kb^2\left(x - x_L^{\text{touch}}\right).$$

   The requirement $\tilde{k}(x) \le k(x)$ becomes

$$\frac{N(x)}{b^2} + \alpha \le k(x) \quad \Longleftrightarrow \quad N(x) \le b^2(k(x) - \alpha).$$

Plugging the expression for $N(x)$ gives

$$\left(x - x_L^{\text{touch}}\right)\left(x - x_R^{\text{touch}}\right) + 2Kb^2\left(x - x_L^{\text{touch}}\right) \le b^2(k(x) - \alpha),$$

or equivalently

$$b^2\left(k(x) - \alpha - 2K\left(x - x_L^{\text{touch}}\right)\right) \ge \left(x - x_L^{\text{touch}}\right)\left(x - x_R^{\text{touch}}\right).$$

Define

$$D(x) := k(x) - \alpha - 2K\left(x - x_L^{\text{touch}}\right).$$

Then for each x the inequality reads

$$b^2 D(x) \ge \left(x - x_L^{\text{touch}}\right)\left(x - x_R^{\text{touch}}\right)$$

Now consider three cases depending on the sign of $D(x)$:

- If $D(x) = 0$, the constraint reduces to $0 \ge \left(x - x_L^{\text{touch}}\right)\left(x - x_R^{\text{touch}}\right)$, which must hold for that x.
- If $D(x) > 0$, dividing by $D(x)$ results in

$$b^2 \ge \frac{\left(x - x_L^{\text{touch}}\right)\left(x - x_R^{\text{touch}}\right)}{D(x)}$$

- If $D(x) < 0$, dividing by the negative $D(x)$ flips the inequality:

$$b^2 \le \frac{\left(x - x_L^{\text{touch}}\right)\left(x - x_R^{\text{touch}}\right)}{D(x)}$$

To satisfy $\tilde{k}(x) \le k(x)$ for all $x \in [x_{\text{start}}, x_{\text{end}}]$ we therefore need to find b such that the above inequalities hold simultaneously for every x in the domain. Equivalently, letting

$$\underline{R}(x) := \frac{\left(x - x_L^{\text{touch}}\right)\left(x - x_R^{\text{touch}}\right)}{D(x)} \quad \text{for } D(x) > 0,$$

$$\overline{R}(x) := \frac{\left(x - x_L^{\text{touch}}\right)\left(x - x_R^{\text{touch}}\right)}{D(x)} \quad \text{for } D(x) < 0,$$

a feasible b^2 must satisfy

$$b^2 \ge \sup_{x: D(x) > 0} \underline{R}(x) \quad \text{and} \quad b^2 \le \inf_{x: D(x) < 0} \overline{R}(x)$$

and additionally all x with $D(x) = 0$ must satisfy $\left(x - x_L^{\text{touch}}\right)\left(x - x_R^{\text{touch}}\right) \le 0$. Hence, our k_{ext} construction instead, that works for any p.

Discussion on higher-degree polynomials When $p > 2$, however, the problem is more challenging and we are unaware on how to solve this problem in these settings. For instance, in Eq. (31), a was isolated by explicitly exploiting that $p = 2$, and for larger p this isolation of a in terms of b is no longer evident.

F.6 A bound on the Between-Grid Violation of $\tilde{k}$

This section belongs with the ℓ_p-ellipsoid construction because it turns the grid-based fit into a guaranteed outer approximation. We first derive a local between-grid bound when a global minimizer lies on the grid, then generalize it to the practical case where only a lower bound ϕ on the minimum is known, and finally derive the safe construction used for $\mathcal{C}^{\ell_p\text{safe}}$.

F.6.1 Setup and Assumptions

Let $k : [x_{\text{start}}, x_{\text{end}}] \to \mathbb{R}$ and let $x_1 < \cdots < x_n$ be a uniform grid with spacing h:

$$x_i = x_{\text{start}} + (i-1)h, \qquad h = \frac{x_{\text{end}} - x_{\text{start}}}{n-1}.$$

Let $\tilde{k}(x) = \frac{|x-a|^p}{b^p} - c$ with $a \in \mathbb{R}$, $b > 0$, $c \in \mathbb{R}$ and integer $p \ge 1$. Assume:

(A1) k is quasiconvex and continuous on $[x_{\text{start}}, x_{\text{end}}]$ (follows from Lem. C.2).

(A2) $\tilde{k}$ is quasiconvex (Lem. C.1).

(A3) Grid lower bound: $\tilde{k}(x_i) \leq k(x_i)$ for all $i \in \{1, \ldots, n\}$.

Define the pointwise error $\epsilon(x) := \tilde{k}(x) - k(x)$.

F.6.2 Error Bound when a Global Minimizer Lies on the Grid

Assumption F.1. Assume there exists a global minimizer $x^\star \in [x_{\text{start}}, x_{\text{end}}]$ that corresponds with a grid point x_m.

Theorem F.1 (Local interval bound)**.** *Under (A1)–(A3) and Asm. F.1, for any $i \in \{1, \ldots, n-1\}$ and any $x \in [x_i, x_{i+1}]$,*

$$\epsilon(x) \leq \max\{\tilde{k}(x_i), \tilde{k}(x_{i+1})\} - \begin{cases} k(x_{i+1}), & i < m, \\ k(x_i), & i \geq m. \end{cases}$$

Proof. Fix an interval $I = [x_i, x_{i+1}]$ and an arbitrary $x \in I$. Since $x^\star = x_m$ is a global minimizer and k is quasiconvex, k is non-increasing on $[x_{\text{start}}, x_m]$ and non-decreasing on $[x_m, x_{\text{end}}]$. We distinguish two cases based on the grid index m.

1. **Case A: $i \geq m$ (interval to the right of x_m).** Then $I \subseteq [x_m, x_{\text{end}}]$ and k is non-decreasing on I, hence $k(x) \geq k(x_i)$. Therefore,

$$\begin{aligned} \epsilon(x) &= \tilde{k}(x) - k(x) \\ &\leq \tilde{k}(x) - k(x_i) \quad \text{since } k(x) \geq k(x_i) \text{ on } I \\ &\leq \max\{\tilde{k}(x_i), \tilde{k}(x_{i+1})\} - k(x_i) \quad \text{since } \tilde{k} \text{ is quasiconvex.} \end{aligned}$$

2. **Case B: $i < m$ (interval to the left of x_m).** Then $I \subseteq [x_{\text{start}}, x_m]$ and k is non-increasing on I, hence $k(x) \geq k(x_{i+1})$. Thus, analogously

$$\epsilon(x) \leq \tilde{k}(x) - k(x_{i+1}) \leq \max\{\tilde{k}(x_i), \tilde{k}(x_{i+1})\} - k(x_{i+1}).$$

This proves the bound with $k(x_i)$ for $i \geq m$ and $k(x_{i+1})$ for $i < m$. □

F.6.3 When the Minimizer is Not a Grid Point

The bound in Thm. F.1 assumes that a global minimizer $x^\star$ corresponds to a grid point. If $x^\star$ is unknown, but a global lower bound ϕ on k is available (i.e. $\forall x : \; k(x) \geq \phi$), we can still obtain a valid upper bound on $\epsilon(x)$. In practice, this lower bound comes directly from the global-minimum procedure in App. D, by setting $\phi \leftarrow \underline{k_\star}$.

Theorem F.2 (Bound using a global lower bound)**.** *Assume (A1)–(A3) and that a global lower bound ϕ is known such that $k(x) \geq \phi$ for all $x \in [x_{\text{start}}, x_{\text{end}}]$. Let $m := \arg\min_i k(x_i)$ be any index achieving the grid minimum. Let $l < m$ be the largest index with $k(x_l) > k(x_m)$ (if it exists) and $r > m$ the smallest index with $k(x_r) > k(x_m)$ (if it exists). Then every minimizer $x^\star$ lies in (x_l, x_r) whenever both l and r exist. Moreover, for any interval $[x_i, x_{i+1}]$,*

$$\epsilon(x) \leq \max\{\tilde{k}(x_i), \tilde{k}(x_{i+1})\} - L_i \qquad \forall x \in [x_i, x_{i+1}],$$

where

$$L_i := \begin{cases} k(x_{i+1}), & i \leq l-1, \\ k(x_i), & i \geq r, \\ \phi, & l \leq i \leq r-1. \end{cases}$$

Proof. By construction, we have $x_l < x_m < x_r$ and $k(x_m) < k(x_l), k(x_r)$. The quasiconvexity of $k(\cdot)$ then implies that all true minimizers $x^\star$ must lie in the open interval (x_l, x_r): otherwise, if $x^\star$ were to the right of x_r, applying the quasiconvexity inequality (that for any x, y and any $\lambda \in [0, 1]$,

one has $k(\lambda x + (1-\lambda)y) \leq \max\{k(x), k(y)\}$) to x_m and $x^\star$ and choosing the convex combination that gives x_r would force $k(x_r) \leq \max\{k(x_m), k(x^\star)\}$, which is a contradiction. The argument for $x^\star$ being left of x_l is analogous.

