BOOTSTRAP-ZOOM: INFERENCE AFTER CHECKPOINT SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Validation data are often used both to select a checkpoint and to estimate its population risk. This reuse can invalidate ordinary confidence intervals, especially when many correlated checkpoints compete. Bootstrap-Zoom extends the zoom correction by replacing its known joint tail with a studentized multiplier-bootstrap tail estimated from the validation loss matrix used for selection. Because the tail, the variance scales, the observed gaps, and the selected checkpoint all depend on one sample, approximation guarantees for fixed regions do not by themselves justify the resulting confidence set. Conditional on a training path fixed before validation, we bound the probability of abstaining or covering from below, accounting for this dependence, finitely many bootstrap draws, and grid discretization. When population standard deviations are bounded away from zero and the checkpoint count and grid satisfy growth conditions, we derive coverage-deficit rates at both positive and zero variance floors. An exact endpoint search returns the same interval as testing every grid value, without assuming that the accepted set is connected. On seven combinations of training setting and loss, we separate the width saved by checkpoint dependence from the width saved by gap discounts. Selected-risk overcoverage can coexist with joint undercoverage, and a rare-loss counterexample shows that positive empirical variances alone do not justify Gaussian calibration.

1 INTRODUCTION

Papers often report the validation loss of the checkpoint that minimizes it. This is an optimistic estimate of that checkpoint's population risk (its expected loss on fresh data). On CIFAR-10 runs whose population risk reaches its minimum early in training, the *uncorrected interval* (the normal-approximation interval that ignores selection) covers 0.639 at nominal 0.90. In this setting $K = 3000$ checkpoints compete on $m = 100$ validation examples, and the interval we develop covers 0.957. Our guarantee conditions on the *training path*, the set of checkpoints fixed before validation, so it excludes validation-based early stopping. The guarantee is asymptotic in m with an unspecified constant, so Section 4 measures coverage at our sample sizes.

A simultaneous *band* covers every checkpoint's risk with one shared threshold in standard errors. Correlation between nearby checkpoints makes this threshold smaller than Bonferroni's. Many checkpoints also have much higher validation loss and could be selected only through a large estimation error. The zoom correction of Zrnic & Fithian (2025) therefore ignores a checkpoint's error at or below a *gap discount* that grows with its validation-loss gap. It sets its threshold from a known bound on the *joint tail*, the probability that some checkpoint's absolute error in standard errors exceeds both the threshold and its discount.

Bootstrap-Zoom (B-Zoom) keeps the gap discounts and estimates the joint tail with a multiplier bootstrap on the same validation losses that select the checkpoint. Its *Gaussian multipliers*, independent standard normal weights on the validation examples, are shared across checkpoints and so preserve the sample correlations of the losses. This captures the dependence that Zrnic & Fithian (2025, Section 6.3) found difficult to characterize on ImageNet and handled with a union bound.

B-Zoom tests each candidate value t of the selected checkpoint's risk. For each t it discounts checkpoints whose validation loss is far above t (Figure 1). It then keeps the largest bootstrap threshold over the checkpoints that plausibly have the smallest population risk.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The sample that selects the checkpoint also estimates the joint tail and sets the gap discounts. A Gaussian approximation for validation errors over fixed regions need not control their probability of falling in a region chosen from the same sample. In a two-coordinate Gaussian example (Section 3), a fresh draw falls in each chosen region with probability 0.90, while the draw that chose it falls there with probability 0.81. Theorem 1 is, to our knowledge, the first coverage guarantee for inference with gap discounts under this reuse.

B-Zoom *abstains* (returns no interval) when the selected checkpoint's validation losses are all equal. Its guarantee bounds the *coverage deficit*, the shortfall below the nominal level $1-\alpha$ of the probability of abstaining or covering. B-Zoom sets its widths from a Gaussian approximation to the joint tail, which we call *Gaussian calibration*.

We make three contributions.

- We prove that B-Zoom is asymptotically valid although its selection, joint tail, scales and gap discounts all come from one validation sample. With population standard deviations bounded away from zero and a grid of candidate values t that refines with m, Theorem 1 bounds the coverage deficit by $O(m^{-1/4})$ up to logarithmic factors. It allows up to $\exp(o(m^{1/5}))$ checkpoints and finitely many bootstrap draws. It needs no positive lower bound on empirical standard deviations (Section 3).
- B-Zoom needs $O(mK + BK)$ memory for B bootstrap draws and never forms or factorizes a $K \times K$ covariance, so thousands of checkpoints are practical. An exact endpoint search returns the same endpoints as testing every grid value, without assuming that the accepted values of t form an interval (Section 2.4). At $K = 3000$ and $m = 1000$ it takes a median of 1.94 seconds on one thread. The zoom correction with a plug-in covariance factorizes a $K \times K$ matrix and was run only up to $K = 600$ on our training settings (Section 4.2).
- We evaluate seven *arms*, each a training setting scored by one loss. At nominal 0.90, B-Zoom covers at least 0.899 in every combination of arm, K and m, while the uncorrected interval falls as low as 0.529 (Section 4.1). Pooled over K and m, its mean width is 13.9% to 29.0% below Bonferroni's (Section 4.2). Dependence between checkpoints supplies most of this saving on six of the seven arms, and the gap discounts add a share that grows with m. Under rare large losses at $m = 100$, B-Zoom with the exact multiplier tail (infinitely many bootstrap draws) covers at most 0.634 at nominal 0.90 (Section 4.3). Every empirical variance is positive there, so positive variance alone cannot justify Gaussian calibration.

2 Setup and Method

For each hypothetical selected risk t, checkpoints far above t are discounted, and the multiplier bootstrap sets the joint threshold for the remaining plausible competitors.

2.1 The Selected Risk

Training produces checkpoints $\theta_1, \dots, \theta_K$ from the training information $\mathcal{G}_{\mathrm{tr}}$, which includes the training sample and optimizer randomness. A quantity is *training-measurable* when $\mathcal{G}_{\mathrm{tr}}$ determines it, as the checkpoints and their number are (Appendix B.1).

A loss $\ell(z, \theta) \in [0, M]$, with M known, scores example z at checkpoint θ. Independently of training, $m \geq 2$ validation examples $z'_1, \dots, z'_m$ are drawn i.i.d. from a distribution $\mathcal{D}$. The *validation loss matrix* L and its column statistics are

$$L_{jk} = \ell(z'_j, \theta_k), \quad \hat{\mu}_k = m^{-1} \sum_j L_{jk}, \quad \tilde{L}_{jk} = L_{jk} - \hat{\mu}_k, \quad \hat{\sigma}_k^2 = m^{-1} \sum_j \tilde{L}_{jk}^2.$$

Checkpoint k has population risk $\mu_k = \mathbb{E}[L_{1k} \mid \mathcal{G}_{\mathrm{tr}}]$ and population variance $\sigma_k^2 = \mathrm{Var}(L_{1k} \mid \mathcal{G}_{\mathrm{tr}})$. Validation selects $\hat{k} = \min \arg\min_k \hat{\mu}_k$, the smallest index among the minimizers. Our target is the *selected risk* $\mu_{\hat{k}}$, and we seek a confidence set for it with coverage $1-\alpha$ for an error level $\alpha \in (0, 1)$.

Write $[K] = \{1, \dots, K\}$. Let $Z = \{k : \hat{\sigma}_k = 0\}$ index the constant columns and $P = [K] \setminus Z$ the rest. The procedure abstains when $\hat{k} \in Z$. Otherwise it uses the scales $s_k = \max(\hat{\sigma}_k, \sigma_{\min})$ for $k \in P$, with a training-measurable *floor* $\sigma_{\min} \geq 0$. Experiments use $\sigma_{\min} = 0$.

2.2 A SHARED ESTIMATE OF THE JOINT TAIL

The *studentized errors* $\sqrt{m}(\hat{\mu}_k - \mu_k)/s_k$ express estimation errors in units of $s_k/\sqrt{m}$. We approximate their joint law with i.i.d. multipliers $e_1, \dots, e_m \sim N(0,1)$, independent of the data and shared across checkpoints. For $k \in P$ and thresholds $v \in [0,\infty)^P$, form

$$G_k^* = m^{-1} \sum_j e_j \tilde{L}_{jk}, \qquad T_k^* = \sqrt{m}\, G_k^*/s_k, \qquad \hat{S}(v) = \mathbb{P}_e\{\exists k \in P : |T_k^*| > v_k\},$$

where $\mathbb{P}_e$ is probability over the multipliers given L and $\mathcal{G}_{\mathrm{tr}}$. The function $\hat{S}(v)$ estimates the probability that some checkpoint k has a studentized error larger than v_k in absolute value. Given $(L, \mathcal{G}_{\mathrm{tr}})$, the vector $(T_k^*)_{k\in P}$ is centered Gaussian with covariance

$$\hat{R}_{kl} = \frac{m^{-1} \sum_j \tilde{L}_{jk} \tilde{L}_{jl}}{s_k s_l}, \qquad k, l \in P,$$

so T^* inherits the sample covariance of the losses.

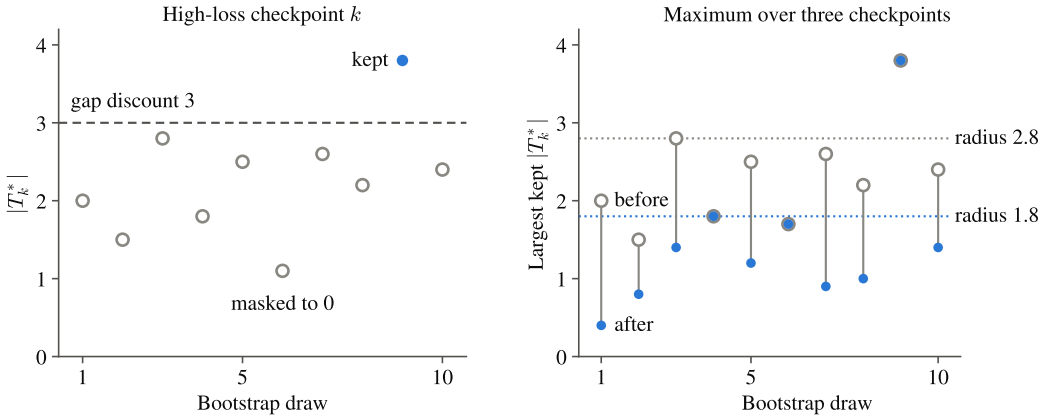

Figure 1: Discounting one high-loss checkpoint lowers the threshold (radius) from 2.8 to 1.8 at $\alpha = 0.2$. With gap discount 3, it survives masking in only draw 9 of ten illustrative bootstrap draws. The right panel takes each draw's maximum over this checkpoint and two undiscounted ones, before and after masking, and each radius is the ninth smallest of the ten maxima.

B-Zoom replaces a threshold r shared by all checkpoints, as in the band, by the larger of r and a data-dependent discount y_k at each checkpoint k. *Masking* by $y \in [0,\infty)^P$ sets each $|T_k^*| \le y_k$ to zero (Figure 1) and leaves the *masked maximum*

$$\max_{k\in P} |T_k^*|\, \mathbf{1}\{|T_k^*| > y_k\}.$$

Its survival function at $r \ge 0$ is $\hat{S}(r\mathbf{1} \vee y)$, with $\vee$ the coordinate-wise maximum and $\mathbf{1}$ the all-ones vector, so its smallest $(1-\alpha)$ quantile under $\mathbb{P}_e$ is the *exact-tail radius*

$$\psi(y) = \inf\{r \ge 0 : \hat{S}(r\mathbf{1} \vee y) \le \alpha\}.$$

The band's radius is $r_{\mathrm{sim}} = \psi(0)$.

Computation replaces $\mathbb{P}_e$ by $B \ge \lceil 1/\alpha \rceil - 1$ independent bootstrap draws of the multiplier vector. Draw b gives the values T_{bk}^* of T_k^*, stored as $A_{bk} = |T_{bk}^*|$ in one $B \times |P|$ array A. The *B-draw radius* $\psi_B(y)$ is the masked row maximum of A with zero-based rank

$$k_\alpha(B) = \lceil (1-\alpha)(B+1) \rceil - 1$$

in increasing order (Appendix A.2). Masking can only lower each maximum, so $\psi(y) \le r_{\mathrm{sim}}$ and $\psi_B(y) \le \psi_B(0)$ for every y.

2.3 FROM GAP DISCOUNTS TO AN INTERVAL

To test a candidate value t for $\mu_{\hat{k}}$, B-Zoom posits a *reference* $i \in P$ as a *population winner* (a checkpoint of smallest population risk) with gap discounts

$$g_j^{(i)}(t) = \frac{\sqrt{m}(\hat{\mu}_j - t)_+}{2s_j + s_i}, \quad j \in P \setminus \{\hat{k}, i\}, \qquad g_{\hat{k}}^{(i)}(t) = g_i^{(i)}(t) = 0, \tag{1}$$

where $x_+ = \max(x, 0)$. If reference i has risk t and $\mu_j = t + \Delta$, a confidence set for all checkpoint risks lets $\hat{\mu}_j$ exceed μ_j by $\Delta s_j/(s_j + s_i)$ (Appendix A.1). The discount is $\sqrt{m}\,\Delta/(s_j + s_i)$ for the smallest $\Delta \geq 0$ at which μ_j plus this allowance reaches $\hat{\mu}_j$ (Appendix B.4).

Because the population winner is unknown, B-Zoom maximizes over a set $W(t)$ of plausible references,

$$W(t) = \{i \in P : \hat{\mu}_i - t \leq r_{\mathrm{sim}} s_i/\sqrt{m}\} \cup \{\hat{k}\}, \qquad \hat{r}_\alpha^W(t) = \max_{i \in W(t)} \psi(g^{(i)}(t)). \tag{2}$$

Including $\hat{k}$ keeps $W(t)$ nonempty. At $t = \mu_{\hat{k}}$, a population winner $i \in P$ lies outside $W(t)$ only if its studentized error exceeds r_{sim}, and one inside gives $\hat{r}_\alpha^W(t) \geq \psi(g^{(i)}(t))$. Without abstention, the accepted set is

$$\hat{C}_\alpha^W = \{t \in \mathbb{R} : |\hat{\mu}_{\hat{k}} - t| \leq \hat{r}_\alpha^W(t)\, s_{\hat{k}}/\sqrt{m}\}.$$

At a positive floor, it contains every value of $\mu_{\hat{k}}$ allowed by the confidence set of Appendix A.1 (Theorem 10).

Since $\psi(y) \leq r_{\mathrm{sim}}$, every accepted t lies in the *search interval*, the band's interval for $\mu_{\hat{k}}$:

$$\hat{C}_\alpha^W \subseteq [\hat{\mu}_{\hat{k}} \pm r_{\mathrm{sim}} s_{\hat{k}}/\sqrt{m}]. \tag{3}$$

As t rises, the discounts mask no more entries and $W(t)$ only grows, so $\hat{r}_\alpha^W(t)$ is non-decreasing in t. The accepted set is therefore an interval below $\hat{\mu}_{\hat{k}}$ but can be disconnected above it (Appendix B.4).

On a finite grid $\mathcal{T}$ whose hull (smallest interval containing it) contains the search interval, we report the *exact-tail interval*

$$I_m = \mathrm{hull}(\hat{C}_\alpha^W \cap \mathcal{T}). \tag{4}$$

Its computed version replaces ψ by ψ_B throughout, with $\psi_B(0)$ in place of r_{sim}. If no grid value is accepted, we report a *grid failure* and count it as noncoverage.

2.4 COMPUTATION

Algorithm 1 takes its grid from a *grid rule*, a measurable map from the loss matrix and the multipliers to a finite grid. It reuses A at every candidate value, since fresh draws would define a different procedure.

As written, Algorithm 1 is the *full sweep*, which tests every grid value. References with equal scales share work in its step 4 (Appendix A.3). The *exact endpoint search* computes the maximum over all of $W(t)$ at each tested value and returns the same endpoints after testing at most 209 of our 400 grid values (Appendix C.7). It bisects below $\hat{\mu}_{\hat{k}}$ and scans downward above it, where the accepted set need not be connected. At $K = 3000$ and $m = 1000$ it takes a median of 1.94 seconds on one thread, 5.3 times faster than the full sweep. Drawing multipliers in blocks, B-Zoom stores L and A in $O(mK + BK)$ memory and never forms or factorizes a $K \times K$ covariance (Appendix A.3).

3 COVERAGE GUARANTEE

Informally, if every σ_k is at least a fixed fraction of M and the grid refines with m, then up to $\log(K \vee m)$ factors

$$\mathbb{P}\big(\text{B-Zoom abstains or } \mu_{\hat{k}} \in I_m \mid \mathcal{G}_{\mathrm{tr}}\big) \geq 1 - \alpha - O(m^{-1/4}).$$

Formally, let $K_{\max}(m)$ be a deterministic sequence with $K \leq K_{\max}(m)$ almost surely, and write

$$\Lambda_{\max} = \log(K_{\max}(m) \vee m), \qquad q_{\max} = \Lambda_{\max}^5/m.$$

Let $h(m, k)$ be a deterministic bound, valid on every realization with $K = k$, on the largest spacing of a grid that may depend on the data.

Input. $L \in [0, M]^{m \times K}$, error level α, $B \geq \lceil 1/\alpha \rceil - 1$, floor $\sigma_{\min} \geq 0$, and a grid rule.