Consequently, for left-side intervals: $[x_1, x_2], [x_2, x_3], \ldots, [x_{l-1}, x_l]$, the function k is non-increasing, and for the right-side intervals, $[x_r, x_{r+1}], \ldots$, it is non-decreasing, so the local bound derived in Thm. F.1 applies there. This addresses the cases for L_i where $i \leq l-1$ or $i \geq r$. For intervals with $l \leq i \leq r-1$ (i.e., inside $[x_l, x_r]$), the monotonicity assumption fails, but we can use the global lower bound ϕ to obtain a weaker bound: since $k(x) \geq \phi$ for all x,

$$\epsilon(x) = \tilde{k}(x) - k(x) \leq \tilde{k}(x) - \phi,$$

and consequently for an interval $[x_i, x_{i+1}] \subset (x_l, x_r)$ we may bound

$$\epsilon(x) \leq \max\{\tilde{k}(x_i), \tilde{k}(x_{i+1})\} - \phi.$$

Hence, the case for L_i with $l \leq i \leq r-1$ is also addressed. This concludes the proof.

□

Corollary F.1 (Global bound)**.** *Under the assumptions of Thm. F.2,*

$$\epsilon_{\max} := \max_{x \in [x_{\text{start}}, x_{\text{end}}]} \epsilon(x) \leq \max_{i=1,\ldots,n-1} \left(\max\{\tilde{k}(x_i), \tilde{k}(x_{i+1})\} - L_i \right).$$

This bound is easily computable from the grid evaluations and allows verifying whether $\tilde{k} \leq k$ holds everywhere. It also suggests an adaptive refinement scheme: whenever the bound at an interval exceeds a prescribed tolerance, subdivide that interval by adding grid points and recomputing $\tilde{k}$ on the refined grid.

F.6.4 A Conservative Construction with Zero Between-Grid Violation

Proposition F.1 (One-step shifted safe values)**.** *Assume **(A1)–(A3)** and that a global lower bound ϕ is known such that $k(x) \geq \phi$ for all $x \in [x_{\text{start}}, x_{\text{end}}]$. Define modified grid values*

$$k_i^{\text{s}} := \begin{cases} k(x_{i+1}), & i < l, \\ \phi, & l \leq i \leq r, \\ k(x_{i-1}), & i > r, \end{cases}$$

(with $k_1^{\text{s}} := k(x_2)$ if $i > 1$ and $k_n^{\text{s}} := k(x_{n-1})$ if $r < n$). If $\tilde{k}$ is constructed to satisfy $\tilde{k}(x_i) \leq k_i^{\text{s}}$ for all i, then the bound on $\epsilon_{\max}$ from Cor. F.1 is zero or less:

$$\max\{\tilde{k}(x_i), \tilde{k}(x_{i+1})\} - L_i \leq 0 \qquad \forall i,$$

and therefore $\epsilon_{\max} \leq 0$.

Proof. We treat the cases:

1. If $i \geq r$, then the bound from Cor. F.1 is given by

   $$\epsilon(x) \leq \max\{\tilde{k}(x_i), \tilde{k}(x_{i+1})\} - k(x_i).$$

   Then, since by construction $\tilde{k}(x_i) \leq k_i^s$ and $\tilde{k}(x_{i+1}) \leq k_{i+1}^s$, we get:

   $$\max\{\tilde{k}(x_i), \tilde{k}(x_{i+1})\} - k(x_i) \leq \max\{k_i^s, k_{i+1}^s\} - k(x_i).$$

   And since $k_i^{\text{s}} = k(x_{i-1})$ and $k_{i+1}^{\text{s}} = k(x_i)$, we obtain

   $$\begin{aligned} \max\{k_i^s, k_{i+1}^s\} - k(x_i) &= \max\{k(x_{i-1}), k(x_i)\} - k(x_i) \quad \text{By definition of } k^s \\ &= 0 \quad \text{Since } k \text{ is non-decreasing on } [x_r, x_{\text{end}}] \end{aligned}$$

   To understand why k is non-decreasing on $[x_r, x_{\text{end}}]$, we refer to Thm. F.2, first sentence in paragraph two of the proof. Thus, we have that the per-interval bound is ≤ 0 for $i \geq r$.

2. If $i \leq l-1$, then the proof is analogous to 1) by using the non-increasing property of k on the left $[x_{\text{start}}, x_l]$.

3. If $l \leq i \leq r-1$, then the bound is given by

$$\begin{aligned}\epsilon(x) &\leq \max\{\tilde{k}(x_i), \tilde{k}(x_{i+1})\} - \phi \\ &\leq \max\{k_i^s, k_{i+1}^s\} - \phi \quad \text{since by construction } \tilde{k}(x_i) \leq k_i^s \text{ and } \tilde{k}(x_{i+1}) \leq k_{i+1}^s \\ &= \max\{\phi, \phi\} - \phi = 0.\end{aligned}$$

Making the per-interval bound $l \leq i \leq r-1$.

□

This construction can thus be viewed as shrinking the feasible set of $\tilde{k}$ to account for the worst-case between-grid behavior under the quasiconvexity assumptions of k.

Part III: Experiments

G Experimental Setup & Additional Experiments

G.1 Setup

G.1.1 Data and various experiment parameters.

We consider observations generated from two multivariate distributions. The first consists of i.i.d. samples drawn uniformly from the unit ball, so that each observation satisfies $\|\mathbf{y}\| \leq 1$. This setting deliberately favors confidence-sphere baselines, which explicitly assume a global norm bound of this form, whereas our method only requires coordinate-wise bounds $y_d \in [0, 1]$. The second distribution is a Bernoulli mixture: half of the coordinates follow $y_d \sim \text{Bernoulli}(0.5)$, while the remaining coordinates are shifted and rescaled to lie in $[0.5, 1]$, again with success probability 0.5. In this case, Bounded UP and our constructions are advantaged, as the spherical baselines require rescaling the data to satisfy the $\|\mathbf{y}\| \leq 1$ constraint. For the experiments in App. G.4.3, we additionally consider a highly correlated distribution with correlation parameter $\rho = 0.98$. Each coordinate is uniform on $[0, 1]$, but the coordinates within an observation are nearly equal, producing a narrow cloud along the main diagonal.

Across all experiments, methods assuming a norm bound $\|\mathbf{y}\| \leq 1$ (i.e., Banach Sphere and Conf Sphere) receive data that are appropriately shifted and scaled to lie within the unit ball, whereas for our method and Bounded UP, the data are scaled to lie in $[0, 1]^D$. Additional details, such as method parameters and how volume is computed, are summarized in App. G.1.

Generally, we set the number of observations to 1,000, unless stated otherwise. For the volume-versus-time experiments in the main text, we fix the dimension to $D = 5$. For the small-scale experiments involving Bounded UP, we use $D = 2$. All portfolio weights and Bonferroni weights are chosen uniformly as $w_d = 1/D$. Empirical per-time-step coverage is evaluated using 200 independent repetitions for each configuration. The error level α is set to 0.05 across all experiments.

G.1.2 Hedged Capital Process and Capital Stakes

Following the discussion from § 2, we summarize the authors' concrete suggestions for the stakes ($\lambda_i^+(\cdot)$ and $\lambda_i^-(\cdot)$), making them adaptive to the variance of the observed data:

$$\widetilde{\lambda}_t \;=\; \sqrt{\frac{2\log(2/\alpha)}{\widehat{\sigma}_{t-1}^2\, t\, \log(t+1)}},$$

where

$$\widehat{\mu}_t = \frac{\frac{1}{2} + \sum_{i=1}^{t} y_i}{t+1}, \quad \widehat{\sigma}_t^2 = \frac{\frac{1}{4} + \sum_{i=1}^{t} (y_i - \widehat{\mu}_i)^2}{t+1}.$$

Finally, the stakes are then defined as:

$$\lambda_t^+(m) = \min\Big(\widetilde{\lambda}_t,\ \frac{c}{m}\Big),\ \lambda_t^-(m) = \min\Big(\widetilde{\lambda}_t,\ \frac{c}{1-m}\Big), \tag{32}$$

where c is a constant ($c = 1/2$ or $3/4$). In our experiments, we set $c = 3/4$ and its parameter θ in Eq. (3) to $1/2$.

The term stake is inspired by $k_t^{\pm}$'s interpretation as a *betting process*. $k_t^+(m)$ represents the accumulated wealth after t sequential bets on the hypothesis $\mu \geq m$. At round i, the bettor risks a fraction $\lambda_i^+(m)$ of their current wealth, resulting in the multiplicative update $k_i^+(m) = k_{i-1}^+(m)\left(1 + \lambda_i^+(m)\,(y_i - m)\right)$. Here, $\lambda_i^+(m) \in [0, 1/m)$ is the "stake", chosen based on the past $i-1$ observations. An analogous interpretation holds for $k_t^-(m)$, with stakes $\lambda_i^-(m) \in [0, 1/(1-m))$, corresponding to bets on $\mu < m$. This interpretation motivates the term *betting-based* CS.

G.1.3 Ellipsoid construction.

The ℓ_p-ellipsoid construction uses a fixed one-dimensional grid of size $N = 2001$ for each coordinate. N remains fixed independent of the interval width $[x_{\text{start}}, x_{\text{end}}]$. The exponent p is selected from the candidate set $\{6, \ldots, 16\}$. The touching parameter is set to $\gamma_d = 1/(\alpha \sqrt[p]{D})$.

G.1.4 Volume estimation.

When the volume of a region is not available in closed form, for example, for ellipsoid- or sphere-based regions intersecting the domain $[0,1]^D$, or for $\mathcal{C}^{\text{port}}$, $\mathcal{C}^{\text{bbox} \cap \ell_p}$ and Bounded UP, we resort to Monte Carlo estimation. Specifically, volumes are estimated using 5 independent runs of $25{,}000$ samples each. We report the mean volume estimate, and visualize uncertainty via a transparent overlay corresponding to a 95% standard-deviation band.

Whenever possible, sampling is restricted to a tight outer bounding box. For $\mathcal{C}^{\text{port}}$ and $\mathcal{C}^{\text{bbox} \cap \ell_p}$, we sample uniformly from the bounding box defined by $\mathcal{C}^{\text{bbox}}$. For $\mathcal{C}^{\ell_p}$ and the sphere-based baselines, we use the minimal axis-aligned bounding box enclosing the ellipsoid. For Bounded UP, we are not aware of a tighter sampling region, and therefore sample uniformly from the full domain $[0,1]^D$.

Note that as the number of dimensions increases, the accuracy of the Monte Carlo volume estimate decreases, requiring more samples to compensate. We analyze this accuracy both theoretically and empirically in Ablation App. G.4.2.

G.1.5 Compute resources.

All experiments were run on a single laptop, without GPU acceleration or specialized hardware. The machine has an Intel Core i7-13650HX processor and 16 GB of RAM. We used this same environment for all reported runs in the paper and appendix.

G.1.6 Evaluation time.

Tab. 4 reports mean wall-clock times for method construction and volume evaluation at $t = 100$. $\mathcal{C}^{\text{port}}$ and Bounded UP are not included due to their exponential computational cost. But Bounded UP would require the same exponential number of calls as $\mathcal{C}^{\text{port}}$ for its respective wealth function with the difference that each evaluation is much more costly in Bounded UP.

Table 4: Mean wall-clock evaluation time in seconds of computing the volume at $t = 100$.

	Hedged-capital methods				External baselines		
D	$\mathcal{C}^{\text{bbox}}$	$\mathcal{C}^{\text{bonf}}$	$\mathcal{C}^{\ell_p\text{safe}}$	$\mathcal{C}^{\text{bbox}\cap\ell_p\text{safe}}$	Banach Sphere	Conf Sphere	Sub-ψ
2	0.006036	0.006204	1.613209	1.667916	0.000830	0.000841	0.000812
3	0.009015	0.008886	2.383790	2.418203	0.000746	0.000908	0.000954
4	0.011913	0.011997	3.183447	3.186238	0.000731	0.000789	0.000742
5	0.015119	0.015322	4.041727	4.063602	0.000732	0.000796	0.000791
6	0.017873	0.017848	4.827297	4.798591	0.000723	0.008539	0.022119
7	0.020892	0.021427	5.656130	5.695633	0.000740	0.012678	0.071832
8	0.023990	0.023965	6.405785	6.409076	0.000774	0.013652	0.130039
9	0.027374	0.027274	7.164873	7.200621	0.000744	0.013114	0.144640
10	0.030424	0.031048	7.943246	7.972349	0.000745	0.013018	0.163916

G.2 Small-Scale Experiments

Because the Bounded UP CS's $O(t^{D+1})$ complexity of verifying membership is too costly to evaluate for larger settings (recall that we need to estimate its volume through Monte Carlo sampling, requiring another $5 \times 25{,}000$ samples at each t), we present a smaller-scale experiment to allow for comparison with this baseline as well.

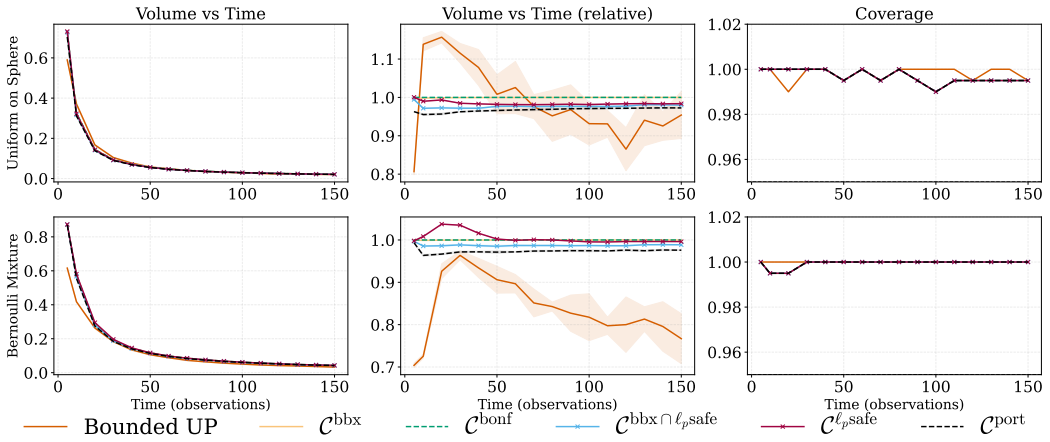


Figure 9: Small-scale comparison on synthetic data that includes Bounded UP.

G.3 $\mathcal{C}^{\ell_p}$ vs $\mathcal{C}^{\ell_p\text{safe}}$

Fig. 10 compares the volumes obtained by the safe and non-safe variants, which perform remarkably similarly. The additional computational cost of the safe variant, measured in the number of $k_t^d(\cdot)$ calls, is quantified in Tab. 5.

G.4 Ablation Studies

G.4.1 Ellipsoid Parameters

We evaluate three parameters: for the ℓ_p-ellipsoid $\mathcal{C}^{\ell_p}$ we consider effect of the touching parameter λ_d and p on the volume. For its safe construction $\mathcal{C}^{\ell_p\text{safe}}$, we look at the impact of the grid size N.

Fig. 11 displays the impact of the parameter p, whereas the impact on the touch level γ_d is reported in Fig. 12. Lastly, the effect on the grid size for the safe ellipsoid is shown in Fig. 13. In all three plots $\mathcal{C}^{\text{bonf}}$ and $\mathcal{C}^{\text{port}}$ are included for reference.