1. Compute $\hat{\mu}_k, \hat{\sigma}_k$ and select $\hat{k} = \min \arg\min_k \hat{\mu}_k$. Abstain if $\hat{\sigma}_{\hat{k}} = 0$. Retain the nonconstant columns P and set $s_k = \max(\hat{\sigma}_k, \sigma_{\min})$.
2. Draw B shared Gaussian multiplier vectors and form $A_{bk} = |T^*_{bk}|$. Set $k_\alpha(B) = \lceil (1-\alpha)(B+1) \rceil - 1$ and take $\psi_B(0)$ as the zero-based $k_\alpha(B)$th order statistic of the row maxima.
3. Form a grid spanning $[\hat{\mu}_{\hat{k}} \pm \psi_B(0) s_{\hat{k}} / \sqrt{m}]$. For each grid value t, form $W(t)$ with $\psi_B(0)$.
4. For every $i \in W(t)$, compute the discounts (1). In each row of A, zero the entries at or below their discounts. The zero-based $k_\alpha(B)$th order statistic of the masked row maxima is $\psi_B(g^{(i)}(t))$.
5. Accept t when $\sqrt{m}\,|\hat{\mu}_{\hat{k}} - t| / s_{\hat{k}} \leq \max_{i \in W(t)} \psi_B(g^{(i)}(t))$. Return the smallest and largest accepted grid values. If no value is accepted, report a grid failure.

Algorithm 1: B-Zoom from a validation loss matrix.

Theorem 1 (Coverage with population variances bounded away from zero)**.** *Suppose $K \geq 2$, $q_{\max} \to 0$, the validation rows are conditionally i.i.d. in $[0, M]^K$, and $\min_k \sigma_k \geq \xi M$ almost surely for a fixed $\xi \in (0, 1/2]$. Fix $\alpha \in (0,1)$ and a finite $B \geq \lceil 1/\alpha \rceil - 1$. Set $a = \xi M/2$. Require the endpoints of each realized grid to span the search interval (3), with $\psi_B(0)$ in place of r_{sim} when the radius is ψ_B, and require*

$$\sup_{2 \leq k \leq K_{\max}(m)} \frac{C_k \sqrt{m}\, h(m,k)}{a} = o(q_{\max}^{1/4}), \qquad C_k = \sqrt{2\log(2k)} + 2. \tag{5}$$

Use the floor $\sigma_{\min} = a$ or $\sigma_{\min} = 0$. Then, for all sufficiently large m,

$$\mathbb{P}\big(\{\hat{k} \in Z\} \cup \{\hat{k} \notin Z,\ \mu_{\hat{k}} \in I_m\} \mid \mathcal{G}_{\text{tr}}\big) \geq 1 - \alpha - O_\xi(\Lambda_{\max}^{5/4} m^{-1/4}).$$

The bound holds for both radii ψ and ψ_B, with a constant that depends only on ξ. The sample size needed depends only on ξ, α, $K_{\max}$ and h, so the bound is uniform over training paths and averaging over training gives the same rate.

The spacing bound $h(m,k) = a/(m \log k)$ meets condition (5), and $q_{\max} \to 0$ is equivalent to $\log K_{\max}(m) = o(m^{1/5})$, which allows $K_{\max} = m^2$. Appendix B.8 derives Theorem 1 from Theorem 13, a bound at any positive training-measurable floor $a \leq \min_k \sigma_k$.

How the proof handles validation reuse. The proof reduces coverage to the event that the studentized errors $\sqrt{m}(\hat{\mu}_k - \mu_k)/s_k$ lie in a rectangle, a product of intervals whose endpoints depend on the sample through the radius and the scales. Data-dependent endpoints matter even when the errors are exactly Gaussian. Let X_1, X_2 be independent standard Gaussians, and constrain the coordinate with larger $|X_i|$ to $[-z_{0.95}, z_{0.95}]$, where z_p is the standard Gaussian p-quantile, leaving the other free. An independent draw lies in each realized rectangle with probability 0.90, but the sample lies in its own rectangle with probability

$$\mathbb{P}\big(\max_i |X_i| \leq z_{0.95}\big) = 0.90^2 = 0.81.$$

A bound uniform over fixed rectangles therefore does not apply to this event, and Lemma 4 states when one does.

The proof instead places the errors divided by population standard deviations in a rectangle that training fixes. It proceeds in four steps.

1. Empirical scales and correlations lie near their population values with conditional probability at least $1 - 1/m$.
2. A Gaussian comparison and a bound on Gaussian probability near the rectangle's boundary supply the fixed rectangle, to which a central limit theorem applies.
3. The scale bounds turn this event into acceptance of $\mu_{\hat{k}}$ by the test whose reference is the population winner, and the maximum over $W(t)$ is at least that test's radius (Theorem 10).

4. Shrinking the fixed rectangle by one grid spacing leaves accepted grid values on both sides of $\mu_{\hat{k}}$.

Finite B adds no error term, and the zero floor costs $1/m$ (Appendices B.7 and B.8).

Scope of Theorem 1. The bound conditions on $\mathcal{G}_{\mathrm{tr}}$ and averages over validation samples and bootstrap draws, without conditioning on returning an interval or on the selected index. The checkpoints may be any training-measurable models, such as models with different hyperparameters. A grid with a fixed number of points, such as our 400-point grid, cannot meet condition (5) as m grows.

4 Experiments

We evaluate B-Zoom on seven arms. The *main arm* trains a ResNet-8 for 3000 SGD steps on 25,000 CIFAR-10 examples. The *interior-optimum arm* trains the same model on 5000 examples without augmentation, and its population risk reaches its minimum early in training. CIFAR-10 with 20% label noise and clean CIFAR-100 each yield a cross-entropy arm and an *error* arm with 0–1 loss, and the seventh arm fine-tunes a CIFAR-100 model on CIFAR-10 (Appendix E.1).

We compare B-Zoom with three intervals for $\mu_{\hat{k}}$: the uncorrected interval, *Bonferroni* and the band's interval computed from B-Zoom's bootstrap draws,

$$\hat{\mu}_{\hat{k}} \pm z_{1-\alpha/2}\frac{s_{\hat{k}}}{\sqrt{m}}, \qquad \hat{\mu}_{\hat{k}} \pm z_{1-\alpha/(2K)}\frac{s_{\hat{k}}}{\sqrt{m}}, \qquad \hat{\mu}_{\hat{k}} \pm \psi_B(0)\frac{s_{\hat{k}}}{\sqrt{m}}.$$

Sample splitting selects on a fixed random half of the rows and forms the uncorrected interval on the other half.

Protocol. By default, $\alpha = 0.10$, $B = 2000$, $\sigma_{\min} = 0$, and the grid $\mathcal{T}$ has 400 equally spaced points spanning 1.02 times the search interval. A *case* is one validation sample of m rows drawn with replacement from a pool of 25,000 examples held out from training, and every method receives the same matrix in it. Each arm has 100 training seeds, $K \in \{30, 150, 600, 3000\}$ and $m \in \{100, 1000, 5000\}$. Each seed contributes ten cases per *cell* (a (K, m) pair of an arm or one setting of a synthetic design), giving 12,000 cases per arm.

Each interval targets the selected risk, computed as the selected checkpoint's mean stored loss over the pool. Cross-entropy is capped at $M = 10$, and 0–1 loss has $M = 1$. Coverage limits are 95% Student-t intervals over per-seed averages on real arms and exact Clopper–Pearson limits on synthetic designs. Appendix E gives further settings.

4.1 Coverage after selection

Figure 2 shows B-Zoom covering at least 0.899 in all 84 cells. At $m = 100$ the uncorrected interval covers less at $K = 3000$ than at $K = 30$ in every arm, and across arms it covers 0.801 to 0.897 at $K = 30$ and 0.529 to 0.885 at $K = 3000$. B-Zoom returns an interval in all 84,000 real-arm cases, so no abstention enters its coverage.

Table 1 pools over K on the main, interior-optimum and CIFAR-100 error arms and adds widths. B-Zoom covers at least 0.955 in every row, and Section 4.2 decomposes the widths.

Bonferroni, the band and B-Zoom over-cover in Table 1. Their constructions account for competition across checkpoints, which can make selected-risk coverage conservative. At $K = 600$ and $m = 100$, the band covers the selected risk in 0.980 of 200 main-arm cases. Its *joint rectangle*, the product of its intervals over all checkpoints, covers every checkpoint's risk at once in only 0.760, with 95% limits $[0.697, 0.823]$ below the nominal 0.90 (Appendix C.2).

Sample splitting covers 0.871 to 0.908 and is narrower than B-Zoom on all three arms at $m \leq 1000$. It targets the checkpoint selected on half the rows, whose population risk exceeds that of the one selected on all m rows by the split cost of Table 1 on average.

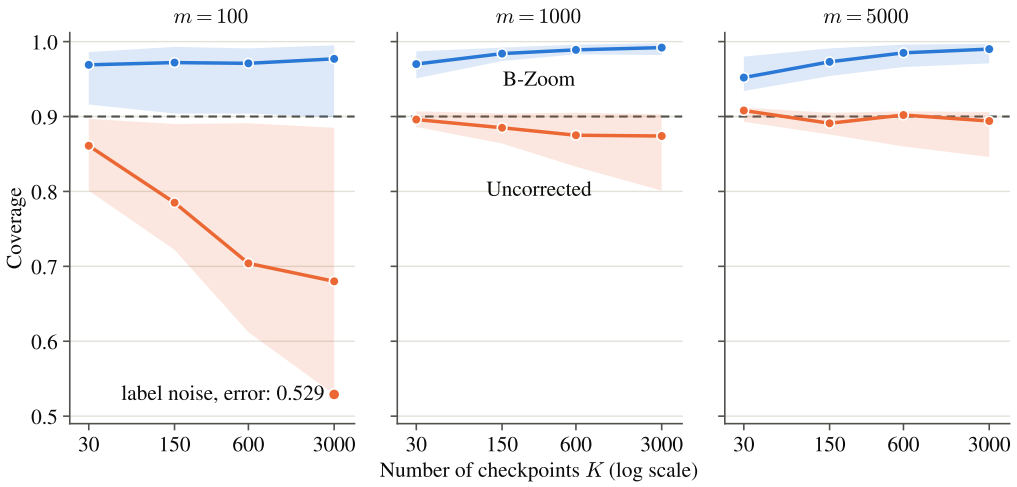

Figure 2: B-Zoom covers at least 0.899 in all 84 cells of the seven real arms, while the uncorrected interval falls as low as 0.529 at $m = 100$ and $K = 3000$. Lines give the median over arms and bands the range over arms, and each cell pools 1000 cases from 100 training seeds. The dashed line marks nominal coverage 0.90.

Table 1: Coverage and mean width at nominal 0.90 on three arms, each row pooling 4000 cases over $K \in \{30, 150, 600, 3000\}$. Bold marks coverage whose 95% upper limit lies below 0.90, and coverage limits have half-width at most 0.024. Split cost is the mean increase in population risk from selecting on half the rows instead of all m. Unc. is uncorrected and Bonf. is Bonferroni.

		Coverage at nominal 0.90					Mean width				Split
Arm	m	Unc.	Split	Bonf.	Band	B-Zoom	Split	Bonf.	Band	B-Zoom	cost
Main	100	**0.839**	**0.871**	0.986	0.974	0.972	0.470	0.677	0.572	0.554	0.0221
	1000	0.902	0.898	0.999	0.997	0.985	0.149	0.227	0.193	0.160	0.0016
	5000	0.905	0.908	1.000	0.998	0.973	0.067	0.102	0.087	0.062	0.0003
Interior optimum	100	**0.730**	0.888	0.980	0.956	0.955	0.693	0.900	0.756	0.749	0.0578
	1000	**0.879**	0.900	0.999	0.994	0.992	0.193	0.289	0.244	0.225	0.0067
	5000	0.891	0.900	1.000	0.997	0.975	0.085	0.130	0.109	0.085	0.0012
CIFAR-100, error	100	**0.752**	0.890	0.995	0.992	0.991	0.229	0.349	0.321	0.305	0.0142
	1000	**0.862**	0.896	0.997	0.994	0.984	0.073	0.112	0.104	0.087	0.0022
	5000	0.886	0.908	0.999	0.998	0.984	0.033	0.050	0.047	0.035	0.0006

4.2 Where the efficiency comes from

The band's saving over Bonferroni comes from dependence between checkpoints and is most of B-Zoom's saving on six of seven arms (Table 2). These width comparisons fix the nominal level, and achieved coverages differ. The *grid band* restricts the band's interval for $\mu_{\hat{k}}$ to the hull of the grid values inside it, and *grid rounding* is the width this restriction saves. *Localization* is the width that B-Zoom's gap discounts save beyond the grid band. In mean widths w, the saving over Bonferroni splits as

$$\underbrace{w_{\text{Bonf}} - w_{\text{B-Zoom}}}_{\text{total}} = \underbrace{w_{\text{Bonf}} - w_{\text{band}}}_{\text{dependence}} + \underbrace{w_{\text{band}} - w_{\text{grid}}}_{\text{grid rounding}} + \underbrace{w_{\text{grid}} - w_{\text{B-Zoom}}}_{\text{localization}},$$

where w_{grid} is the grid band's width.

On every arm and at every K, localization's share of the grid band's width grows with m. On the main arm and across the four values of K, this share is 2.7–4.0% at $m = 100$ and 26.6–29.9% at $m = 5000$ (Table 4 in Appendix C.1). At $m = 5000$ on that arm, B-Zoom covers 0.949–0.987, against the band's 0.991–1.000.

Table 2: Width saving of B-Zoom over Bonferroni on all 12,000 cases per arm, in percent of the mean Bonferroni width, split into dependence (Bonferroni minus the band) and localization (the grid band minus B-Zoom). The two parts sum to the total up to grid rounding, which is 0.033 to 0.042 percentage points on every arm. The last column divides the mean localization saving by the mean grid-band width, and CE denotes cross-entropy.

Arm	Total	Dependence	Localization	Localization, % of grid band
CIFAR-10 main	22.86	15.38	7.44	8.79
Interior optimum	19.66	15.96	3.66	4.35
Label noise, CE	26.08	16.99	9.05	10.91
Label noise, error	13.89	8.66	5.19	5.68
CIFAR-100, CE	28.94	16.55	12.36	14.81
CIFAR-100, error	16.38	7.77	8.57	9.30
Fine-tuning	28.97	25.98	2.95	3.99

The coverage proof needs the maximum over $W(t)$ in (2). Against the *single-reference ablation* $W(t) = \{\hat{k}\}$, the maximum raises mean width by 0.39–0.78% on the main and interior-optimum arms at $K = 600$ and $m \in \{100, 1000\}$ (Appendix C.3). The *zoom baseline*, the correction of Zrnic & Fithian (2025) with reference $\hat{k}$ and Gaussian draws from the estimated covariance, is 0.03–0.56% narrower than B-Zoom where both return an interval. On those cases the two coverages differ by at most 0.003. The zoom baseline factorizes that $K \times K$ covariance and was run only up to $K = 600$ on the real arms (Table 5 in Appendix C.1). It returns no interval when any column is constant, which happens in 1820 of the 9000 cases with $K \leq 600$ on the CIFAR-100 error arm.

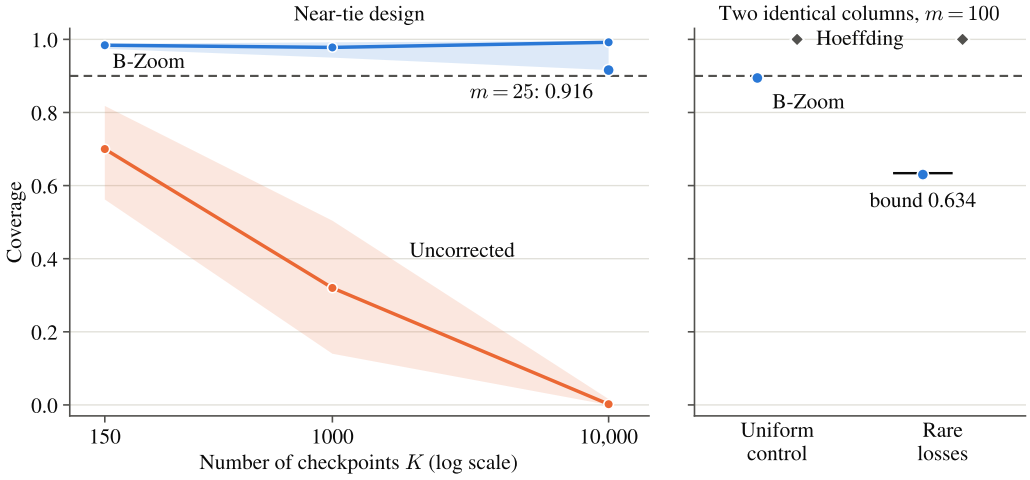

Figure 3: Gaussian calibration survives near ties at K up to 10,000 and a uniform control with two identical Uniform$[0, 1]$ columns. It fails under rare losses, where B-Zoom covers 0.630 against the bound 0.634 for the exact-tail interval. On the left, lines give the median over the five validation sizes m and bands their range, from 500 replications per cell. On the right, Hoeffding is the union-bound interval, and Clopper–Pearson 95% limits over 5000 replications are narrower than the markers, $[0.886, 0.903]$ for B-Zoom on the uniform control.