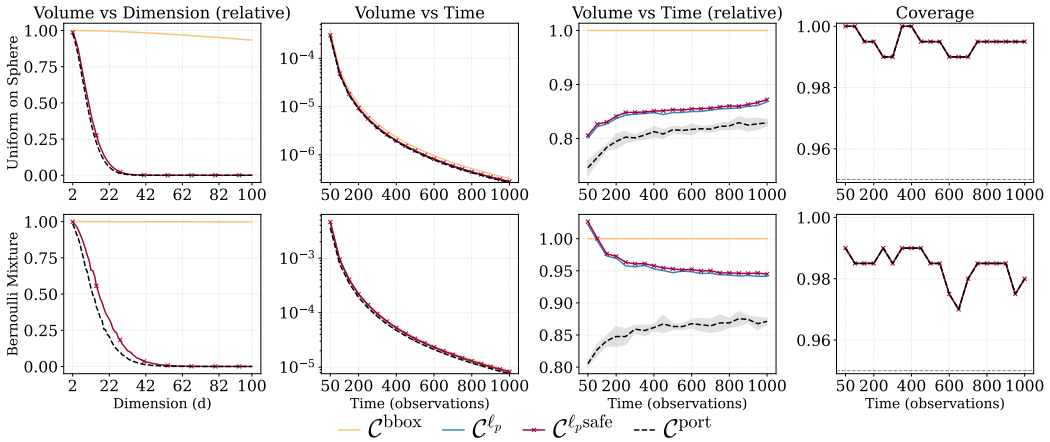


Figure 10: Comparison between the "safe" and non-safe ℓ_p-ellipsoids.

Table 5: Safe vs non-safe approximations: number of $k_t^d(\cdot)$ calls to compute the volume at $t = 100$ (sorted from most to least efficient).

D	$\mathcal{C}^{\ell_p}$	$\mathcal{C}^{\ell_p\text{safe}}$	$\mathcal{C}^{\text{bbox}} \cap \mathcal{C}^{\ell_p}$	$\mathcal{C}^{\text{bbox}} \cap \mathcal{C}^{\ell_p\text{safe}}$
2	1.07×10^5	1.60×10^5	1.08×10^5	1.61×10^5
3	1.62×10^5	2.40×10^5	1.63×10^5	2.41×10^5
4	2.15×10^5	3.20×10^5	2.16×10^5	3.21×10^5
5	2.69×10^5	4.00×10^5	2.70×10^5	4.01×10^5
6	3.23×10^5	4.80×10^5	3.25×10^5	4.82×10^5
7	3.77×10^5	5.61×10^5	3.79×10^5	5.63×10^5
8	4.31×10^5	6.41×10^5	4.33×10^5	6.43×10^5
9	4.85×10^5	7.22×10^5	4.87×10^5	7.24×10^5
10	5.40×10^5	8.03×10^5	5.43×10^5	8.06×10^5

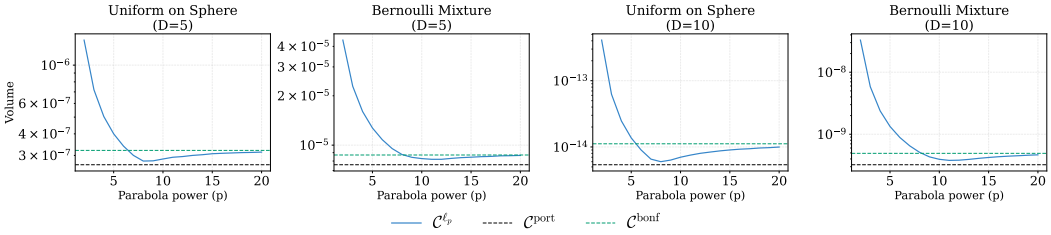


Figure 11: Impact on the exponent p.

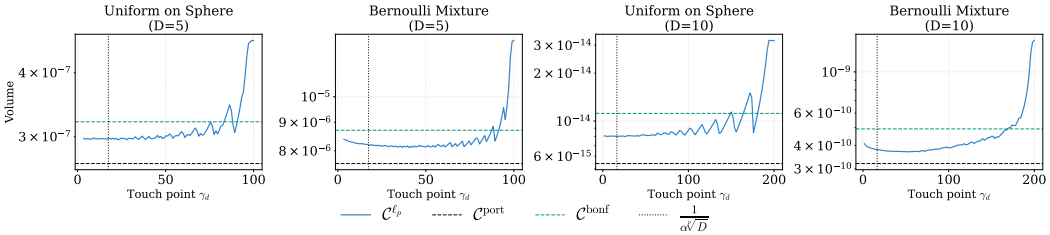


Figure 12: Impact on the touch level γ_d. The vertical indicator marks our default choice across all experiments. For an isolated comparison, the parameter p is kept fixed at $p = 12$.

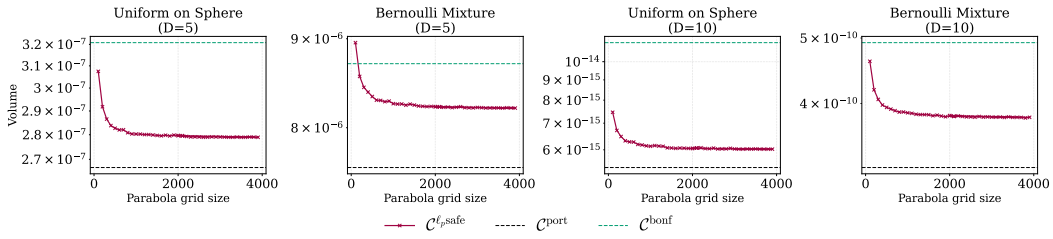


Figure 13: Impact on the grid size N.

G.4.2 Accuracy of Monte Carlo Volume Estimation

Since $\mathcal{C}^{\text{bbox}} \cap \ell_p$ in § 4.3 requires Monte Carlo sampling to estimate the volume, we analyze how accurate the estimate remains as the dimension D increases.

Let

$$\rho_D = \frac{\text{Vol}\big(\mathcal{C}^{\text{bbox}} \cap \mathcal{C}^{\ell_p}\big)}{\text{Vol}(\mathcal{C}^{\text{bbox}})}$$

denote the acceptance probability when sampling uniformly from the bounding box. With M independent Monte Carlo samples, the number of accepted points satisfies

$$H \sim \text{Binomial}(M, \rho_D), \qquad \mathbb{E}[H] = M\rho_D.$$

The Monte Carlo estimate of $\text{Vol}(\mathcal{C}^{\text{bbox}} \cap \mathcal{C}^{\ell_p})$ is:

$$\widehat{V} = \text{Vol}(\mathcal{C}^{\text{bbox}})\frac{H}{M}.$$

Its accuracy can be quantified through the relative standard error, defined as the standard deviation of $\widehat{V}$ divided by the target volume $\text{Vol}(\mathcal{C}^{\text{bbox}} \cap \mathcal{C}^{\ell_p})$. Since $\text{Var}(H) = M \cdot \rho_D(1-\rho_D)$, one can derive this standard error as:

$$\frac{\sqrt{\text{Var}(\widehat{V})}}{\text{Vol}(\mathcal{C}^{\text{bbox}} \cap \mathcal{C}^{\ell_p})} = \sqrt{\frac{1-\rho_D}{M \cdot \rho_D}}.$$

Hence, achieving relative standard error at most ε requires

$$M \geq \frac{1-\rho_D}{\rho_D \varepsilon^2}.$$

Although ρ_D need not decrease exponentially in every setting, it can become very small as D grows, and our method provides no dimension-free bound on it. Thus, the number of samples required for an accurate volume estimate may grow rapidly with the dimension.

Tab. 6 validates empirically how the relative error is affected by the dimension for $M = 25{,}000$ samples.

Thus, box sampling remains effective at $D = 20$ in these examples, but deteriorates substantially by $D = 50$, particularly for the uniform-on-sphere setting. It is nevertheless far better than sampling from $[0,1]^D$.

Table 6: Monte Carlo acceptance rates and resulting estimation accuracy for different dimensions. The number of accepted points is computed for a budget of $25{,}000$ Monte Carlo samples.