4.3 Where Gaussian calibration fails

Rare large losses defeat Gaussian calibration even when every empirical variance is positive. The *rare-loss design* has $m = 100$, $M = 1$ and two identical loss columns

$$L_{j1} = L_{j2} = (1 - \epsilon)A_j + \epsilon U_j, \qquad \epsilon = 10^{-4}, \tag{6}$$

with independent $A_j \sim \text{Bernoulli}(0.01)$ and $U_j \sim \text{Uniform}[0, 1]$. The small term ϵU_j makes both empirical variances positive almost surely. With probability $0.99^{100} \approx 0.366$ every A_j is zero, and

the exact-tail interval then lies below the population risk. Its coverage is therefore at most 0.634 at every floor, and B-Zoom with $B = 2000$ draws covers 0.630 (Figure 3, right panel). The design meets the variance condition of Theorem 1 with $\xi = 0.099$, and the probability 0.99^m that every A_j is zero vanishes as m grows.

The *Hoeffding union-bound interval*

$$[\hat{\mu}_{\hat{k}} \pm r_{\mathrm{H}}] \cap [0, M], \qquad r_{\mathrm{H}} = M\sqrt{\log(2K/\alpha)/(2m)},$$

covers the selected risk with probability at least $1 - \alpha$ for every loss law on $[0, M]$ (Appendix C.4). In the rare-loss design it covers every replication at 5.8 times B-Zoom's mean width.

Gaussian calibration survives the *near-tie design*, which pairs nearly tied checkpoint risks with unequal loss variances at K up to $10{,}000$ and m down to 25 (Appendix C.5). There B-Zoom covers at least 0.916 at every m, while the uncorrected interval covers at most 0.016 at $K = 10{,}000$. Appendix D varies B, the multiplier law, the floor and α on this design, and the left panel of Figure 3 maps it.

5 Related work and discussion

Simultaneous inference protects any later selection by covering all candidates at once (Miller, 1981; Berk et al., 2013), and locally simultaneous inference focuses this protection around the selected candidate (Zrnic & Fithian, 2024). The zoom correction tailors it to the selected candidate through validation-loss gaps and a known bound on the joint tail (Zrnic & Fithian, 2025). Gaussian approximations to maxima of high-dimensional sums justify the multiplier bootstrap uniformly over fixed rectangles (Chernozhukov et al., 2013; 2017a; 2022). B-Zoom estimates the joint tail with this bootstrap from the loss matrix that selects the checkpoint, and Theorem 1 bounds its coverage deficit under this reuse.

Post-selection inference can condition on the selection event (Lee et al., 2016) or split each data point (Leiner et al., 2025). Conditional and hybrid methods condition on which candidate was selected, alone or combined with simultaneous protection, under a known or consistently estimable covariance and a fixed number of candidates (Andrews et al., 2024; McCloskey, 2024; Bakshi et al., 2026). On one synthetic design where the uncorrected interval already covers near nominal, the hybrid interval of Andrews et al. (2024) with shrinkage covariance is narrower than B-Zoom and covers closer to nominal (Appendix C.6). B-Zoom never forms a $K \times K$ covariance, and Theorem 1 lets the number of checkpoints grow with m.

Cross-validation studies measure and correct the downward bias of the minimum cross-validation error (Varma & Simon, 2006; Tibshirani & Tibshirani, 2009; Cawley & Talbot, 2010). The checkpoint chosen on validation often has lower test accuracy than the one with the best test accuracy (Apicella et al., 2026), and validation overfitting also appears in hyperparameter optimization (Schneider et al., 2025). Differential privacy lets a holdout set be reused across adaptive analyses (Dwork et al., 2015), and studies of reused benchmarks find little evidence of adaptive overfitting (Recht et al., 2019; Roelofs et al., 2019).

A reality check tests whether the best of many models beats a benchmark (White, 2000), and a model confidence set contains the best model with a chosen probability (Hansen et al., 2011; Lei, 2020). SIREN targets the fresh-data performance of a selection rule made less sensitive to validation noise (Xu et al., 2026). Guan (2018) proposes bootstrap intervals for test error averaged over the selection rule with a fixed number of candidates, while we let the checkpoint count grow as $\exp(o(m^{1/5}))$.

Discussion. Under the conditions of Theorem 1, the coverage deficit of B-Zoom vanishes as m grows. When rare large losses are plausible, the Hoeffding union-bound interval, chosen before the validation outcomes are seen, has coverage at least $1 - \alpha$ at every sample size (Appendix C.4).

AI USE STATEMENT

We used generative AI tools to draft and revise prose. We also used them for coding experiments. We did not use them to generate research ideas or to produce experimental results. We used AI for improving the rigor and presentation clarity of the proofs. We reviewed all AI-assisted work, and we take responsibility for the final content of this paper.

ETHICS STATEMENT

This work studies confidence intervals for the risk of a checkpoint selected on validation data. Its experiments use CIFAR-10 and CIFAR-100 under their published terms, with no human subjects and no new dataset. The intervals target the population risk under the validation distribution $\mathcal{D}$ and give no guarantee under distribution shift.

REPRODUCIBILITY STATEMENT

Section 4 and Appendix E specify every experimental setting, from training and seed rules to the pool that defines each target. Appendix E.3 describes the released code and per-replication results. Appendix A.1 proves Proposition 2, and Appendix B contains every other proof under the conventions of Appendix A.

REFERENCES

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on winners. *The Quarterly Journal of Economics*, 139(1):305–358, 2024. doi: 10.1093/qje/qjad043.

Andrea Apicella, Francesco Isgrò, Andrea Pollastro, and Roberto Prevete. Don't stop me now: Rethinking validation criteria for model parameter selection, 2026. arXiv:2602.22107.

Soham Bakshi, Lingjun Gao, Zijun Gao, and Snigdha Panigrahi. Flexible inference for winners with conditional validity, 2026. arXiv:2607.18545v1, 20 July 2026.

Yoav Benjamini and Daniel Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005. doi: 10.1198/016214504000001907.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. doi: 10.1214/12-AOS1077.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL https://www.jmlr.org/papers/v11/cawley10a.html.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41 (6):2786–2819, 2013. doi: 10.1214/13-AOS1161.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352, 2017a. doi: 10.1214/16-AOP1113.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Detailed proof of Nazarov's inequality, 2017b. arXiv:1711.10696.

Victor Chernozhukov, Denis Chetverikov, Kengo Kato, and Yuta Koike. Improved central limit theorem and bootstrap approximations in high dimensions. *The Annals of Statistics*, 50(5):2562–2586, 2022. doi: 10.1214/22-AOS2193.

D. R. Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2): 441–444, 1975. doi: 10.1093/biomet/62.2.441.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi: 10.1126/science.aaa9375.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection, 2014. arXiv:1410.2597v4, 18 April 2017.

Leying Guan. Test error estimation after model selection using validation error, 2018. arXiv:1801.02817.

Peter R. Hansen, Asger Lunde, and James M. Nason. The model confidence set. *Econometrica*, 79 (2):453–497, 2011. doi: 10.3982/ECTA5771.

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

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03) 00096-4.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016. doi: 10.1214/ 15-AOS1371.

Jing Lei. Cross-validation with confidence. *Journal of the American Statistical Association*, 115 (532):1978–1997, 2020. doi: 10.1080/01621459.2019.1672556.

James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data fission: Splitting a single data point. *Journal of the American Statistical Association*, 120(549):135–146, 2025. doi: 10. 1080/01621459.2023.2270748.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Adam McCloskey. Hybrid confidence intervals for informative uniform asymptotic inference after model selection. *Biometrika*, 111(1):109–127, 2024. doi: 10.1093/biomet/asad023.

Rupert G. Miller, Jr. *Simultaneous Statistical Inference*. Springer Series in Statistics. Springer-Verlag, New York, 2nd edition, 1981.

Lutz Prechelt. Early stopping – but when? In Genevieve B. Orr and Klaus-Robert Müller (eds.), *Neural Networks: Tricks of the Trade*, Lecture Notes in Computer Science, pp. 55–69. Springer, Berlin, Heidelberg, 1998. doi: 10.1007/3-540-49430-8_3.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400, 2019. URL https://proceedings.mlr.press/v97/recht19a.html. arXiv:1902.10811.

Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005. doi: 10.1111/j.1468-0262.2005.00615.x.

Lennart Schneider, Bernd Bischl, and Matthias Feurer. Overtuning in hyperparameter optimization. In *Proceedings of the Fourth International Conference on Automated Machine Learning*, volume 293 of *Proceedings of Machine Learning Research*, pp. 17/1–43. PMLR, 2025. URL https://proceedings.mlr.press/v293/schneider25a.html.

Xiaoying Tian and Jonathan Taylor. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018. doi: 10.1214/17-AOS1564.

Ryan J. Tibshirani and Robert Tibshirani. A bias correction for the minimum error rate in cross-validation. *The Annals of Applied Statistics*, 3(2):822–829, 2009. doi: 10.1214/08-AOAS224.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006. doi: 10.1186/1471-2105-7-91.

Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000. doi: 10.1111/1468-0262.00152.

Yang Xu, Jiefu Zhang, Haixiang Sun, Zihan Zhou, Tianyu Cao, and Vaneet Aggarwal. Towards reliable LLM evaluation: Correcting the winner's curse in adaptive benchmarking, 2026. arXiv:2605.05973v1, 7 May 2026.

Tijana Zrnic and William Fithian. Locally simultaneous inference. *The Annals of Statistics*, 52(3): 1227–1253, 2024. doi: 10.1214/24-AOS2391.

Tijana Zrnic and William Fithian. A flexible defense against the winner's curse. *The Annals of Statistics*, 53(6):2516–2535, 2025. doi: 10.1214/25-AOS2553. Appendices A–H are in the supplement, doi 10.1214/25-AOS2553SUPP.

Tijana Zrnic and Michael I. Jordan. Post-selection inference via algorithmic stability. *The Annals of Statistics*, 51(4):1666–1691, 2023. doi: 10.1214/23-AOS2303.

GUIDE TO THE APPENDIX

Each row pairs a claim in the main text with the appendix material that supports it.

Claim in the main text	Supporting material
At a positive floor, every selected-risk value allowed by the confidence set for all checkpoint risks lies in $\hat{C}_\alpha^W$ (Section 2.3)	Appendix A.1, Theorem 10
Theorem 1, its four proof steps and its positive-floor form (Section 3)	Appendix B, Theorem 13
A Gaussian bound uniform over fixed rectangles holds at a rectangle chosen from training information (Section 3)	Lemma 4
Arms, synthetic designs, comparators and seeds (Section 4)	Appendix E
All 84 cells and all seven arms, including split cost and the zoom baseline (Figure 2, Sections 4.1 and 4.2)	Tables 4 and 5
Coverage of all checkpoint risks and of the selected risk by the band (Section 4.1)	Appendix C.2
The maximum over references (Section 4.2)	Appendix C.3
The hybrid interval on a synthetic design (Section 5)	Appendix C.6
Rare-loss counterexample, Hoeffding union-bound interval, near ties and estimator ablations (Section 4.3, Figure 3)	Appendices C.4, C.5 and D
Exact endpoint search and timing (Section 2.4)	Appendix C.7

A METHOD DETAILS

This appendix defines the confidence set $\hat{C}_\alpha$ that the coverage proof uses (Appendix A.1) and justifies the conventions for the tail and the draws (Appendix A.2). It then computes the maximum over references exactly (Appendix A.3).

A.1 THE JOINT CONFIDENCE SET

The coverage proof works with a confidence set $\hat{C}_\alpha$ for the whole vector of checkpoint risks, defined in (8). At a positive floor, on every nonabstaining realization, the accepted set $\hat{C}_\alpha^W$ contains the projection of $\hat{C}_\alpha$ onto the selected coordinate (Theorem 10).

The set $\hat{C}_\alpha$ uses a positive floor and extends $s_k = \hat{\sigma}_k \vee \sigma_{\min}$ to every $k \in [K]$. For a candidate risk vector $\mu \in \mathbb{R}^K$, write $\Delta_k(\mu) = \mu_k - \min_l \mu_l$ and let $i^*(\mu)$ be the smallest index attaining the minimum, with $s_{i^*} = s_{i^*(\mu)}$. The threshold vector $F(r,\mu)$, the discount vector $c(\mu)$ and the radius are

$$\begin{aligned} F(r,\mu)_k &= \max\{r, c(\mu)_k\}, \qquad c(\mu)_k = \frac{\sqrt{m}\,\Delta_k(\mu)}{s_k + s_{i^*}}, \\ \hat{r}_\alpha(\mu) &= \inf\{r \geq 0 : \hat{S}(F(r,\mu)) \leq \alpha\}. \end{aligned} \tag{7}$$

Every denominator in (7) is at least $2\sigma_{\min}$, so $c(\mu) \in [0,\infty)^K$. The threshold vector has a coordinate for each $k \in [K]$, of which $\hat{S}$ reads only those in P. The radius is finite because $\hat{S}(F(r,\mu)) \leq \hat{S}(r\mathbf{1})$ and the right side tends to 0 as $r \to \infty$, and Lemma 5 shows that the infimum is attained. Writing $v_P = (v_k)_{k\in P}$ for $v \in \mathbb{R}^K$, the identity $F(r,\mu) = r\mathbf{1} \vee c(\mu)$ and the definition of ψ in Section 2.2 give $\hat{r}_\alpha(\mu) = \psi(c(\mu)_P)$. Replacing α by an error level $p \in (0,1)$ in (7) defines the radius $\hat{r}_p(\mu)$. The acceptance region $\mathcal{A}_\alpha(\mu)$ and the confidence set $\hat{C}_\alpha$ are

$$\begin{aligned} \mathcal{A}_\alpha(\mu) &= \left\{ x \in \mathbb{R}^K : \frac{\sqrt{m}\,|x_k - \mu_k|}{s_k} \leq F(\hat{r}_\alpha(\mu),\mu)_k \text{ for all } k \right\}, \\ \hat{C}_\alpha &= \{\mu \in \mathbb{R}^K : \hat{\mu} \in \mathcal{A}_\alpha(\mu)\}. \end{aligned} \tag{8}$$

The region $\mathcal{A}_\alpha(\mu)$ is the rectangle centered at μ with half-width $s_k F(\hat{r}_\alpha(\mu),\mu)_k/\sqrt{m}$ in coordinate k. It depends on the validation sample through the scales s_k and the radius $\hat{r}_\alpha(\mu)$, and $\hat{C}_\alpha$ depends

on the sample also through $\hat{\mu}$. The threshold $F(\hat{r}_\alpha(\mu), \mu)_k$ equals the radius when $c(\mu)_k \le \hat{r}_\alpha(\mu)$ and equals $c(\mu)_k$ otherwise. In the second case checkpoint k enters the event that defines $\hat{S}$ only when $|T_k^*|$ exceeds $c(\mu)_k$, a discount that grows with the gap $\Delta_k(\mu)$ for fixed scales. With $c(\mu)_P = 0$ the radius is $\psi(0) = r_{\mathrm{sim}}$, the band's radius. The next proposition shows that the discounts never raise the radius above r_{sim} and says when they leave it there.

Proposition 2 (The radius $\hat{r}_\alpha(\mu)$ is at most the band's radius)**.** *Fix the loss matrix. Then* $\hat{r}_\alpha(\mu) \le r_{\mathrm{sim}}$ *for every* $\mu \in \mathbb{R}^K$*. Equality holds if and only if* $\hat{S}(F(r,\mu)) > \alpha$ *for every* $0 \le r < r_{\mathrm{sim}}$*.*

Proof. Step 1. For every $r \ge 0$*,* $\hat{S}(F(r,\mu)) \le \hat{S}(r\mathbf{1})$*.* By (7), every coordinate of $F(r,\mu)$ is at least r. Raising a threshold cannot enlarge the strict-exceedance event that defines $\hat{S}$.

Step 2. Every $r > r_{\mathrm{sim}}$ *is feasible, so* $\hat{r}_\alpha(\mu) \le r_{\mathrm{sim}}$*.* By definition $r_{\mathrm{sim}} = \psi(0) = \inf\{r \ge 0 : \hat{S}(r\mathbf{1}) \le \alpha\}$. This set is nonempty because $\hat{S}(r\mathbf{1}) = \mathbb{P}_e\{\max_{k\in P} |T_k^*| > r\}$ tends to 0 as $r \to \infty$ and $\alpha > 0$. Fix $r > r_{\mathrm{sim}}$. The definition of the infimum gives some $r' \in [r_{\mathrm{sim}}, r)$ with $\hat{S}(r'\mathbf{1}) \le \alpha$. Then

$$\hat{S}(F(r,\mu)) \le \hat{S}(r\mathbf{1}) \le \hat{S}(r'\mathbf{1}) \le \alpha,$$

by Step 1, by monotonicity of $\hat{S}(r\mathbf{1})$ in r with $r' < r$, and by the choice of r'. So r is feasible in the infimum defining $\hat{r}_\alpha(\mu)$, and hence $\hat{r}_\alpha(\mu) \le r_{\mathrm{sim}}$.

Step 3. The equality case. If some $0 \le r < r_{\mathrm{sim}}$ has $\hat{S}(F(r,\mu)) \le \alpha$, this r is feasible, so $\hat{r}_\alpha(\mu) \le r < r_{\mathrm{sim}}$. Otherwise every feasible radius is at least r_{sim}. Then $\hat{r}_\alpha(\mu) \ge r_{\mathrm{sim}}$, and Step 2 gives equality. □

A.2 Tail and rank conventions

The zoom correction of Zrnic & Fithian (2025) gives an interval for the mean of the coordinate with the largest observed value. We apply the variance-adaptive form of their Section 5.4 to $-\hat{\mu}$ and negate the output. Under the smaller-index tie rule of Section 2.1, $\hat{k}$ indexes the largest coordinate of $-\hat{\mu}$, and the gap of $-\mu_k$ below $\max_l(-\mu_l)$ equals $\Delta_k(\mu)$. With their σ_k set to $s_k/\sqrt{m}$ and their tail bound $S(v)$ set to $\hat{S}$ at the thresholds $\sqrt{m}\, v_k/s_k$, their region and the radius of their equation (9) become (8) and (7).

Selection favors downward deviations of $\hat{\mu}_{\hat{k}}$, so the correction reaches further upward. This downward bias is the selection bias that Efron (2011) describes for the largest of many estimates, with the sign reversed. The radius $\hat{r}_\alpha^W(t)$ is non-decreasing in t (Appendix B.4). Hence every $\delta \ge 0$ with $\hat{\mu}_{\hat{k}} - \delta \in \hat{C}_\alpha^W$ satisfies

$$\delta \le \hat{r}_\alpha^W(\hat{\mu}_{\hat{k}} - \delta)\, s_{\hat{k}}/\sqrt{m} \le \hat{r}_\alpha^W(\hat{\mu}_{\hat{k}} + \delta)\, s_{\hat{k}}/\sqrt{m},$$

where the first inequality is membership in $\hat{C}_\alpha^W$ and the second is monotonicity. So $\hat{\mu}_{\hat{k}} + \delta \in \hat{C}_\alpha^W$, and the accepted set reaches at least as far above $\hat{\mu}_{\hat{k}}$ as below it.

The constant columns Z of Section 2.1 are removed before studentization, so no quotient $0/0$ arises. Replacing every s_k by cs_k with $c > 0$ divides each T_k^*, each discount and each radius by c. The products $r_{\mathrm{sim}} s_i$ and $\hat{r}_\alpha^W(t) s_{\hat{k}}$ are unchanged, and so are $W(t)$, the search interval (3) and $\hat{C}_\alpha^W$. The rescaling therefore leaves the grid, and hence every B-Zoom endpoint, unchanged when the grid rule sees the scales only through the search interval, as a uniform grid over it does. Dividing by $m-1$ instead of m in $\hat{\sigma}_k^2$ multiplies each $\hat{\sigma}_k$ by $\sqrt{m/(m-1)}$, so it gives the same B-Zoom interval when $\hat{\sigma}_k \ge \sigma_{\min}$ for every $k \in P$.

An indicator column with n ones has $\hat{\sigma}_k^2 = n(m-n)/m^2$ and is nonconstant exactly when $1 \le n \le m-1$. Then $n(m-n) - (m-1) = (n-1)(m-1-n) \ge 0$, so $\hat{\sigma}_k \ge \sqrt{m-1}/m$. This bound decreases in m and exceeds 10^{-3} for every $m \le 10^5$, which covers all our validation sizes. At $\sigma_{\min} = 10^{-3}$, every 0–1 loss column in P therefore keeps $s_k = \hat{\sigma}_k$. Since Z does not depend on the floor, every interval for a 0–1 loss at this floor equals its zero-floor interval exactly.

A sequence $v^{(n)}$ of threshold vectors *decreases to* v when each coordinate is non-increasing in n and converges to that of v. A tail S is *continuous from above* when $S(v^{(n)}) \to S(v)$ for every such

sequence. Strict-exceedance tails have this property, so the infimum defining each radius is itself feasible, and hence attained, even when a draw equals a threshold (Lemma 5). The weak tail, with $\geq$ in place of $>$, lacks this property, so the infimum defining its radius need not be attained.

At a fixed threshold vector, a *fresh draw* is one additional independently generated masked maximum. Its exceedance probability at rank k is the chance, conditional on the data, that it strictly exceeds the order statistic of rank k of the first B draws. Ranks count from zero in increasing order. We bound this probability and then find the smallest rank that keeps it at most α.

Step 1. The $B+1$ values are exchangeable. Conditional on the data, the B draws and the fresh draw are i.i.d.

Step 2. On any outcome, at most $B-k$ of the $B+1$ values strictly exceed at least $k+1$ of the others. If j values have this property, the smallest of them strictly exceeds at least $k+1$ values outside the group, so $j+k+1 \leq B+1$.

Step 3. The exceedance probability at rank k is at most $(B-k)/(B+1)$, and (9) gives the smallest rank whose bound is at most α. The fresh draw strictly exceeds the value of rank k exactly when it strictly exceeds at least $k+1$ of the other B values. By Step 1, each of the $B+1$ values has the same probability of the event in Step 2, and by Step 2 these probabilities sum to at most $B-k$. The exceedance probability at rank k is therefore at most $(B-k)/(B+1)$, with equality when ties have probability zero. This bound is at most α exactly when $k \geq (1-\alpha)(B+1)-1$, and some rank in $\{0,\ldots,B-1\}$ meets this condition exactly when $B \geq \lceil 1/\alpha \rceil - 1$. For such B, the smallest zero-based rank whose fresh-draw exceedance bound is at most α is

$$k_\alpha(B) = \lceil (1-\alpha)(B+1) \rceil - 1, \qquad \frac{B-k_\alpha(B)}{B+1} \leq \alpha. \tag{9}$$

At $B = 2000$ and $\alpha = 0.10$ the rank is 1800 and the bound is $200/2001 \approx 0.09995$. The usual empirical $(1-\alpha)$ quantile has rank $\lceil (1-\alpha)B \rceil - 1 = 1799$, and when ties have probability zero its fresh-draw exceedance is $201/2001 \approx 0.10045$, above α. At most $B - k_\alpha(B) - 1$ draws lie strictly above the order statistic of rank $k_\alpha(B)$. The fraction of the B draws strictly above it is therefore at most $\alpha_B = (B - k_\alpha(B) - 1)/B$, which equals 0.0995 at these settings. Appendix B.6 shows that $\psi_B(y)$ is the smallest $r \geq 0$ at which the fraction of masked maxima above r is at most α_B.

A.3 Computing the maximum over references

At each grid value t, the released code (Appendix E.3) evaluates $\hat{r}^{W}_{\alpha,B}(t) = \max_{i \in W(t)} \psi_B(g^{(i)}(t))$, the radius (2) from the B draws, with $W(t)$ formed using $\psi_B(0)$ in place of r_{sim}. For a reference scale c, the largest entry of draw b that exceeds its discount among the columns $j \in P \setminus \{\hat{k}\}$ is

$$N_b(c) = \max_{j \in P \setminus \{\hat{k}\}} |T^*_{bj}| \, \mathbf{1}\left\{ |T^*_{bj}| > \frac{\sqrt{m}(\hat{\mu}_j - t)_+}{2s_j + c} \right\},$$

with an empty maximum equal to zero. For reference i, the discounts (1) vanish at $\hat{k}$ and i, so the masked maximum of draw b is the largest of $|T^*_{b\hat{k}}|$, $|T^*_{bi}|$ and $N_b(s_i)$. The term of $N_b(s_i)$ with $j = i$ is at most $|T^*_{bi}|$ and so leaves this maximum unchanged. Since $N_b(c)$ involves the reference only through c, the code computes it once per distinct scale in $W(t)$.

B-Zoom never inverts $\hat{R}$, which may be singular. The code generates the multipliers in blocks of bounded size, so memory is $O(mK + BK)$ and no covariance matrix is formed or factorized.

B Proofs

The coverage proof follows the four steps sketched in Section 3. Appendix B.1 fixes symbols and proves three preliminary lemmas. Appendix B.2 bounds the covariance-estimation error of the first step and the two Gaussian approximation errors of the second. Appendix B.3 completes the second step by bounding Gaussian probability near rectangle boundaries. Appendices B.4 and B.5 carry out the third and fourth steps, showing that the reported interval contains the selected risk whenever the studentized errors lie in a slightly shrunk rectangle.

Appendix B.6 extends these results to the empirical tail of the B draws and bounds the exceedance probability of a fresh draw. Theorem 13 in Appendix B.7 assembles these bounds, and Appendix B.8 derives Theorem 1 from it and remarks on a shrinking floor.

B.1 Notation and preliminary lemmas

Formally, $\mathcal{G}_{\mathrm{tr}}$ is the σ-field generated by the training data and optimizer randomness, and training-measurable (Section 2.1) means $\mathcal{G}_{\mathrm{tr}}$-measurable. The example space $\mathcal{Z}$ and parameter space Θ are standard Borel spaces, that is, measurable spaces isomorphic to Borel subsets of complete separable metric spaces. The loss is a jointly measurable map from $\mathcal{Z} \times \Theta$ to $[0, M]$, and the checkpoints and their count are training-measurable, so the conditional laws below are well defined. Unless stated otherwise, probabilities condition on $\mathcal{G}_{\mathrm{tr}}$, under which the validation rows are i.i.d. in $[0, M]^K$ (Lemma 3). We assume $K \geq 2$ and $m \geq 3$, the range of the central limit theorem of Chernozhukov et al. (2022, Theorem 2.1), although the method needs only $m \geq 2$. Every logarithm is natural.

Write $a = \sigma_{\min}$ for a positive, training-measurable floor with $a \leq \min_k \sigma_k$, so that $s_k = \hat{\sigma}_k \vee a$ for every $k \in [K]$, constant columns included. Write $\Lambda = \log(K \vee m)$, and let C_K be the constant C_k of (5) at $k = K$. For training-measurable $\delta, \epsilon \in (0, 1)$, the error terms b, ℓ and g of Theorem 13 and the intermediate bounds ρ and β are

$$\rho = 2\sqrt{2}\, M^2 \sqrt{\frac{\log(4K^2/\delta)}{m}} + 2M^2 \frac{\log(4K^2/\delta)}{m}, \qquad \beta = \frac{\rho}{2a^2}, \tag{10}$$

$$b = C_P(1) \left(\frac{4\rho}{a^2} \log^2(2K) \right)^{1/2}, \qquad \ell = C_\diamond \left\{ \left(\frac{M}{a} \right)^2 \frac{\Lambda^5}{m} \right\}^{1/4}, \tag{11}$$

$$g = 16 C_K \beta \sqrt{2 \log(2K/\epsilon)} + \epsilon. \tag{12}$$

Here $C_P(1)$ is the constant of Chernozhukov et al. (2022, Prop. 2.1) when the second Gaussian vector in that proposition has variances at least one. The constant $C_\diamond$ is $2^{5/4}(\log 2)^{-1/2}$ times the constant C of their Theorem 2.1 at $b_1 = b_2 = 1$, the moment constants of their Condition M (Step 7 of Appendix B.2). Both are absolute constants, which do not depend on the data. Table 3 lists the remaining standing symbols and the quantity that each of ρ, β, b, ℓ and $g/2$ bounds. Lemmas 7–9, Theorem 10 and Lemmas 11 and 12 rely only on the hypotheses they state.

Table 3: Standing symbols of the proofs. The event Ω of Appendix B.2 has conditional probability at least $1 - \delta$. The last block gives the quantity that each of ρ, β, b, ℓ and $g/2$ bounds (Appendices B.2 and B.3), with $\mathcal{B}$ ranging over rectangles.

Σ	$\mathrm{Cov}(L_{1\cdot} \mid \mathcal{G}_{\mathrm{tr}})$	R	$R_{kl} = \Sigma_{kl}/(\sigma_k \sigma_l)$, the correlation
$\hat{V}$	$\hat{V}_{kl} = m^{-1} \sum_j \tilde{L}_{jk} \tilde{L}_{jl}$	$\hat{Z}$	a vector with conditional law $N(0, R)$
$S^{\hat{Z}}(v)$	$\mathbb{P}(\exists k : \lvert\hat{Z}_k\rvert > v_k \mid \mathcal{G}_{\mathrm{tr}})$	$\bar{\Phi}(x)$	$\mathbb{P}(N > x)$ for $N \sim N(0,1)$
$\mathcal{B}_v$	$\prod_k [-v_k, v_k]$, a rectangle of Section 3	$i^*(\mu)$	the smallest index minimizing μ
d_k	$\sqrt{m}(\hat{\mu}_k - \mu_k)/\sigma_k$, the *standardized error*	$\hat{d}_k$	$\sqrt{m}(\hat{\mu}_k - \mu_k)/s_k$, the *studentized error*
T^*_k, $k \in Z$	0, the *zero extension* of T^* to $[K]$	$\hat{S}^{[K]}(v)$	$\mathbb{P}_e(\exists k \in [K] : \lvert T^*_k\rvert > v_k)$, $v \in [0, \infty)^K$
$\hat{C}^{[K]}_\alpha$	$\hat{C}_\alpha$ of (8) with $\hat{S}^{[K]}$ in place of $\hat{S}$	$\psi^{[K]}(y)$	$\inf\{r \geq 0 : \hat{S}^{[K]}(r\mathbf{1} \vee y) \leq \alpha\}$
ρ	$\max_{k,l} \lvert\hat{V}_{kl} - \Sigma_{kl}\rvert$ on Ω	β	$\max_k \lvert s_k/\sigma_k - 1\rvert$ on Ω
b	$\sup_{\mathcal{B}} \lvert\mathbb{P}_e(T^* \in \mathcal{B}) - \mathbb{P}(\hat{Z} \in \mathcal{B})\rvert$ on Ω	ℓ	$\sup_{\mathcal{B}} \lvert\mathbb{P}(d \in \mathcal{B} \mid \mathcal{G}_{\mathrm{tr}}) - \mathbb{P}(\hat{Z} \in \mathcal{B})\rvert$
$g/2$	$\mathbb{P}(\hat{Z} \in \mathcal{B}_u \setminus \mathcal{B}_{\varrho u})$ for every $u \in [0, \infty)^K$, if $\beta \leq 1/8$ and $\varrho \in [1 - 4\beta, 1]$		

Lemma 3 (Conditional validation model and averaging)**.** (i) *Conditional on $\mathcal{G}_{\mathrm{tr}}$, the rows $(L_{j\cdot})_{j \leq m}$ are i.i.d. in $[0, M]^K$ with mean μ and covariance Σ.* (ii) *Let E be an event and $\eta \geq 0$ an integrable $\mathcal{G}_{\mathrm{tr}}$-measurable random variable. If $\mathbb{P}(E \mid \mathcal{G}_{\mathrm{tr}}) \geq 1 - \alpha - \eta$ almost surely, then $\mathbb{P}(E) \geq 1 - \alpha - \mathbb{E}[\eta]$.*

Proof. Step 1. Part (i) holds. The validation sample is independent of $\mathcal{G}_{\mathrm{tr}}$, so its conditional law given $\mathcal{G}_{\mathrm{tr}}$ may be taken to be $\mathcal{D}^{\otimes m}$, the law of m i.i.d. draws from $\mathcal{D}$. The count K is $\mathcal{G}_{\mathrm{tr}}$-measurable, so we may argue on each event $\{K = K_0\}$ with K_0 fixed, where the checkpoints $\theta_1, \ldots, \theta_{K_0}$ are

$\mathcal{G}_{\mathrm{tr}}$-measurable. Since the loss is jointly measurable on $\mathcal{Z} \times \Theta$, each row $L_{j\cdot}$ given $\mathcal{G}_{\mathrm{tr}}$ is the image of z'_j under one fixed measurable map, the loss vector at $\theta_1, \dots, \theta_{K_0}$. The z'_j are i.i.d. under $\mathcal{D}^{\otimes m}$, so the rows are conditionally i.i.d., with conditional mean μ and conditional covariance Σ by definition. The loss takes values in $[0, M]$, so the rows lie in $[0, M]^{K_0}$.

Step 2. Part (ii) holds. Taking expectations in the hypothesis of (ii) and using the tower property gives

$$\mathbb{P}(E) = \mathbb{E}[\mathbb{P}(E \mid \mathcal{G}_{\mathrm{tr}})] \geq 1 - \alpha - \mathbb{E}[\eta].$$

□

For $v, y \in [0, \infty)^K$, the zero-extended tail $\hat{S}^{[K]}$ and radius $\psi^{[K]}$ of Table 3 reduce to the tail $\hat{S}$ and radius ψ of Section 2.2. A zero coordinate never strictly exceeds a non-negative threshold, so $\hat{S}^{[K]}(v) = \hat{S}(v_P)$. For $r \geq 0$ the vector $r\mathbf{1} \vee y$ is non-negative with P coordinates $r\mathbf{1} \vee y_P$, so the two infima run over the same radii and

$$\psi^{[K]}(y) = \psi(y_P), \qquad y \in [0, \infty)^K. \tag{13}$$

Write $T_k^{*(b)} = T_{bk}^*$ for the value at checkpoint k of bootstrap draw b of Section 2.2. For $v \in [0, \infty)^P$, the empirical tail of the B draws is

$$\hat{S}_B(v) = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\{\exists k \in P : |T_k^{*(b)}| > v_k\}. \tag{14}$$

The same two arguments, applied draw by draw, show that the tail $\hat{S}_B^{[K]}$ of the zero-extended draws satisfies $\hat{S}_B^{[K]}(v) = \hat{S}_B(v_P)$. Its radius $\psi_B^{[K]}(y) = \inf\{r \geq 0 : \hat{S}_B^{[K]}(r\mathbf{1} \vee y) \leq \alpha_B\}$, for $y \in [0, \infty)^K$, therefore equals $\inf\{r \geq 0 : \hat{S}_B(r\mathbf{1} \vee y_P) \leq \alpha_B\}$. Appendix B.6 identifies this radius with the order statistic $\psi_B(y_P)$ of Section 2.2.