Distribution	D	Estimated acceptance rate ρ_D	Accepted points	Rel. std. error
Bernoulli mixture	20	0.359	$8{,}987$	0.844%
	50	1.12×10^{-2}	280	5.94%
	100	3.13×10^{-6}	7.84×10^{-2}	357%
Uniform on Sphere	20	0.118	$2{,}947$	1.73%
	50	7.17×10^{-5}	1.79	74.7%
	100	2.90×10^{-12}	7.25×10^{-8}	$3.71 \times 10^{5}\%$

G.4.3 Influence on Significance Level α

Figs. 14 to 16 compare the CSs for $\alpha \in \{0.01, 0.05, 0.10\}$, respectively. Across all settings, the estimated cumulative probability of excluding the true mean by $t = 1000$ remains below the level α. The relative-volume rankings are also stable across significance levels: our approximations remain close to the portfolio region and improve substantially over previous tractable baselines.

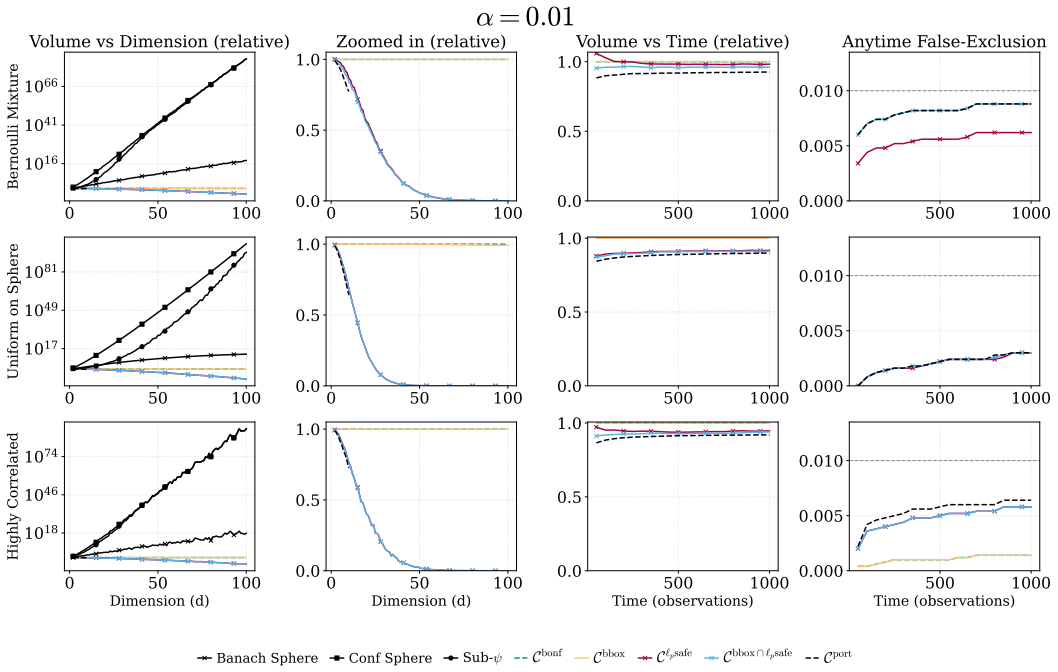


Figure 14: Results for $\alpha = 0.01$. Rows correspond to the Bernoulli-mixture, uniform-sphere, and highly correlated settings. Columns show relative volume versus dimension (with baselines and a zoom-in), relative volume versus time, and the cumulative false-exclusion rate; the dashed horizontal line marks the nominal α.

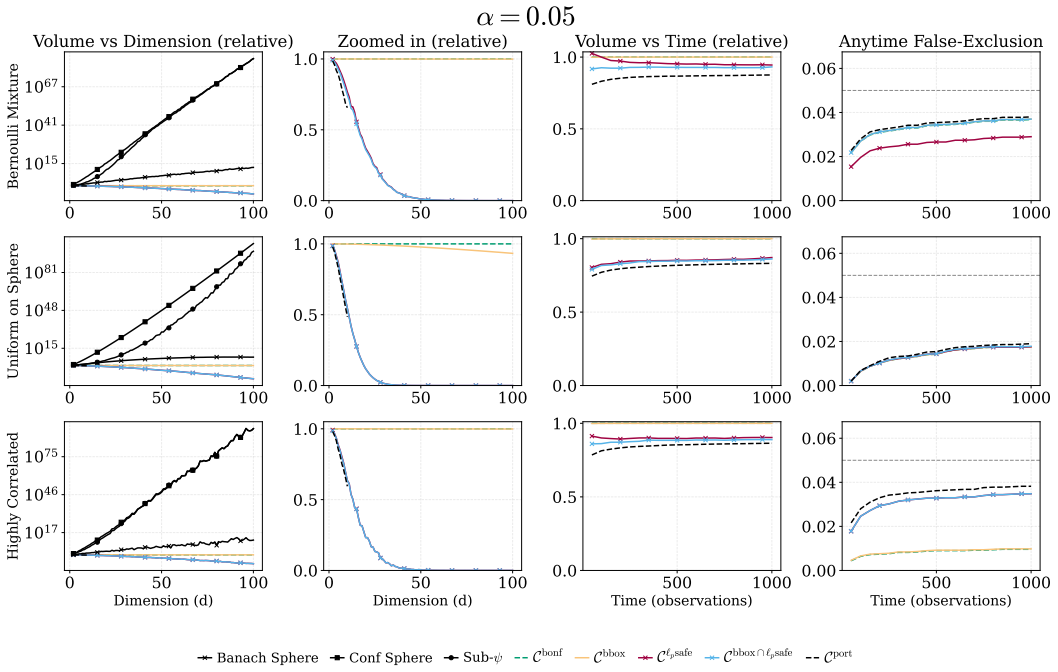


Figure 15: Results for $\alpha = 0.05$. Rows and columns follow Fig. 14; the dashed horizontal line marks the nominal α.

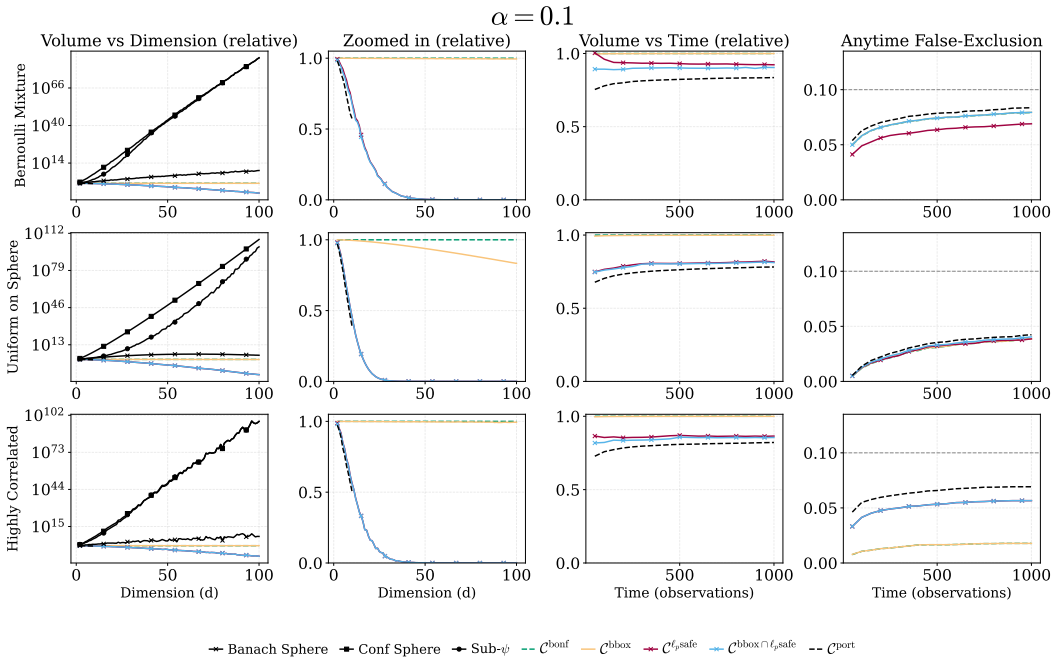


Figure 16: Results for $\alpha = 0.10$. Rows and columns follow Fig. 14; the dashed horizontal line marks the nominal α.

G.5 A/B-Testing Experiment: Synthetic Web Metrics

Short Overview We simulate a sequential A/B test comparing two variants of a web shop to determine whether there is a statistically significant difference between them. Let $\mathbf{y}_t^{(A)}$ denote the vector of observed metrics for website A at step t, where t may correspond to an individual user (and analogously for B), and define $\boldsymbol{\Delta}_t := \mathbf{y}_t^{(B)} - \mathbf{y}_t^{(A)}$. A CS then gives a region $\mathcal{C}_t$ for $\mathbb{E}[\boldsymbol{\Delta}_t]$ at every time t. If the interval for a particular coordinate lies entirely above or below 0, we can conclude that the variants differ significantly on that metric. On the other hand, if the region becomes sufficiently small without excluding 0 (measured through volume), we may conclude that there is no practically meaningful difference and stop collecting data.

Dataset generation. The simulation process is designed to mimic a real-world setting. Time is indexed by batches $t = 1, \ldots, T$. At each batch t, we sample $n_{t,A}$ independent sessions from variant A and $n_{t,B}$ independent sessions from variant B, where $n_{t,A}$ and $n_{t,B}$ vary by $\pm 15\%$ around a nominal batch size $n = 400$ (with a minimum of 100 sessions per variant).

Each session is represented by a $D = 10$-dimensional metric vector in $[0,1]^D$. The first eight coordinates mimic typical web-analytics metrics: **(i)** time-on-site (seconds, clipped at 300s and divided by 300), **(ii)** number of clicks (clipped at 20 and divided by 20), **(iii)** number of pages viewed (clipped at 10 and divided by 10), **(iv)** scroll depth (already in $[0,1]$), **(v)** add-to-cart (Bernoulli), **(vi)** purchase (Bernoulli), **(vii)** non-bounce (Bernoulli; equals $1 -$ bounce so that larger is better), and **(viii)** 7-day return (Bernoulli). The final two coordinates are added as independent noise variables sampled uniformly from $[0,1]$. All coordinates are therefore bounded in $[0,1]$.

To create correlation between metrics, we draw a latent *engagement* variable $E \in [0,1]$ per session. Conditional on E, continuous metrics are generated from simple parametric distributions (log-normal for time-on-site, Poisson for counts, Beta for scroll depth), and binary metrics are generated by Bernoulli draws whose success probabilities increase with E. This produces realistic correlations (e.g., sessions with larger E tend to have more time, more clicks, deeper scroll, and higher conversion probability).

Fig. 17 visualizes what a single batch looks like (i.e., the $n = 400$ samples). For each coordinate, we plot a histogram of samples for variants A and B within one batch (not the distribution of batch

averages over time). The latent engagement variable E and its shifted treatment variant are shown in the top-left panel.

Two experiments: with effect vs. no effect. We run two variants of the simulation:

- **With effect:** variant B is improved on a subset of metrics by shifting the engagement signal upward for those metrics only. Concretely, for the affected coordinates, we replace E by $\tilde{E} = \min(E + \delta, 1)$ when generating sessions from B, with shift $\delta = 0.06$. This means that the two variants differ only in a small set of coordinates (time, clicks, scroll, and add-to-cart), while the remaining metrics (pages, purchase, non-bounce, return, and the two noise dimensions) are approximately unchanged.
- **No effect:** we set $\delta = 0$ and apply no coordinate-specific shift, so that variants A and B are generated from the same distribution.

From two variants to one bounded stream (batched pairing). Our CSs are defined for a single stream of bounded observations. We therefore combine the two variants within each batch. Let $\mathbf{y}_t^{(A)}, \mathbf{y}_t^{(B)} \in [0,1]^D$ denote the per-batch mean metric vectors for variants A and B. We form the per-batch difference

$$\boldsymbol{\Delta}_t \;=\; \mathbf{y}_t^{(B)} - \mathbf{y}_t^{(A)} \in [-1,1]^D,$$

and rescale it coordinate-wise into $[0,1]$ by

$$\mathbf{z}_t \;=\; \tfrac{1}{2}(\boldsymbol{\Delta}_t + \mathbf{1}) \in [0,1]^D.$$

Under the no-effect experiment, $\mathbb{E}[\mathbf{z}_t] = \frac{1}{2}\mathbf{1}$, i.e., $\mathbb{E}[z_t^d] = 0.5$ for every coordinate d. In the with-effect experiment, only the affected coordinates satisfy $\mathbb{E}[z_t^d] > 0.5$; the unaffected and noise coordinates remain close to 0.5.

We run all confidence region methods at level $\alpha = 0.05$ on the sequential stream $\{\mathbf{z}_t\}_{t=1}^T$.

Stopping and decision rule. At each time t, let $\mathcal{C}_t$ be the confidence region for the mean of $\mathbf{z}_t$. We conclude coordinate-wise significance for metric d as soon as $\mathcal{C}_t$ excludes 0.5 on that axis:

$$\underline{\mu}_{t,d} > 0.5 \quad \text{or} \quad \overline{\mu}_{t,d} < 0.5,$$

where $[\underline{\mu}_{t,d}, \overline{\mu}_{t,d}]$ denotes the coordinate interval for μ_d. If no coordinate is significant, we stop and decide nothing once the region is sufficiently small in volume. Concretely, we stop without significance when $\sqrt[D]{V(\mathcal{C}_t)} \le r_{\mathrm{tol}}$.

In the main experiments, we use $D = 10$, batch size $n = 400$, $r_{\mathrm{tol}} = 0.02$, and run up to $T = 10{,}000$ batches. The main results are shown in Tab. 3.

Ablation: sensitivity to batch size. We repeat the experiment for $n \in \{50, 100, 200, 400, 800\}$. As shown in Tab. 7, the relative performance of the methods remains consistent. Interestingly, although larger batches reduce the variance of each batch mean, stopping times measured in batches change little: our methods stop after 170–181 batches with an effect and 355–428 batches without an effect. Consequently, smaller batches reach the same decisions with substantially fewer users in this experiment.

Table 7: **Sensitivity of A/B stopping outcomes to batch size.** Arrows report the results as the nominal batch size increases from $n = 50$ to $n = 800$; the intermediate values follow the same trend. Approximate users are reported per variant.

	With effect		No effect	
Method	Stop_t	Approx. users	Stop_t	Approx. users
ConfSphere	$10{,}000 \to 10{,}000$	$500{,}000 \to 8{,}000{,}000$	$10{,}000 \to 10{,}000$	$500{,}000 \to 8{,}000{,}000$
BanachSphere	$5{,}227 \to 5{,}046$	$261{,}350 \to 4{,}036{,}800$	$8{,}273 \to 8{,}036$	$413{,}650 \to 6{,}428{,}800$
Sub-ψ	$6{,}352 \to 5{,}243$	$317{,}600 \to 4{,}194{,}400$	$9{,}259 \to 6{,}650$	$462{,}950 \to 5{,}320{,}000$
$\mathcal{C}^{\text{bonf}}$	$171 \to 176$	$8{,}550 \to 140{,}800$	$428 \to 397$	$21{,}400 \to 317{,}600$
$\mathcal{C}^{\text{bbox}}$	$170 \to 175$	$8{,}500 \to 140{,}000$	$427 \to 396$	$21{,}350 \to 316{,}800$
$\mathcal{C}^{\ell_p\text{safe}}$	$171 \to 176$	$8{,}550 \to 140{,}800$	$378 \to 355$	$18{,}900 \to 284{,}000$
$\mathcal{C}^{\text{bbox} \cap \ell_p\text{safe}}$	$\mathbf{170 \to 175}$	$\mathbf{8{,}500 \to 140{,}000}$	$\mathbf{378 \to 355}$	$\mathbf{18{,}900 \to 284{,}000}$