Since $F(r, \mu)$ lies in $[0, \infty)^K$, $\hat{S}^{[K]}(F(r, \mu)) = \hat{S}(F(r, \mu)_P)$ at every $r \geq 0$, so the two sets share the radius $\hat{r}_\alpha(\mu)$ and $\hat{C}_\alpha^{[K]} = \hat{C}_\alpha$. For a fixed candidate μ, the first equality below is the definition of $\hat{C}_\alpha$ combined with $\hat{C}_\alpha^{[K]} = \hat{C}_\alpha$. Setting $x = \hat{\mu}$ in (8) turns the kth constraint into $|\hat{d}_k| \leq F(\hat{r}_\alpha(\mu), \mu)_k$, which gives the second:

$$\{\hat{\mu} \in \mathcal{A}_\alpha(\mu)\} = \{\mu \in \hat{C}_\alpha^{[K]}\} = \{\hat{d} \in \mathcal{B}_v\}, \qquad v = F(\hat{r}_\alpha(\mu), \mu). \tag{15}$$

The index $i^*(\mu)$ depends on μ alone, so the threshold vector v depends on the validation sample only through the radius and through every denominator in (7).

Given L and $\mathcal{G}_{\mathrm{tr}}$, the bound b compares two fixed Gaussian laws, so it holds at a rectangle chosen by the data. The central limit bound (17) holds only at rectangles fixed before validation, and it can fail at $\mathcal{B}_v$, which the same sample chooses (Section 3). The next lemma extends a bound over fixed rectangles to a training-measurable rectangle.

Lemma 4 (Substitution at a training-measurable rectangle)**.** *Let $\varkappa$ be a regular conditional distribution of d given $\mathcal{G}_{\mathrm{tr}}$. That is, for each point ω of the underlying probability space, $\varkappa(\omega, \cdot)$ is a probability measure on $\mathbb{R}^K$, and for each Borel set $\mathcal{H} \subseteq \mathbb{R}^K$, $\varkappa(\cdot, \mathcal{H})$ is a version of $\mathbb{P}(d \in \mathcal{H} \mid \mathcal{G}_{\mathrm{tr}})$. One choice of $\varkappa(\omega, \cdot)$ is the law of d under $\mathcal{D}^{\otimes m}$ with the checkpoints of ω held fixed. Write $\Gamma_\omega = N(0, R(\omega))$ for the conditional law of $\hat{Z}$.*

Let U be a $\mathcal{G}_{\mathrm{tr}}$-measurable random vector in $[0, \infty)^K$, and write $\varkappa(\cdot, \mathcal{B}_U)$ and $\Gamma_\cdot(\mathcal{B}_U)$ for the maps $\omega \mapsto \varkappa(\omega, \mathcal{B}_{U(\omega)})$ and $\omega \mapsto \Gamma_\omega(\mathcal{B}_{U(\omega)})$. Then $\varkappa(\cdot, \mathcal{B}_U)$ is $\mathcal{G}_{\mathrm{tr}}$-measurable and $\mathbb{P}(d \in \mathcal{B}_U \mid \mathcal{G}_{\mathrm{tr}}) = \varkappa(\cdot, \mathcal{B}_U)$ almost surely. Let $\eta \geq 0$ be $\mathcal{G}_{\mathrm{tr}}$-measurable, and suppose that outside one null set, $|\varkappa(\omega, \mathcal{B}_u) - \Gamma_\omega(\mathcal{B}_u)| \leq \eta(\omega)$ for every $u \in [0, \infty)^K$. Then $|\mathbb{P}(d \in \mathcal{B}_U \mid \mathcal{G}_{\mathrm{tr}}) - \mathbb{P}(\hat{Z} \in \mathcal{B}_U \mid \mathcal{G}_{\mathrm{tr}})| \leq \eta$ almost surely.

Proof. The identity $\mathbb{P}(d \in \mathcal{B}_U \mid \mathcal{G}_{\mathrm{tr}}) = \varkappa(\cdot, \mathcal{B}_U)$ means that $\varkappa(\cdot, \mathcal{B}_U)$ is $\mathcal{G}_{\mathrm{tr}}$-measurable and

$$\mathbb{E}[\mathbf{1}_E \mathbf{1}\{d \in \mathcal{B}_U\}] = \mathbb{E}[\mathbf{1}_E \varkappa(\cdot, \mathcal{B}_U)] \qquad \text{for every } E \in \mathcal{G}_{\mathrm{tr}}.$$

Steps 1 and 2 prove the identity, first for threshold vectors with countably many values and then in the limit from above. Step 3 derives the Gaussian bound, and Step 4 checks the stated choice of $\varkappa$. Each event $\{K = K_0\}$ is training-measurable, so we argue on one such event at a time.

Step 1. The identity holds when U is replaced by a $\mathcal{G}_{\mathrm{tr}}$-measurable vector V with countably many values. Fix $E \in \mathcal{G}_{\mathrm{tr}}$, and put $E_v = E \cap \{V = v\}$ for each value v of V. The defining property of $\varkappa$ at the fixed rectangle $\mathcal{B}_v$ gives $\mathbb{E}[\mathbf{1}_{E_v}\mathbf{1}\{d \in \mathcal{B}_v\}] = \mathbb{E}[\mathbf{1}_{E_v}\varkappa(\cdot, \mathcal{B}_v)]$, and summing over v by monotone convergence gives

$$\mathbb{E}[\mathbf{1}_E\mathbf{1}\{d \in \mathcal{B}_V\}] = \mathbb{E}[\mathbf{1}_E\varkappa(\cdot, \mathcal{B}_V)].$$

The integrand $\varkappa(\cdot, \mathcal{B}_V) = \sum_v \mathbf{1}\{V = v\}\varkappa(\cdot, \mathcal{B}_v)$ on the right is $\mathcal{G}_{\mathrm{tr}}$-measurable.

Step 2. The identity passes to U along dyadic upper approximations. For $n \geq 1$ put $U_k^{(n)} = 2^{-n}\lceil 2^n U_k \rceil$, a training-measurable vector with countably many values, so the display of Step 1 holds at $V = U^{(n)}$. Since $\lceil 2x \rceil \leq 2\lceil x \rceil$ and $0 \leq U_k^{(n)} - U_k < 2^{-n}$, the vectors $U^{(n)}$ decrease to U. The closed rectangles $\mathcal{B}_{U^{(n)}}$ are therefore nested with intersection $\mathcal{B}_U$, which gives

$$\mathbf{1}\{d \in \mathcal{B}_{U^{(n)}}\} \downarrow \mathbf{1}\{d \in \mathcal{B}_U\}, \qquad \varkappa(\omega, \mathcal{B}_{U^{(n)}}) \downarrow \varkappa(\omega, \mathcal{B}_U).$$

The first limit holds at every outcome, and the second is continuity of each probability measure $\varkappa(\omega, \cdot)$ along these nested sets.

As a pointwise limit of $\mathcal{G}_{\mathrm{tr}}$-measurable functions, $\varkappa(\cdot, \mathcal{B}_U)$ is $\mathcal{G}_{\mathrm{tr}}$-measurable. Dominated convergence on both sides of the display of Step 1 gives the identity at U for every $E \in \mathcal{G}_{\mathrm{tr}}$. The approximation must come from above because d may put mass on the faces of $\mathcal{B}_U$.

Step 3. The Gaussian bound holds. The correlation R is training-measurable, so $\omega \mapsto \Gamma_\omega$ is a regular conditional distribution of $\hat{Z}$, and the identity for $\hat{Z}$ gives $\mathbb{P}(\hat{Z} \in \mathcal{B}_U \mid \mathcal{G}_{\mathrm{tr}}) = \Gamma_\cdot(\mathcal{B}_U)$ almost surely. Outside the null set of the hypothesis, the bound at $u = U(\omega)$ reads

$$|\varkappa(\omega, \mathcal{B}_{U(\omega)}) - \Gamma_\omega(\mathcal{B}_{U(\omega)})| \leq \eta(\omega).$$

The two almost-sure identities turn this into the stated bound.

Step 4. The law of d under $\mathcal{D}^{\otimes m}$ with the checkpoints of ω held fixed is a regular conditional distribution of d. The loss, the checkpoints and the population means and standard deviations are measurable, so d is a jointly measurable function of ω and the validation sample. That sample has law $\mathcal{D}^{\otimes m}$ and is independent of $\mathcal{G}_{\mathrm{tr}}$, so Tonelli's theorem gives measurability of this law in ω and its defining property. □

The coverage proof evaluates the tail at the radius (7), an infimum, and the next lemma shows that this infimum is attained, using continuity from above in the sense of Appendix A.2.

Lemma 5 (Attainment of the radius)**.** *Let $S : [0, \infty)^K \to [0, 1]$ be non-increasing in each coordinate and continuous from above. Let $c \in [0, \infty)^K$ and write $F_c(r)_k = \max\{r, c_k\}$. Fix $\lambda \in [0, 1]$ and suppose $S(F_c(r)) \leq \lambda$ for some $r \geq 0$. Then $S(F_c(\hat{r})) \leq \lambda$ at $\hat{r} = \inf\{r \geq 0 : S(F_c(r)) \leq \lambda\}$, and the same holds at every $r \geq \hat{r}$.*

The tails $\hat{S}$, $\hat{S}^{[K]}$, $S^{\hat{Z}}$ and the empirical tail $\hat{S}_B$ of (14) satisfy these hypotheses for every loss matrix and every set of B bootstrap draws. Here $S^{\hat{Z}}$ is taken at a fixed ω, as the $N(0, R(\omega))$ probability that $|\hat{Z}_k| > v_k$ for some k. The tails $\hat{S}$ and $\hat{S}_B$ read only the P coordinates of v. All four tend to 0 as $\min_k v_k \to \infty$, so the hypothesis on λ holds for each of them at every $\lambda > 0$.

Proof. Step 1 proves attainment, and Steps 2 to 4 check the hypotheses for the four tails.

Step 1. The infimum is attained. Each coordinate $\max\{r, c_k\}$ of F_c is continuous and non-decreasing in r, so $\phi(r) = S(F_c(r))$ is non-increasing. The feasible set $\{r \geq 0 : \phi(r) \leq \lambda\}$ is nonempty by hypothesis, so its infimum $\hat{r}$ is finite. Each $r > \hat{r}$ exceeds some feasible r' and therefore satisfies $\phi(r) \leq \phi(r') \leq \lambda$, so $r_n = \hat{r} + 1/n$ is feasible for every n. The vectors $F_c(r_n)$ decrease to $F_c(\hat{r})$ in each coordinate, and

$$\phi(r) \leq \phi(\hat{r}) = \lim_{n \to \infty} \phi(r_n) \leq \lambda \qquad \text{for every } r \geq \hat{r}.$$

The first inequality is monotonicity of ϕ, the equality is continuity from above, and the last inequality is feasibility of each r_n.

Step 2. The four tails share one form and are non-increasing. Each has the form $S(v) = \mathbb{Q}(\exists k : |X_k| > v_k)$ for a random vector X with law $\mathbb{Q}$. For $\hat{S}$ and $\hat{S}^{[K]}$, X is $(T^*_k)_{k\in P}$ and $(T^*_k)_{k\in[K]}$ respectively, both under $\mathbb{P}_e$. For $S^{\hat{Z}}$ it is $\hat{Z}$ under $N(0, R(\omega))$, and for $\hat{S}_B$ it is $(T^{*(b)}_k)_{k\in P}$ with b uniform on $\{1, \dots, B\}$ and the B draws held fixed. Every tail is non-increasing because the event $\{\exists k : |X_k| > v_k\}$ shrinks as v grows.

Step 3. Each tail is continuous from above. Fix a value x of X and let $v^{(n)}$ decrease to v. If $|x_k| > v_k$ for some k, then $|x_k| > v^{(n)}_k$ for all large n. If $|x_k| \le v_k$ for every k, then $|x_k| \le v^{(n)}_k$ for every k and n. The indicators $\mathbf{1}\{\exists k : |x_k| > v^{(n)}_k\}$ therefore increase to $\mathbf{1}\{\exists k : |x_k| > v_k\}$, and continuity from below of $\mathbb{Q}$ gives $S(v^{(n)}) \to S(v)$.

Step 4. Each tail tends to 0 as $\min_k v_k \to \infty$. A union bound gives $S(v) \le \sum_k \mathbb{Q}(|X_k| > v_k)$, and each term tends to 0 as $v_k \to \infty$ because X_k is real-valued. No further property of $\mathbb{Q}$ enters, which allows a singular R. □

B.2 Concentration and Gaussian comparison

This subsection builds an event Ω of conditional probability at least $1 - \delta$ and proves the Gaussian approximations (16) and (17). Steps 1 to 5 show that T^* is Gaussian given $(L, \mathcal{G}_{\mathrm{tr}})$, with a covariance close to R on Ω. Step 6 turns this closeness into b, and Step 7 bounds the error ℓ of the Gaussian approximation to the sampling law of d.

Step 1. Given $(L, \mathcal{G}_{\mathrm{tr}})$, the vector T^* is $N(0, \hat{R})$ on all of $[K]$, where $\hat{R}$ has diagonal at most one and rank at most $m - 1$. Once $(L, \mathcal{G}_{\mathrm{tr}})$ is fixed, the only randomness left is the multiplier vector $e \sim N(0, I_m)$, which is independent of L and $\mathcal{G}_{\mathrm{tr}}$. The formula for G^*_k in Section 2.2 extends to every $k \in [K]$ and reads $\sqrt{m}\, G^* = m^{-1/2}\tilde{L}^\top e$. Hence $\sqrt{m}\, G^* \mid (L, \mathcal{G}_{\mathrm{tr}}) \sim N(0, \hat{V})$ with $\hat{V} = m^{-1}\tilde{L}^\top\tilde{L}$. Dividing coordinate k by $s_k \ge a > 0$ gives $T^* \mid (L, \mathcal{G}_{\mathrm{tr}}) \sim N(0, \hat{R})$ with $\hat{R}_{kl} = \hat{V}_{kl}/(s_k s_l)$, the matrix of Section 2.2 extended to all of $[K]$.

On a constant column $k \in Z$, $\tilde{L}_{jk} = 0$ for every j, so $T^*_k = 0$, in agreement with the zero extension. Since $\hat{V}_{kk} = \hat{\sigma}^2_k$ and $s^2_k = \hat{\sigma}^2_k \vee a^2$, the diagonal is $\hat{R}_{kk} = \min(1, \hat{\sigma}^2_k/a^2) \le 1$. The centering $\tilde{L} = (I_m - m^{-1}\mathbf{1}\mathbf{1}^\top)L$ uses a projector of rank $m - 1$, so $\hat{R}$ has rank at most $m - 1$, and no later step needs full rank.

Step 2. The covariance $\hat{V}$ is an average centered at the population means, minus a correction. Write $\|\cdot\|_{\max}$ for the largest absolute entry of a matrix, $\hat{\Sigma}_{kl} = m^{-1}\sum_j (L_{jk} - \mu_k)(L_{jl} - \mu_l)$ and $\zeta_k = \hat{\mu}_k - \mu_k$. Write $u_j = L_{jk} - \mu_k$ and $v_j = L_{jl} - \mu_l$, whose averages are ζ_k and ζ_l. The identity $m^{-1}\sum_j (u_j - \bar{u})(v_j - \bar{v}) = m^{-1}\sum_j u_j v_j - \bar{u}\bar{v}$ gives $\hat{V} = \hat{\Sigma} - \zeta\zeta^\top$, so

$$\|\hat{V} - \Sigma\|_{\max} \le \|\hat{\Sigma} - \Sigma\|_{\max} + \max_k \zeta^2_k.$$

Step 3. There is an event Ω, which does not involve the floor a, with $\mathbb{P}(\Omega^c \mid \mathcal{G}_{\mathrm{tr}}) \le \delta$ and $\|\hat{V} - \Sigma\|_{\max} \le \rho$ on Ω. Given $\mathcal{G}_{\mathrm{tr}}$, the rows are i.i.d. in $[0, M]^K$ by Lemma 3, and the training-measurable δ is a constant. For fixed (k, l), the summands of $\hat{\Sigma}_{kl}$ are i.i.d. with mean Σ_{kl} and values in $[-M^2, M^2]$. The deviations $L_{jk} - \mu_k$ lie in $[-M, M]$ and average to ζ_k.