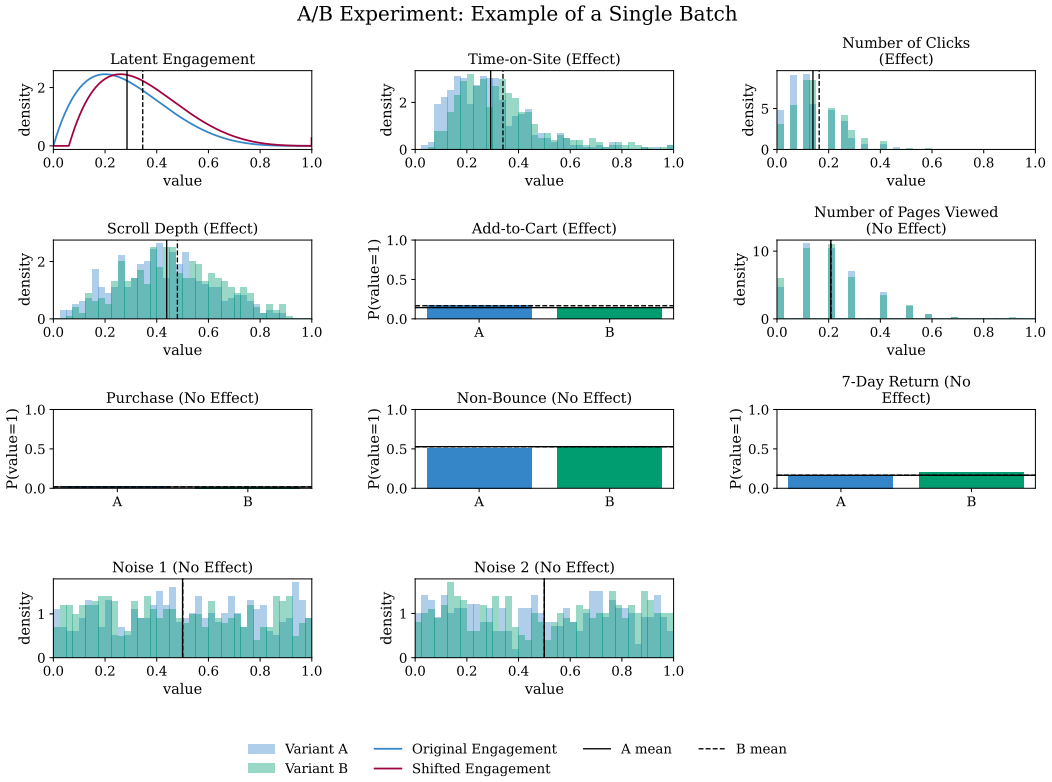


Figure 17: **Session-level distributions within one batch (synthetic A/B web metrics).** We show one representative batch from the *with-effect* simulation. Top-left: A distribution of the latent engagement variable E, and its shifted variant. Other: Affected coordinates (time, clicks, scroll, add-to-cart) show a systematic shift from A to B, while the remaining coordinates (pages, purchase, non-bounce, return-7d, noise_1, noise_2) show little to no shift.

G.6 Model Comparison on Adult

Short Overview We use the Adult dataset to illustrate how a multivariate CS can compare two classifiers while monitoring several performance quantities at once. Model A is the baseline and model B is the challenger. The target decision is whether B improves overall accuracy; however, the same CS also reports uncertainty for the Brier score and for subgroup accuracies. The datasets' features include (among others) gender, education, and age. The task is to predict whether a person's income exceeds \$50,000. Tab. 8 summarizes the empirical performance of the two trained classifiers before applying any CS.

Table 8: **Summary of the two classifiers on the Adult dataset.** Model A is logistic regression and model B is histogram gradient boosting. The table reports the performance of both models on the tracked metrics, together with the mean gain of model B over model A. For overall accuracy and subgroup-specific accuracies, the gain is computed as $B - A$. For the Brier score, since lower values are better, we instead report the improvement as $A - B$, so that a positive value indicates that model B outperforms model A.

Metric	Model A (%)	Model B (%)	Mean gain (%)
Overall accuracy	85.09	87.30	2.24
Overall Brier score	10.25	8.83	1.42
Female accuracy (conditional)	92.84	93.80	0.96
Male accuracy (conditional)	81.19	84.06	2.87
Age < 40 accuracy (conditional)	89.13	90.70	1.57
Age ≥ 40 accuracy (conditional)	79.89	82.98	3.09

The setup is similar to the A/B experiment, where each $\mathbf{\Delta}_t$ is defined as the difference between two metric vectors, with $\mathbf{y}_t^{(A)}$ representing the metrics of model A and $\mathbf{y}_t^{(B)}$ those of model B. The **main technical challenge lies in handling conditional accuracies**, such as $\Pr(f_B(\mathbf{x}) = \text{correct} \mid \mathbf{x} = \text{woman})$ while preserving anytime validity. This is achieved by adding an extra coordinate to $\mathbf{\Delta}_t$ that represents the proportion of the relevant subgroup, for example, $\Pr(\mathbf{x} = \text{woman})$, and then dividing by this quantity.

Setup. The test set is split into 509 batches of size 32 (with the last batch smaller). For a batch I_t, we define the batch-level metric vector for model A as

$$\mathbf{y}_t^{(A)} = \frac{1}{|I_t|} \sum_{i \in I_t} \begin{pmatrix} \mathbf{1}\{f_A(\mathbf{x}_i) = f_i\} \\ (c_A(\mathbf{x}_i) - f_i)^2 \\ \mathbf{1}\{f_A(\mathbf{x}_i) = f_i \text{ and } \mathbf{x}_i \text{ is woman}\} \\ \mathbf{1}\{f_A(\mathbf{x}_i) = f_i \text{ and } \mathbf{x}_i \text{ is man}\} \\ \mathbf{1}\{f_A(\mathbf{x}_i) = f_i \text{ and } \mathbf{x}_i\text{'s age} < 40\} \\ \mathbf{1}\{f_A(\mathbf{x}_i) = f_i \text{ and } \mathbf{x}_i\text{'s age} \geq 40\} \\ \mathbf{1}\{\mathbf{x}_i \text{ is woman}\} \\ \mathbf{1}\{\mathbf{x}_i\text{'s age} < 40\} \end{pmatrix} \in [0,1]^8. \tag{33}$$

Here $f_A(\mathbf{x}_i)$ is model A's predicted class and $c_A(\mathbf{x}_i)$ is its predicted probability for the true class; $\mathbf{y}_t^{(B)}$ is defined analogously. The first coordinate is accuracy, the second is Brier score, the next four are joint subgroup accuracies, and the final two are subgroup proportions. We only track the proportions of women and individuals with age < 40, since the proportions of men and individuals with age ≥ 40 are their complements.

We compare model B to model A by subtracting these batch metrics:

$$\mathbf{\Delta}_t = \begin{pmatrix} y_{t,1}^{(B)} - y_{t,1}^{(A)} \\ y_{t,2}^{(A)} - y_{t,2}^{(B)} \\ y_{t,3}^{(B)} - y_{t,3}^{(A)} \\ \vdots \\ y_{t,6}^{(B)} - y_{t,6}^{(A)} \\ y_{t,7}^{(A)} \\ y_{t,8}^{(A)} \end{pmatrix}.$$

Note that A and B are flipped in the Brier-score coordinate, since larger is better for accuracy but smaller is better for Brier score. Thus positive values in the first six coordinates favor model B. The gain coordinates lie in $[-1, 1]$ and are normalized to $[0, 1]$ before applying the CS, via $(\Delta_{t,d} + 1)/2$. The subgroup-proportion coordinates already lie in $[0, 1]$.

The expectations of the first six coordinates correspond to the overall accuracy gain, the Brier-score improvement, and the joint subgroup-accuracy gains, whereas the final two coordinates correspond to

subgroup proportions. For example, for the "women gain" coordinate (coordinate 3), we have

$$\mathbb{E}[\Delta_{\text{gain woman}}] = \Pr(f_B(\mathbf{x}) = f,\ \mathbf{x} \text{ is woman}) - \Pr(f_A(\mathbf{x}) = f,\ \mathbf{x} \text{ is woman}),$$

and for the woman subgroup proportion (coordinate 7), we have

$$\mathbb{E}[\Delta_{\text{prop woman}}] = \Pr(\mathbf{x} \text{ is woman}).$$

Therefore, the conditional accuracy gain is obtained by dividing by the subgroup probability:

$$\Pr(f_B(\mathbf{x}) = f \mid \mathbf{x} \text{ is woman}) - \Pr(f_A(\mathbf{x}) = f \mid \mathbf{x} \text{ is woman}) = \frac{\mathbb{E}[\Delta_{\text{gain woman}}]}{\Pr(\mathbf{x} \text{ is woman})}. \tag{34}$$