Hoeffding's inequality (Hoeffding, 1963, Theorem 2), with union bounds over the K^2 pairs and the K coordinates, gives

$$\mathbb{P}\big(\|\hat{\Sigma} - \Sigma\|_{\max} > t \mid \mathcal{G}_{\mathrm{tr}}\big) \le 2K^2 e^{-mt^2/(2M^4)}, \qquad \mathbb{P}\big(\max_k |\zeta_k| > t \mid \mathcal{G}_{\mathrm{tr}}\big) \le 2Ke^{-mt^2/(2M^2)}.$$

The right sides equal $\delta/2$ at $\bar{\Delta}_1 := M^2\sqrt{2\log(4K^2/\delta)/m}$ and $\bar{\zeta} := M\sqrt{2\log(4K/\delta)/m}$ respectively. Let Ω be the event that $\|\hat{\Sigma} - \Sigma\|_{\max} \le \bar{\Delta}_1$ and $\max_k |\zeta_k| \le \bar{\zeta}$, so that $\mathbb{P}(\Omega^c \mid \mathcal{G}_{\mathrm{tr}}) \le \delta$. On Ω, Step 2 gives $\|\hat{V} - \Sigma\|_{\max} \le \bar{\Delta}_1 + \bar{\zeta}^2 \le \rho$. The last inequality holds because $\bar{\Delta}_1$ is half the first term of ρ in (10) and $\bar{\zeta}^2 = 2M^2\log(4K/\delta)/m$ is at most the second.

Step 4. On Ω, $|s_k/\sigma_k - 1| \le \beta$ for every k. If also $\beta \le 1/8$, then Z is empty and $1 - 2\beta \le \hat{R}_{kk} \le 1$. Since $\hat{V}_{kk} = \hat{\sigma}_k^2$ and $\Sigma_{kk} = \sigma_k^2$, Step 3 gives $|\hat{\sigma}_k^2 - \sigma_k^2| \le \rho$ on Ω. Because $(x \vee a)^2 = x^2 \vee a^2$ and $t \mapsto t \vee a^2$ is 1-Lipschitz, every $x, y \ge 0$ satisfy $|x \vee a - y \vee a| \le |x^2 - y^2|/(x \vee a + y \vee a)$. Take $x = \hat{\sigma}_k$ and $y = \sigma_k$. Then $x \vee a = s_k$, and the floor condition $a \le \sigma_k$ gives $y \vee a = \sigma_k$. Dividing by σ_k and using $s_k, \sigma_k \ge a$ gives, on Ω,

$$\left|\frac{s_k}{\sigma_k} - 1\right| \le \frac{\rho}{\sigma_k(s_k + \sigma_k)} \le \frac{\rho}{2a^2} = \beta.$$

Under the condition $\beta \le 1/8$ of Theorem 13, every k satisfies $\hat{\sigma}_k^2 \ge \sigma_k^2 - \rho \ge a^2 - \rho = (1-2\beta)a^2 > 0$ on Ω. Hence no column is constant, $P = [K]$, and $\hat{R}_{kk} = \min(1, \hat{\sigma}_k^2/a^2)$ lies in $[1 - 2\beta, 1]$.

Step 5. On Ω, $\|\hat{R} - R\|_{\max} \le 4\rho/a^2$ at every value of β. The Cauchy–Schwarz inequality gives $|R_{kl}| \le 1$ and, since $|\hat{V}_{kl}| \le \hat{\sigma}_k\hat{\sigma}_l \le s_k s_l$, also $|\hat{R}_{kl}| \le 1$. If $\beta > 1/2$, these bounds give $|\hat{R}_{kl} - R_{kl}| \le 2 < 8\beta = 4\rho/a^2$. If $\beta \le 1/2$, write

$$\hat{R}_{kl} - R_{kl} = \frac{\hat{V}_{kl} - \Sigma_{kl}}{s_k s_l} + R_{kl}\Big[\frac{\sigma_k\sigma_l}{s_k s_l} - 1\Big].$$

On Ω, the first term is at most ρ/a^2 in absolute value, since $s_k, s_l \ge a$. Step 4 places each σ_k/s_k in $[1/(1+\beta), 1/(1-\beta)] \subseteq [1-2\beta, 1+2\beta]$. The product of two such ratios lies in $[(1-2\beta)^2, (1+2\beta)^2]$, so the bracket is at most $4\beta + 4\beta^2 \le 6\beta = 3\rho/a^2$ in absolute value. Adding the two terms proves the claim. The bound $4\rho/a^2$ does not depend on the validation sample.

Step 6. On Ω, the rectangle probabilities of T^* and $\hat{Z}$ differ by at most b, which is (16). Let Y_1 and Y_2 be centered Gaussian vectors in $\mathbb{R}^p$ whose covariances differ entrywise by at most Δ, where every variance of Y_2 is at least $c > 0$. The Gaussian comparison of Chernozhukov et al. (2022, Prop. 2.1) states that

$$\sup_{y \in \mathbb{R}^p} \big|\mathbb{P}(Y_1 \le y) - \mathbb{P}(Y_2 \le y)\big| \le C_P(c)\big(\Delta \log^2 p\big)^{1/2},$$

where $C_P(c)$ depends only on c. The comparison places no condition on the covariance of Y_1, which may be singular or have variances below one.

A closed rectangle $\prod_k [l_k, u_k]$ contains x exactly when $(x, -x) \le (u, -l)$ coordinate-wise. Rectangle events for T^* and $\hat{Z}$ are therefore orthant events for $(T^*, -T^*)$ and $(\hat{Z}, -\hat{Z})$ in $\mathbb{R}^{2K}$. Their covariances have blocks $\pm\hat{R}$ and $\pm R$, so they differ entrywise by at most $\|\hat{R} - R\|_{\max}$, and the second has unit diagonal. We apply the comparison with $p = 2K$ and $c = 1$, and $C_P(1)$ is absolute. Monotone limits of both probabilities extend the comparison from closed bounded rectangles to all rectangles.

Since Step 5 bounds $\|\hat{R} - R\|_{\max}$ by $4\rho/a^2$, this gives on Ω

$$\sup_{\mathcal{B}} \big|\mathbb{P}_e(T^* \in \mathcal{B}) - \mathbb{P}(\hat{Z} \in \mathcal{B})\big| \;\le\; C_P(1)\big((4\rho/a^2)\log^2(2K)\big)^{1/2} \;=\; b, \tag{16}$$

where the supremum is over all rectangles in $\mathbb{R}^K$. Both laws are fixed once $(L, \mathcal{G}_{\mathrm{tr}})$ is fixed, so the bound holds at any rectangle determined by the validation sample and training. It does not apply to a rectangle that depends on the multipliers e.

Step 7. The sampling law of d satisfies the central limit bound (17). The rows $X_{j\cdot}$ with entries $X_{jk} = (L_{jk} - \mu_k)/\sigma_k$ have normalized sum $m^{-1/2}\sum_{j \le m} X_{j\cdot} = d$. By Lemma 3, they are conditionally i.i.d. with mean zero and covariance R, the covariance of $\hat{Z}$. Fix $B_n \ge 1$ and constants $0 < b_1 \le b_2$, and take sample size $n = m$ and dimension $p = K$.

Condition E of Chernozhukov et al. (2022) asks that $\mathbb{E}\exp(|X_{jk}|/B_n) \le 2$ for all j, k. Their Condition M asks that $\mathbb{E}X_{jk}^2 \ge b_1^2$ and $\mathbb{E}X_{jk}^4 \le B_n^2 b_2^2$ for every k. Under both conditions, their Theorem 2.1 bounds $|\mathbb{P}(d \in \mathcal{B} \mid \mathcal{G}_{\mathrm{tr}}) - \mathbb{P}(\hat{Z} \in \mathcal{B})|$ by $C\{B_n^2 \log^5(Km)/m\}^{1/4}$ for every closed rectangle $\mathcal{B}$. The constant C depends only on b_1 and b_2, and neither condition restricts R, which may be singular. Applied at almost every ω, the theorem bounds $\varkappa(\omega, \cdot)$ of Lemma 4.

Take $B_n = M/(a \log 2)$ and $b_1 = b_2 = 1$, so that C is absolute. Since L_{jk} and μ_k lie in $[0, M]$ and each entry has unit variance, both conditions hold:

$$|X_{jk}| \le M/\sigma_k \le M/a, \qquad \mathbb{E}\exp(|X_{jk}|/B_n) \le \exp(\log 2) = 2,$$
$$\mathbb{E}X_{jk}^2 = 1, \qquad \mathbb{E}X_{jk}^4 \le (M/a)^2\,\mathbb{E}X_{jk}^2 = (\log 2)^2 B_n^2 \le B_n^2.$$

A variable in $[0, M]$ has standard deviation at most $M/2$, so $a \le M/2$ and $B_n \ge 2/\log 2 > 1$. Since $Km \le (K \vee m)^2$, we have $\log(Km) \le 2\Lambda$. Absorbing the factors $2^{5/4}$ and $(\log 2)^{-1/2}$ into $C_\diamond$, and extending from closed rectangles by monotone limits, gives

$$\sup_{\mathcal{B}} \left|\mathbb{P}(d \in \mathcal{B} \mid \mathcal{G}_{\mathrm{tr}}) - \mathbb{P}(\hat{Z} \in \mathcal{B})\right| \;\le\; C_\diamond \left\{(M/a)^2 \Lambda^5/m\right\}^{1/4} = \ell, \tag{17}$$

where the supremum is over all rectangles in $\mathbb{R}^K$. Scaling coordinate k by $\sigma_k > 0$ maps rectangles onto rectangles, so the same bound holds for $\sqrt{m}(\hat{\mu} - \mu)$ against $N(0, \Sigma)$.

B.3 Anti-concentration

A *shell* is the region between a rectangle and a smaller rectangle inside it, and its probability is the coverage lost when the rectangle's boundaries move inward. Estimated scales multiply each threshold of (7) by a factor in $[1 - 2\beta, 1 + 2\beta]$, and a threshold equal to its discount can reach $\sqrt{m}\, M/(2a)$. Such a threshold can then move by $\beta\sqrt{m}\, M/a$, which does not vanish as m grows, so the lemma below splits the coordinates by whether their threshold exceeds a fixed $\tau > 0$.

Lemma 6 (Shell). *Let $\hat{Z} \sim N(0, R)$ in $\mathbb{R}^K$ with unit diagonal and otherwise arbitrary R. For every $u \in [0, \infty)^K$, every $\varrho \in [1/2, 1]$ and every $\tau > 0$, with $T = \{k : u_k > \tau\}$,*

$$0 \;\le\; S^{\hat{Z}}(\varrho u) - S^{\hat{Z}}(u) \;=\; \mathbb{P}\big(\hat{Z} \in \mathcal{B}_u \setminus \mathcal{B}_{\varrho u}\big) \;\le\; C_K\,(1 - \varrho)\,\tau + 2|T|\,\bar{\Phi}(\varrho\tau).$$

The bound also holds with $|T|$ replaced by K, and then its right side does not depend on u.

Proof. We split the shell by where a coordinate leaves the inner rectangle $\mathcal{B}_{\varrho u}$. Nazarov's inequality bounds the part where such a coordinate lies in T^c, and a union bound over Gaussian marginals bounds the part where it lies in T.

Step 1. The shell splits into two events. By definition $S^{\hat{Z}}(v) = 1 - \mathbb{P}(\hat{Z} \in \mathcal{B}_v)$, and $\mathcal{B}_{\varrho u} \subseteq \mathcal{B}_u$ since $\varrho \le 1$ and $u \ge 0$. So the displayed equality holds and the quantity is non-negative. On the shell $\mathcal{B}_u \setminus \mathcal{B}_{\varrho u}$, $|\hat{Z}_k| \le u_k$ for every k and $|\hat{Z}_k| > \varrho u_k$ for at least one k. Let E_1 be the part of the shell on which such a coordinate lies in $T^c = [K] \setminus T$, and E_2 the part on which it lies in T. The shell is $E_1 \cup E_2$, so

$$\mathbb{P}\big(\hat{Z} \in \mathcal{B}_u \setminus \mathcal{B}_{\varrho u}\big) \;\le\; \mathbb{P}(E_1) + \mathbb{P}(E_2).$$

Step 2. We show $\mathbb{P}(E_1) \le C_K(1 - \varrho)\tau$. If T^c is empty or $\varrho = 1$, then E_1 is empty and $\mathbb{P}(E_1) = 0$, so assume neither. Write $\mathcal{B}_v^{T^c} = \prod_{k \in T^c}[-v_k, v_k]$, and let $\hat{Z}^{T^c}$ be the subvector of $\hat{Z}$ indexed by T^c, so that $\hat{Z}^{T^c} \in \mathcal{B}_u^{T^c} \setminus \mathcal{B}_{\varrho u}^{T^c}$ on E_1. The stacked vector $\hat{Z}^{\pm}_{T^c} = (\hat{Z}^{T^c}, -\hat{Z}^{T^c})$ is a centered Gaussian vector in $\mathbb{R}^{2|T^c|}$ with unit variances. With $\tilde{u} = (u_{T^c}, u_{T^c})$, we have $\{\hat{Z}^{T^c} \in \mathcal{B}_{\varrho u}^{T^c}\} = \{\hat{Z}^{\pm}_{T^c} \le \varrho\tilde{u}\}$. Each face of $\mathcal{B}_{\varrho u}^{T^c}$ moves out by $(1 - \varrho)u_k \le (1 - \varrho)\tau$, since $u_k \le \tau$ for $k \in T^c$. Raising every coordinate of $\varrho\tilde{u}$ by $(1 - \varrho)\tau$ therefore gives

$$\{\hat{Z}^{T^c} \in \mathcal{B}_u^{T^c}\} \;\subseteq\; \{\hat{Z}^{\pm}_{T^c} \le \varrho\tilde{u} + (1 - \varrho)\tau\mathbf{1}\}.$$

Nazarov's inequality (Chernozhukov et al., 2017b, Theorem 1) states that every centered Gaussian vector Y in $\mathbb{R}^p$ with unit variances, singular covariance allowed, satisfies

$$\mathbb{P}(Y \le y + x\mathbf{1}) - \mathbb{P}(Y \le y) \;\le\; x\big(\sqrt{2\log p} + 2\big) \qquad \text{for every } y \in \mathbb{R}^p \text{ and } x > 0.$$

We apply it to $\hat{Z}^{\pm}_{T^c}$, whose covariance is singular, with $p = 2|T^c|$, $y = \varrho\tilde{u}$ and $x = (1 - \varrho)\tau > 0$:

$$\begin{aligned}\mathbb{P}(E_1) \;&\le\; \mathbb{P}(\hat{Z}^{T^c} \in \mathcal{B}_u^{T^c}) - \mathbb{P}(\hat{Z}^{T^c} \in \mathcal{B}_{\varrho u}^{T^c}) \;\le\; \mathbb{P}\big(\hat{Z}^{\pm}_{T^c} \le \varrho\tilde{u} + (1 - \varrho)\tau\mathbf{1}\big) - \mathbb{P}\big(\hat{Z}^{\pm}_{T^c} \le \varrho\tilde{u}\big) \\ &\le\; (1 - \varrho)\tau\big(\sqrt{2\log(2|T^c|)} + 2\big) \;\le\; C_K\,(1 - \varrho)\,\tau.\end{aligned}$$

The first inequality holds because E_1 lies in the difference of the two events, and the second follows from the two event relations above. The third is Nazarov's inequality, and the last uses $|T^c| \le K$ and the definition of C_K.

Step 3. We show $\mathbb{P}(E_2) \le 2|T|\bar{\Phi}(\varrho\tau)$. Since $u_k > \tau$ for $k \in T$, on E_2 some $k \in T$ has $|\hat{Z}_k| > \varrho u_k \ge \varrho\tau$. Each marginal is $\hat{Z}_k \sim N(0,1)$, so a union bound over the $|T|$ coordinates gives

$$\mathbb{P}(E_2) \;\le\; \sum_{k\in T} \mathbb{P}(|\hat{Z}_k| > \varrho\tau) \;=\; 2|T|\,\bar{\Phi}(\varrho\tau).$$

Step 4. Adding the bounds of Steps 2 and 3 in the inequality of Step 1 gives the claim, and $|T| \le K$ gives the version with K. □

A calibrated shell bound. The proof of Theorem 13 needs a shell bound that depends on neither u nor ϱ, which (18) obtains by fixing τ in Lemma 6. Assume $\beta \le 1/8$, as in Theorem 13.

Step 1. We fix the constants and state the bound. Put $\varrho_0 = 1-4\beta$, so that $\varrho_0 \in [1/2, 1]$ and $\varrho_0^{-1} \le 2$. For $\epsilon \in (0,1)$ put $\tau_\epsilon = \varrho_0^{-1}\sqrt{2\log(2K/\epsilon)}$, and for $u \in [0,\infty)^K$ let $T = \{k : u_k > \tau_\epsilon\}$. For every $u \in [0,\infty)^K$ and every $\varrho \in [\varrho_0, 1]$, Lemma 6 with $\tau = \tau_\epsilon$ gives the first inequality of

$$\begin{aligned} S^{\hat{Z}}(\varrho u) - S^{\hat{Z}}(u) \;&\le\; C_K(1-\varrho)\tau_\epsilon + 2|T|\,\bar{\Phi}(\varrho\tau_\epsilon) \\ &\le\; 8\,C_K\,\beta\sqrt{2\log(2K/\epsilon)} + \epsilon/2 \;=:\; H, \end{aligned} \tag{18}$$

and Steps 2 and 3 below prove the second by bounding the two terms separately.

Step 2. The first term is at most $8C_K\beta\sqrt{2\log(2K/\epsilon)}$. Since $1-\varrho \le 1-\varrho_0 = 4\beta$ and $\varrho_0^{-1} \le 2$,

$$C_K(1-\varrho)\tau_\epsilon \;\le\; C_K \cdot 4\beta \cdot 2\sqrt{2\log(2K/\epsilon)} \;=\; 8C_K\beta\sqrt{2\log(2K/\epsilon)}.$$

Step 3. The second term is at most $\epsilon/2$. Since $\bar{\Phi}$ decreases and $\varrho\tau_\epsilon \ge \varrho_0\tau_\epsilon = \sqrt{2\log(2K/\epsilon)}$, the bound $\bar{\Phi}(x) \le \frac{1}{2}e^{-x^2/2}$ for $x \ge 0$ gives

$$\bar{\Phi}(\varrho\tau_\epsilon) \;\le\; \bar{\Phi}\big(\sqrt{2\log(2K/\epsilon)}\big) \;\le\; \tfrac{1}{2}\,e^{-\log(2K/\epsilon)} \;=\; \epsilon/(4K).$$

With $|T| \le K$, the second term is at most $2K \cdot \epsilon/(4K) = \epsilon/2$.

Step 4. The two applications cost $2H = g$. Because $S^{\hat{Z}}$ is a fixed function given $\mathcal{G}_{\mathrm{tr}}$, bound (18) also holds at data-determined u. The proof of Theorem 13 applies it twice, with $\varrho = 1/(1+2\beta)$ and $\varrho = (1-2\beta)^2$, both in $[\varrho_0, 1]$ when $\beta \le 1/8$, so the two applications cost at most $2H = g$ of (12).

B.4 Containment of the joint confidence set

At a positive floor, whenever the procedure does not abstain, Theorem 10 places every selected-risk value permitted by $\hat{C}_\alpha$ in the accepted set $\hat{C}_\alpha^W$. Lemmas 7–9 supply the steps of its proof.

Recall $s_k = \hat{\sigma}_k \vee a \ge a > 0$, the constant columns $Z = \{k : \hat{\sigma}_k = 0\}$ and $P = [K] \setminus Z$. Lemma 8 evaluates threshold vectors at a minimizer that need not be the smallest index attaining the minimum, so we index them by the minimizer. With the gap $\Delta_k(\mu)$ of Appendix A.1, for $\mu \in \mathbb{R}^K$ and any $i \in \arg\min_j \mu_j$ set

$$G_k(\mu,i) = \frac{\Delta_k(\mu)}{s_k + s_i}, \qquad F(r;\mu,i)_k = \max\{r, \sqrt{m}\,G_k(\mu,i)\},$$

$$\hat{r}_\alpha(\mu,i) = \inf\{r \ge 0 : \hat{S}^{[K]}(F(r;\mu,i)) \le \alpha\} = \psi^{[K]}(\sqrt{m}\,G(\mu,i)),$$

$$w_k(\mu,i) = \max\Big\{\frac{\hat{r}_\alpha(\mu,i)s_k}{\sqrt{m}},\; \frac{\Delta_k(\mu)s_k}{s_k+s_i}\Big\},$$

so that $w_k(\mu,i) = s_k F(\hat{r}_\alpha(\mu,i);\mu,i)_k/\sqrt{m}$. Since i minimizes μ and $s_k + s_i \ge 2a$, the vector $G(\mu,i)$ is well defined and non-negative. The second expression for $\hat{r}_\alpha(\mu,i)$ follows from $F(r;\mu,i) = r\mathbf{1} \vee \sqrt{m}\,G(\mu,i)$ and the definition of $\psi^{[K]}$.

At $i = i^*(\mu)$ the vector $\sqrt{m}\, G(\mu, i)$ is $c(\mu)$ of (7). Hence $F(r; \mu, i^*(\mu)) = F(r, \mu)$, and $\hat{r}_\alpha(\mu, i^*(\mu))$ is the radius that $\hat{C}_\alpha^{[K]}$ uses. Multiplying the kth inequality of (8) by $s_k/\sqrt{m}$ gives

$$\hat{C}_\alpha^{[K]} = \{\mu : |\hat{\mu}_k - \mu_k| \le w_k(\mu, i^*(\mu))\ \forall k\}.$$

We call the kth inequality in this set *admissibility at* k, and $w_k(\mu, i^*(\mu))$ its half-width. The set uses only the index $i^*(\mu)$, and other minimizers enter only the comparison below. The band's radius of Section 2.2 is $r_{\mathrm{sim}} = \psi(0)$, equal to $\psi^{[K]}(0)$ by (13).

Admissibility yields the discount formula (1) in two steps. Let $\mu \in \hat{C}_\alpha^{[K]}$ with $\mu_{\hat{k}} = t$, set $i = i^*(\mu)$, so $\mu_i \le t$, and write $G_k = G_k(\mu, i)$. The formula bounds $\sqrt{m}\, G_k$ from below wherever $\sqrt{m}\, G_k \ge \hat{r}_\alpha(\mu, i)$.

Step 1. The estimation error and the gap add up. Take a coordinate $k \notin \{\hat{k}, i\}$ at which the second entry of $w_k(\mu, i)$ is the larger. Admissibility at k bounds the estimation error $\hat{\mu}_k - \mu_k$ by $G_k s_k$, and the gap satisfies $\mu_k - \mu_i = G_k(s_k + s_i)$ by definition. Adding the two gives

$$\hat{\mu}_k - \mu_i = (\hat{\mu}_k - \mu_k) + (\mu_k - \mu_i) \ \le\ G_k s_k + G_k(s_k + s_i) = G_k(2s_k + s_i).$$

Step 2. The discount with reference i. Since $G_k \ge 0$ and $\mu_i \le t$, Step 1 yields

$$\sqrt{m}\, G_k \ \ge\ \frac{\sqrt{m}(\hat{\mu}_k - t)_+}{2s_k + s_i},$$

the formula with reference i. Lemma 8 treats the coordinates at which the first entry is the larger. The discount is zero at $\hat{k}$ and at i, so masking leaves their entries unchanged. Theorem 10 proves that these discounts suffice.

Lemma 7 (Monotone comparison)**.** *Take $\hat{S}^{[K]}$, coordinate-wise non-increasing by Lemma 5, and its radius $\psi^{[K]}$ from Appendix B.1. The lemma and its proof also hold for $\hat{S}_B^{[K]}$ and $\psi_B^{[K]}$, with α_B in place of α. Fix $x \in [0, \infty)^K$ with $r := \psi^{[K]}(x) < \infty$, and let $x' \in [0, \infty)^K$ satisfy, for every k, either $x'_k \le x_k$ or $x'_k < r$. Then $\psi^{[K]}(x') \ge r$.*

Proof. We show that every $r' \in [0, r)$ has $\hat{S}^{[K]}(r'\mathbf{1} \vee x') > \alpha$. There is nothing to show when $r = 0$.

Step 1. A larger radius below r. Fix $r' \in [0, r)$ and put

$$J = \{k : x'_k > x_k\}, \qquad r'' = \max\big(\{r'\} \cup \{x'_k\}_{k \in J}\big).$$

Every $k \in J$ has $x'_k < r$ by hypothesis, so r'' lies in $[0, r)$ as the maximum of finitely many numbers in that interval.

Step 2. Coordinate-wise comparison: $r'\mathbf{1} \vee x' \le r''\mathbf{1} \vee x$. For $k \notin J$ this follows from $r' \le r''$ and $x'_k \le x_k$. For $k \in J$ both r' and x'_k are at most r''.

Step 3. Monotonicity. Since $\hat{S}^{[K]}$ is coordinate-wise non-increasing, Step 2 gives

$$\hat{S}^{[K]}(r'\mathbf{1} \vee x') \ \ge\ \hat{S}^{[K]}(r''\mathbf{1} \vee x) \ >\ \alpha,$$

where the strict inequality holds because $r'' \in [0, r)$ lies below the infimum $r = \psi^{[K]}(x)$. Hence the feasible set for x' lies in $[r, \infty)$, which gives $\psi^{[K]}(x') \ge r$. □

The strict inequality $x'_k < r$ cannot be relaxed to $x'_k \le r$: when $r \in (0, \infty)$, which holds at $x = 0$ whenever $r_{\mathrm{sim}} > 0$, the vector $x' = x \vee r\mathbf{1}$ meets the weakened hypothesis, yet Lemma 5 gives $\hat{S}^{[K]}(x') \le \alpha$ and so $\psi^{[K]}(x') = 0 < r$.

Lemma 8 (Two alternatives for each gap)**.** *Let $\mu \in \hat{C}_\alpha^{[K]}$, set $i = i^*(\mu)$, and put $t = \mu_{\hat{k}}$, $r = \hat{r}_\alpha(\mu, i)$ and*

$$\mu_j^w = \begin{cases} t, & j = \hat{k}, \\ \mu_i, & j = i, \\ \max\left\{\dfrac{\hat{\mu}_j(s_j + s_i) + \mu_i s_j}{2s_j + s_i},\ \mu_i\right\}, & \text{otherwise.} \end{cases}$$

Then $i \in \arg\min_j \mu_j^w$, and for every k either $G_k(\mu^w, i) \le G_k(\mu, i)$ or $\sqrt{m}\, G_k(\mu^w, i) < r$.

Proof. We prove the first claim from the construction and then check the two alternatives one coordinate at a time.

Step 1. The index i minimizes μ^w. When $\hat{k} = i$ the first two cases defining μ^w agree, since then $t = \mu_i$. Because i minimizes μ, the coordinate $\mu^w_{\hat{k}} = t$ is at least μ_i, and the maximum in the third case makes every other coordinate at least μ_i. Together with $\mu^w_i = \mu_i$ this gives the first claim.

Step 2. The coordinates $\hat{k}$ and i satisfy the first alternative. At $\hat{k}$ the gap is unchanged, since $\Delta_{\hat{k}}(\mu^w) = t - \mu_i = \Delta_{\hat{k}}(\mu)$, and at i both gaps vanish.

Step 3. For the remaining coordinates, the case split reduces to comparing $\hat{\mu}_k$ with an increasing map Ψ_k. Fix $k \notin \{\hat{k}, i\}$ and let $\Psi_k(y) = y + (y - \mu_i)s_k/(s_k + s_i)$, which is strictly increasing because $s_k, s_i > 0$. The inequality $\hat{\mu}_k - \mu_k \le \Delta_k(\mu)s_k/(s_k + s_i)$, which compares the estimation error with the second entry of $w_k(\mu, i)$, is equivalent to $\hat{\mu}_k \le \Psi_k(\mu_k)$. Put $\bar{\mu}_k = [\hat{\mu}_k(s_k + s_i) + \mu_i s_k]/(2s_k + s_i)$, so that $\mu^w_k = \max\{\bar{\mu}_k, \mu_i\}$. Direct algebra gives $\Psi_k(\bar{\mu}_k) = \hat{\mu}_k$, so $\bar{\mu}_k$ is the smallest y with $\hat{\mu}_k \le \Psi_k(y)$.

Step 4. The case $\hat{\mu}_k \le \Psi_k(\mu_k)$ satisfies the first alternative. Monotonicity of Ψ_k gives $\bar{\mu}_k \le \mu_k$, hence $\mu^w_k \le \max\{\mu_k, \mu_i\} = \mu_k$ and $\Delta_k(\mu^w) \le \Delta_k(\mu)$. Both gaps share the denominator $s_k + s_i$, so $G_k(\mu^w, i) \le G_k(\mu, i)$, the first alternative.

Step 5. The case $\hat{\mu}_k > \Psi_k(\mu_k)$ gives the second alternative. Here $\Psi_k(\mu_k) = \mu_k + \Delta_k(\mu)s_k/(s_k + s_i)$, and $\bar{\mu}_k > \mu_k$ because Ψ_k is strictly increasing with $\Psi_k(\bar{\mu}_k) = \hat{\mu}_k$. The estimation error $\hat{\mu}_k - \mu_k$ then exceeds the second entry of $w_k(\mu, i)$. Admissibility at k therefore forces $w_k(\mu, i) = rs_k/\sqrt{m}$ and gives $\hat{\mu}_k \le \mu_k + rs_k/\sqrt{m}$.

Step 6. In the case of Step 5, the coordinate k satisfies $\sqrt{m}\, G_k(\mu^w, i) < r$. The bound on $\hat{\mu}_k$ gives

$$\bar{\mu}_k + (\bar{\mu}_k - \mu_i)\frac{s_k}{s_k + s_i} \;=\; \hat{\mu}_k \;\le\; \mu_k + \frac{rs_k}{\sqrt{m}} \;<\; \bar{\mu}_k + \frac{rs_k}{\sqrt{m}},$$

where the equality is $\Psi_k(\bar{\mu}_k) = \hat{\mu}_k$ and the strict inequality is $\mu_k < \bar{\mu}_k$. Subtracting $\bar{\mu}_k$ and dividing by $s_k > 0$ gives $\sqrt{m}(\bar{\mu}_k - \mu_i)/(s_k + s_i) < r$. Since $\bar{\mu}_k > \mu_k \ge \mu_i$, the third case gives $\mu^w_k = \bar{\mu}_k$. The last inequality is therefore $\sqrt{m}\, G_k(\mu^w, i) < r$, the second alternative. □

Lemma 9 (Selected-coordinate half-width)**.** *Let $\mu \in \hat{C}^{[K]}_\alpha$ and set $i = i^*(\mu)$. Then $\sqrt{m}\,\Delta_{\hat{k}}(\mu)/(s_{\hat{k}} + s_i) \le \hat{r}_\alpha(\mu, i)$, so $w_{\hat{k}}(\mu, i) = \hat{r}_\alpha(\mu, i)s_{\hat{k}}/\sqrt{m}$.*

Proof. Write $r = \hat{r}_\alpha(\mu, i)$. We prove the inequality by contradiction and then read off the half-width.

Step 1. The contrary assumption fixes $w_{\hat{k}}(\mu, i)$. Suppose that $\sqrt{m}\,\Delta_{\hat{k}}(\mu)/(s_{\hat{k}} + s_i) > r$. Then $\Delta_{\hat{k}}(\mu)s_{\hat{k}}/(s_{\hat{k}} + s_i) > rs_{\hat{k}}/\sqrt{m}$, so $w_{\hat{k}}(\mu, i) = \Delta_{\hat{k}}(\mu)s_{\hat{k}}/(s_{\hat{k}} + s_i)$.

Step 2. A chain through $\hat{\mu}_{\hat{k}}$. Since $\Delta_i(\mu) = 0$, admissibility at i gives $\hat{\mu}_i \le \mu_i + rs_i/\sqrt{m}$. With admissibility at $\hat{k}$ and the minimality of $\hat{\mu}_{\hat{k}}$, this gives

$$\mu_i + \frac{\Delta_{\hat{k}}(\mu)s_i}{s_{\hat{k}} + s_i} \;=\; \mu_{\hat{k}} - w_{\hat{k}}(\mu, i) \;\le\; \hat{\mu}_{\hat{k}} \;\le\; \hat{\mu}_i \;\le\; \mu_i + \frac{rs_i}{\sqrt{m}}.$$

Step 3. Conclusion. Subtracting μ_i and dividing by $s_i > 0$ gives $\sqrt{m}\,\Delta_{\hat{k}}(\mu)/(s_{\hat{k}} + s_i) \le r$, a contradiction. Hence $\Delta_{\hat{k}}(\mu)s_{\hat{k}}/(s_{\hat{k}} + s_i) \le rs_{\hat{k}}/\sqrt{m}$, and the definition of $w_{\hat{k}}(\mu, i)$ as a maximum gives $w_{\hat{k}}(\mu, i) = rs_{\hat{k}}/\sqrt{m}$. □

Theorem 10 (Containment)**.** *Let $\mathrm{proj}_{\hat{k}}(\hat{C}^{[K]}_\alpha) = \{\mu_{\hat{k}} : \mu \in \hat{C}^{[K]}_\alpha\}$, and let $\hat{C}^W_\alpha$ use the tail $\hat{S}$ of Section 2.2. At any positive floor a, on $\{\hat{k} \in P\}$,*

$$\mathrm{proj}_{\hat{k}}(\hat{C}^{[K]}_\alpha) \;\subseteq\; \hat{C}^W_\alpha \;\subseteq\; \mathrm{hull}(\hat{C}^W_\alpha).$$

Hence, at the population risk vector μ,

$$\{\mu \in \hat{C}^{[K]}_\alpha\} \;\subseteq\; \{\hat{k} \in Z\} \cup \{\hat{k} \in P,\ \mu_{\hat{k}} \in \hat{C}^W_\alpha\}.$$

In particular, $\mathbb{P}(\hat{k} \in P,\ \mu_{\hat{k}} \notin \hat{C}^W_\alpha) \le \mathbb{P}(\mu \notin \hat{C}^{[K]}_\alpha)$ under any probability measure.

Proof. The proof shows $\hat{r}_\alpha(\mu, i^*(\mu)) \le \hat{r}_\alpha^W(\mu_{\hat{k}})$ through a threshold vector that depends on μ only through $\mu_{\hat{k}}$ and $i^*(\mu)$, and Lemma 9 turns this radius bound into membership of $\mu_{\hat{k}}$. There is nothing to prove on $\{\hat{k} \in Z\}$, where the procedure abstains. Work on $\{\hat{k} \in P\}$, let $\mu \in \hat{C}_\alpha^{[K]}$ and set $i = i^*(\mu)$. Put $t = \mu_{\hat{k}}$, $r = \hat{r}_\alpha(\mu, i)$ and μ^w as in Lemma 8. Then $\mu_i \le t$ because i minimizes μ.