In practice, the quantities in (34) are unknown, so we replace them by their CS bounds. Let

$$[\underline{\mu}_{\text{gain woman}}, \overline{\mu}_{\text{gain woman}}] := \mathcal{I}_t(\Delta_{\text{gain woman}})$$

denote the CS interval for $\mathbb{E}[\Delta_{\text{gain woman}}]$, and analogously let

$$[\underline{\mu}_{\text{prop woman}}, \overline{\mu}_{\text{prop woman}}] := \mathcal{I}_t(\Delta_{\text{prop woman}})$$

denote the interval for $\Pr(\mathbf{x} \text{ is woman})$. Since the denominator is positive, we report the interval for (34) as

$$\left[\min_{\substack{a \in \{\underline{\mu}_{\text{gain woman}}, \overline{\mu}_{\text{gain woman}}\} \\ b \in \{\underline{\mu}_{\text{prop woman}}, \overline{\mu}_{\text{prop woman}}\}}} \frac{a}{b},\ \max_{\substack{a \in \{\underline{\mu}_{\text{gain woman}}, \overline{\mu}_{\text{gain woman}}\} \\ b \in \{\underline{\mu}_{\text{prop woman}}, \overline{\mu}_{\text{prop woman}}\}}} \frac{a}{b} \right].$$

For men, we use the complement $1 - \mathcal{I}_t(\Delta_{\text{prop woman}})$. The same construction is used for the other subgroup gains. Since this is a deterministic transformation of a jointly valid CS, it preserves anytime validity.

Training Setup. Both models are implemented using sklearn's implementation. For logistic regression, categorical features are one-hot encoded and numeric features are standardized; we use `LogisticRegression(solver='lbfgs', max_iter=2000, random_state=7)`. For the boosted-tree model, categorical features are ordinal-encoded with unknown categories mapped to -1, and we use `HistGradientBoostingClassifier(max_depth=6, learning_rate=0.05, max_iter=300, min_samples_leaf=20, random_state=7)`.

Fig. 18 displays the resulting intervals. The hedged-capital regions all detect a positive overall-accuracy gain after roughly 356 batches, while ConfSphere, BanachSphere, and Sub-ψ remain inconclusive after all 509 test batches.

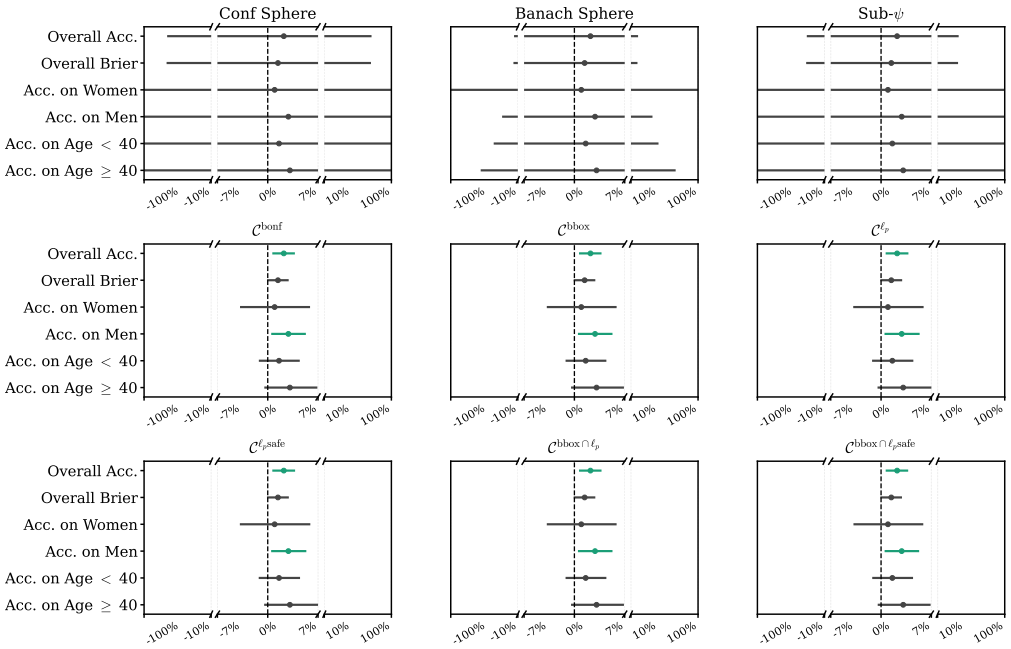


Figure 18: **Model-comparison intervals on the Adult dataset.** Each panel shows one CS method and each row shows one reported metric. Intervals represent the gain for that particular metric, with the dashed vertical line marking zero. Positive values favor the histogram-gradient-boosting challenger (model B) over the logistic-regression baseline (model A). The hedged-capital regions certify a positive overall-accuracy gain, whereas the sphere and normalized-ellipsoid baselines still include zero after all available test batches.

Part IV: Varia

H Additional Related Work

In the context of sequential testing for multi-armed bandits, Cho et al. [14] similarly argue for combining wealth processes by averaging rather than through Bonferroni corrections. Their goal, however, is different: they design sequential tests for bandit decision problems such as best-arm and threshold identification. In doing so, they modify the stake choices (i.e., $\lambda_t(\cdot)$ in § 2) so that the corresponding rejection region is convex, enabling the use of convex optimization tools. In contrast, our focus is on explicitly describing the resulting confidence region through tractable outer approximations that make quantities such as volume or maximum width easy to compute. Convexity alone still does not make these properties straightforward to quantify. Our constructions are also more general, as they apply to any legal stake choice λ_t.

In the context of trustworthy machine learning, Bayesian neural networks (BNNs) provide a principled framework for representing predictive uncertainty, which is particularly valuable in high-stakes applications. Their predictions, however, are commonly approximated using a fixed number of Monte Carlo samples per input, without controlling the resulting error due to the finite number of samples [18–22]. Future work could investigate whether CSs could be applied to control this error as well.

I Broader Impacts

This work aims to improve transparency and reliability in sequential statistical inference, which can support more trustworthy decision-making in applications such as online experimentation and model evaluation. By enabling valid anytime monitoring, the proposed methods can reduce unnecessary data collection and help practitioners detect effects (or lack of effects) earlier. We foresee no direct negative societal impacts from the methods themselves, as they are purely statistical tools that do not

interact with individuals or make decisions on their own. However, as with any statistical method, there is a risk of misuse or misinterpretation.

J Declaration of LLM Usage

LLMs were used to aid during core method development, code writing, and manuscript drafting/editing.

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 (§ 1) state the paper's core claims, and these are matched by the theoretical set-relation results and empirical results in § 6 plus the appendix.

   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: The paper includes an explicit "Limitations and Future Work" paragraph in § 7 discussing discretization cost, conservativeness of the safe bound, and lack of closed-form volume for the intersection region.

   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: Theoretical assumptions are stated in the main text (notably §§2–4), and full proofs and derivations are provided in the appendix sections on validity, set relations, and constructions.

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: The paper provides algorithms, experimental protocols, hyperparameters, stopping rules, and evaluation procedures in § 6 and the appendix's experimental setup, which are sufficient to reproduce the reported experiments. The code is also publicly available and does not require special hardware to run.

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: The code is available via GitHub. It is sufficiently documented to reproduce the main experimental results, and can easily be run by executing a single script. A practical tutorial in the form of a blog post is also included.

   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 experimental setup is specified in § 6 and detailed in the appendix (data generation, dimensions, alpha, repetitions, grid sizes, and method-specific parameters).

   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: Where appropriate, the paper reports confidence intervals. For example, for Monte Carlo volume estimates, such as those in Fig. 9, we report 95% standard-deviation bands. For methods with closed-form solutions, no intervals are reported.

   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: Tab. 2 reports algorithmic complexity and evaluation counts and the appendix includes a dedicated compute-resources subsection (App. G.1.5).

   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 theoretical/empirical on confidence sequences and does not involve procedures that conflict with the NeurIPS Code of Ethics to our knowledge.

   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 appendix includes a dedicated broader-impacts subsection (App. I).

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]

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: Existing assets and datasets are cited (e.g., Adult and prior baselines). Other assets, such as code, are either created by the authors or are synthetic. A license is included in the GitHub repository.

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: [Yes]

    Justification: The only new assets introduced in the paper are code, which is well-documented and publicly available on GitHub.

    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 experiments do not involve crowdsourcing 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 work does not involve human-subject studies requiring IRB or equivalent approval.

    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: [Yes]

    Justification: See App. J.

    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.