Step 1. The radius $\psi^{[K]}$ is non-increasing and finite, and $r \le r_{\mathrm{sim}}$. Let $y \le y'$ coordinate-wise. Then $r'\mathbf{1} \vee y \le r'\mathbf{1} \vee y'$ at every $r' \ge 0$, and Lemma 5 gives $\hat{S}^{[K]}(r'\mathbf{1} \vee y') \le \hat{S}^{[K]}(r'\mathbf{1} \vee y)$. Every radius feasible for y is therefore feasible for y', so $\psi^{[K]}(y') \le \psi^{[K]}(y)$. The radius $\psi^{[K]}$ is finite because $\hat{S}^{[K]}(r'\mathbf{1} \vee y) \to 0$ as $r' \to \infty$ by the same lemma and $\alpha > 0$. Since $G(\mu, i) \ge 0$, these give $r \le \psi^{[K]}(0) = r_{\mathrm{sim}} < \infty$.

Step 2. The comparison vector $x' = \sqrt{m}\, G(\mu^w, i)$ has $\psi^{[K]}(x') \ge r$. The vector x' is non-negative because $i \in \arg\min_j \mu_j^w$ by Lemma 8. With $r < \infty$ from Step 1, that lemma verifies the remaining hypothesis of Lemma 7 for $x = \sqrt{m}\, G(\mu, i)$ and this x', and Lemma 7 gives $\psi^{[K]}(x') \ge r$.

Step 3. Replacing μ_i by t in x' can only lower the thresholds. For $j \notin \{\hat{k}, i\}$, subtracting μ_i from the third case of μ^w gives

$$\Delta_j(\mu^w) = \frac{\max\{(\hat{\mu}_j - \mu_i)(s_j + s_i), 0\}}{2s_j + s_i} = \frac{(\hat{\mu}_j - \mu_i)_+(s_j + s_i)}{2s_j + s_i}.$$

Dividing by $s_j + s_i$ and multiplying by $\sqrt{m}$ gives $\sqrt{m}\, G_j(\mu^w, i) = \sqrt{m}(\hat{\mu}_j - \mu_i)_+/(2s_j + s_i)$. Define the auxiliary threshold vector on all K coordinates

$$\nu_j^{(i)}(t) = \begin{cases} 0, & j \in \{\hat{k}, i\}, \\ \sqrt{m}(\hat{\mu}_j - t)_+/(2s_j + s_i), & j \notin \{\hat{k}, i\}. \end{cases}$$

At $j \notin \{\hat{k}, i\}$ we have $\sqrt{m}\, G_j(\mu^w, i) \ge \nu_j^{(i)}(t)$ because $\hat{\mu}_j - \mu_i \ge \hat{\mu}_j - t$ and the positive part is non-decreasing. At $j = \hat{k}$ the threshold $\nu_{\hat{k}}^{(i)}(t)$ is zero and $\sqrt{m}\, G_{\hat{k}}(\mu^w, i) \ge 0$, and at $j = i$ both sides vanish. Hence $\nu^{(i)}(t) \le x'$ coordinate-wise, and Steps 1 and 2 give

$$\psi^{[K]}(\nu^{(i)}(t)) \;\ge\; \psi^{[K]}(x') \;\ge\; r. \tag{19}$$

Step 4. It remains to show $\hat{r}_\alpha^W(t) \ge \psi^{[K]}(\nu^{(i)}(t))$, first when $i \in P$. Then $\nu^{(i)}(t)_P = g^{(i)}(t)$. Since $\Delta_i(\mu) = 0$, the half-width at i is $w_i(\mu, i) = r s_i/\sqrt{m}$. Admissibility at i, $\mu_i \le t$ and $r \le r_{\mathrm{sim}}$ give

$$\hat{\mu}_i \;\le\; \mu_i + \frac{r s_i}{\sqrt{m}} \;\le\; t + \frac{r_{\mathrm{sim}} s_i}{\sqrt{m}}.$$

Thus $i \in W(t)$, and (13) and (19) give

$$\hat{r}_\alpha^W(t) \;\ge\; \psi(g^{(i)}(t)) \;=\; \psi^{[K]}(\nu^{(i)}(t)) \;\ge\; r.$$

Step 5. The same bound when $i \in Z$. Then $i \notin W(t)$ because $W(t) \subseteq P$, and the proof uses $\hat{k}$ in its place, which lies in $W(t)$ by definition. The scales satisfy $s_i = \hat{\sigma}_i \vee a = a \le s_{\hat{k}}$. Both $g^{(\hat{k})}(t)$ and $\nu^{(i)}(t)_P$ force their $\hat{k}$ coordinate to zero, and every other coordinate $j \in P \setminus \{\hat{k}\}$ satisfies

$$g_j^{(\hat{k})}(t) = \frac{\sqrt{m}(\hat{\mu}_j - t)_+}{2s_j + s_{\hat{k}}} \;\le\; \frac{\sqrt{m}(\hat{\mu}_j - t)_+}{2s_j + s_i} = \nu_j^{(i)}(t).$$

Coordinate-wise monotonicity, (13) and (19) yield

$$\hat{r}_\alpha^W(t) \;\ge\; \psi(g^{(\hat{k})}(t)) \;\ge\; \psi(\nu^{(i)}(t)_P) \;=\; \psi^{[K]}(\nu^{(i)}(t)) \;\ge\; r.$$

Step 6. Conclusion. In either case, admissibility at $\hat{k}$ and Lemma 9 give

$$|\hat{\mu}_{\hat{k}} - t| \;\le\; \frac{r\, s_{\hat{k}}}{\sqrt{m}} \;\le\; \frac{\hat{r}_\alpha^W(t)\, s_{\hat{k}}}{\sqrt{m}}.$$

Hence $t \in \hat{C}_\alpha^W$, which is the first inclusion. The second inclusion holds because every set lies in its hull. On $\{\hat{k} \in P\}$, the event inclusion is the first inclusion at the population risk vector μ. □

We now prove the claims of Section 2.3 about the shape of $\hat{C}_\alpha^W$. Write $\mathcal{R}(t) = \hat{r}_\alpha^W(t)s_{\hat{k}}/\sqrt{m}$ for the half-width at t, so that $t \in \hat{C}_\alpha^W$ exactly when $|\hat{\mu}_{\hat{k}} - t| \le \mathcal{R}(t)$.

Step 1. The half-width is non-decreasing in t. Each discount $g_j^{(i)}(t)$ in (1) is non-increasing in t. A larger threshold vector zeroes more entries, so ψ is non-increasing in it, and each $\psi(g^{(i)}(t))$ is non-decreasing in t. The condition defining $W(t)$ in (2) weakens as t grows, so $W(t)$ grows with t. A maximum of non-decreasing functions over a growing set is non-decreasing. The same steps apply to ψ_B, with $\psi_B(0)$ in place of r_{sim} in (2).

Step 2. Below $\hat{\mu}_{\hat{k}}$ the accepted values form an interval. For $t \le \hat{\mu}_{\hat{k}}$, the value t is accepted when $\hat{\mu}_{\hat{k}} - t \le \mathcal{R}(t)$. The left side decreases in t while the right side does not, so the accepted values below $\hat{\mu}_{\hat{k}}$ form an interval ending at $\hat{\mu}_{\hat{k}}$.

Step 3. Above $\hat{\mu}_{\hat{k}}$ acceptance can fail and resume. For $t > \hat{\mu}_{\hat{k}}$, acceptance reads $t - \hat{\mu}_{\hat{k}} \le \mathcal{R}(t)$ with both sides non-decreasing. Acceptance can then fail and resume where $\mathcal{R}$ grows faster than the distance. The half-width can jump when a reference enters $W(t)$, and under ψ_B also when a discount crosses a stored entry of A.

Step 4. An example under ψ. Take $K = 3$, $\hat{R} = I$, $\alpha = 0.1$, $s = (1, 10, 1)$ and $\sqrt{m}(\hat{\mu} - \hat{\mu}_1\mathbf{1}) = (0, 23.2, 20)$. We express candidate values relative to $\hat{\mu}_1$. Then $\hat{k} = 1$ and $r_{\mathrm{sim}} = 2.114$. Checkpoint 2 enters $W(t)$ at $\sqrt{m}(t - \hat{\mu}_1) = 2.059$, where the radius rises from 1.949 to r_{sim}. The accepted values of $\sqrt{m}(t - \hat{\mu}_1)$ are $[-1.949, 1.949] \cup [2.059, 2.114]$. A search that stops at the first rejected value can therefore miss the outer endpoint. The algorithm returns the outermost accepted grid points, and its exact endpoint search keeps them without assuming that the accepted set is connected.

B.5 Grid discretization

The grid lemma below places accepted grid points on both sides of $\mu_{\hat{k}}$ whenever the studentized errors $\hat{d}$ lie in a slightly shrunk rectangle. It needs neither continuity of the radius nor connectedness of the accepted set. It places two conditions on the realized grid $\mathcal{T}$ and search interval (3), which may both depend on the bootstrap draws. The *span condition* requires the hull of $\mathcal{T}$ to contain the search interval, and the *spacing condition* requires consecutive grid points to lie at most h apart. The margin γ below is the spacing bound h in studentized units at the floor.

Lemma 11 (Grid). *Fix a candidate value $\mu \in \mathbb{R}^K$ and a positive floor a. For a spacing bound h, put $\gamma = \sqrt{m}\,h/a$. Let $v = F(\hat{r}_\alpha(\mu), \mu)$ be the threshold vector in (15). Its smallest entry is $v_{i^*(\mu)} = \hat{r}_\alpha(\mu)$, because the gap at $i^*(\mu)$ is zero. Suppose the span and spacing conditions hold. Then on a nonabstaining realization with $\gamma \le \hat{r}_\alpha(\mu)$ and $\hat{d} \in \mathcal{B}_{v-\gamma\mathbf{1}}$, the interval I_m of (4) contains $\mu_{\hat{k}}$.*

Proof. Steps 1 and 2 show that every value within h of $\mu_{\hat{k}}$ is accepted, and Step 3 finds grid points in that range on either side of $\mu_{\hat{k}}$.

Step 1. Every common shift $\mu' = \mu + c\mathbf{1}$ with $|c| \le h$ lies in $\hat{C}_\alpha^{[K]}$. A common shift leaves every gap and the smallest index attaining the minimum unchanged, so

$$i^*(\mu') = i^*(\mu), \qquad F(\cdot, \mu') = F(\cdot, \mu), \qquad \hat{r}_\alpha(\mu') = \hat{r}_\alpha(\mu).$$

Comparing the definitions of F and w then gives $w_k(\mu', i^*(\mu')) = w_k(\mu, i^*(\mu)) = v_k s_k/\sqrt{m}$ for every k. Multiplying $|\hat{d}_k| \le v_k - \gamma$ by $s_k/\sqrt{m}$ and using $s_k \ge a$ gives

$$|\hat{\mu}_k - \mu_k| \;\le\; \frac{v_k s_k}{\sqrt{m}} - \frac{h\, s_k}{a} \;\le\; w_k(\mu, i^*(\mu)) - h.$$

Hence, for every k,

$$|\hat{\mu}_k - \mu'_k| \;\le\; |\hat{\mu}_k - \mu_k| + |c| \;\le\; w_k(\mu', i^*(\mu')),$$

where the first relation is the triangle inequality and the second combines the previous display, $|c| \le h$ and the identity for w_k. Thus $\mu' \in \hat{C}_\alpha^{[K]}$.

Step 2. The interval $[\mu_{\hat{k}} - h, \mu_{\hat{k}} + h]$ lies in $\hat{C}_\alpha^W$ and in the search interval. A nonabstaining realization has $\hat{k} \in P$, so Theorem 10 applies to each μ' of Step 1 and gives $\mu'_{\hat{k}} = \mu_{\hat{k}} + c \in \hat{C}_\alpha^W$ for every $|c| \le h$. By (3), the interval lies in the search interval.

Step 3. Two accepted grid points bracket $\mu_{\hat{k}}$. Since $\mu_{\hat{k}}$ lies in the search interval by Step 2, the span condition places it in the hull of the finite grid $\mathcal{T}$, so the grid points

$$t' = \max\{t \in \mathcal{T} : t \le \mu_{\hat{k}}\}, \qquad t'' = \min\{t \in \mathcal{T} : t \ge \mu_{\hat{k}}\}$$

exist. They are equal or consecutive grid points, so the spacing condition gives $t'' - t' \le h$. Both therefore lie in $[\mu_{\hat{k}} - h, \mu_{\hat{k}} + h]$ and, by Step 2, are accepted. Their hull $[t', t'']$ contains $\mu_{\hat{k}}$ and lies in I_m, the hull of the accepted grid values (4). □

B.6 Finite bootstrap draws

Theorem 13 in the next subsection also covers the interval computed from B draws. For that case, this subsection transfers Theorem 10 and Lemma 11 to the empirical tail (14) and proves Lemma 12. The B multiplier vectors are independent of each other and of the data, so given L the rows $(T_k^{*(b)})_{k\in P}$, $b = 1, \dots, B$, are independent copies of T^*. Every threshold vector reuses the same array A of Section 2.2. Assume $B \ge \lceil 1/\alpha \rceil - 1$, as in Section 2.2. Then the rank $k_\alpha(B)$ of (9) satisfies $0 \le k_\alpha(B) \le B - 1$, and the bound in (9) places the level $\alpha_B = (B - k_\alpha(B) - 1)/B$ of Appendix A.2 in $[0, \alpha)$.

The B-draw radius $\psi_B(y)$ of Section 2.2 attains the defining infimum for $(\hat{S}_B, \alpha_B)$:

$$\psi_B(y) = \min\{r \ge 0 : \hat{S}_B(r\mathbf{1} \vee y) \le \alpha_B\}, \qquad y \in [0, \infty)^P,$$

even when masked maxima tie.

Step 1. The tail counts masked maxima. Fix $y \in [0, \infty)^P$ and write

$$U_b(y) = \max_{k\in P} |T_k^{*(b)}| \, \mathbf{1}\{|T_k^{*(b)}| > y_k\}$$

for the masked maximum of draw b. For $r \ge 0$, some $|T_k^{*(b)}|$ exceeds $\max\{r, y_k\}$ exactly when $U_b(y) > r$, so $\hat{S}_B(r\mathbf{1} \vee y)$ is the fraction of draws b with $U_b(y) > r$.

Step 2. The rank of $\psi_B(y)$ bounds how many masked maxima exceed it and how many exceed each smaller r. The radius $\psi_B(y)$ is the value at zero-based rank $k_\alpha(B)$ of $U_1(y), \dots, U_B(y)$ sorted in increasing order. Only ranks $k_\alpha(B) + 1, \dots, B - 1$ can exceed $\psi_B(y)$, so at most $B - k_\alpha(B) - 1$ masked maxima exceed it. Ranks $k_\alpha(B), \dots, B-1$ hold values at least $\psi_B(y)$, so at least $B - k_\alpha(B)$ masked maxima exceed every $r < \psi_B(y)$.

Step 3. The two counts give the identity. Dividing both by B gives, for every $0 \le r < \psi_B(y)$,

$$\hat{S}_B(\psi_B(y)\mathbf{1} \vee y) \le \alpha_B < \frac{B - k_\alpha(B)}{B} \le \hat{S}_B(r\mathbf{1} \vee y),$$

so $\psi_B(y)$ belongs to the set in the identity and no smaller $r \ge 0$ does.

By Lemma 5, $\hat{S}_B$ is non-increasing, continuous from above and vanishing at infinity, so every definition of the method applies to the pair $(\hat{S}_B, \alpha_B)$. Write $r_{\mathrm{sim},B} = \psi_B(0)$ for the band's radius, and $\hat{r}_{\alpha,B}(\mu) = \psi_B^{[K]}(c(\mu))$ and $\hat{C}_{\alpha,B}^{[K]}$ for the radius of (7) and the confidence set of (8), all computed from the B draws. The radius $\hat{r}_{\alpha,B}^W$ of Appendix A.3 and the hull I_m of (4) use the same draws.

Theorem 10 and Lemma 11 hold for $(\hat{S}_B, \alpha_B)$ on every realization of A. Their proofs use four properties of the tail and its level, and each holds for this pair.

1. The empirical tail is non-increasing in each coordinate by its definition.
2. By the identity above, each value $\psi_B(y)$ is the infimum that defines a radius, with the common level α_B in place of α.
3. Since $\alpha_B \ge 0$ and $\hat{S}_B$ vanishes above the largest entry of A, every such infimum is finite.

4. With the paragraph after (14), the identity gives $\psi_B^{[K]}(y) = \psi_B(y_P)$, the analog of (13) for ψ_B.

The proof of Theorem 10 compares $\psi_B^{[K]}(\sqrt{m}\, G(\mu, i))$ with $\psi_B^{[K]}(\nu^{(i)}(t))$ and $\psi_B(g^{(i)}(t))$, which is valid because all three read the one array A. In Lemma 11 the threshold vector becomes $v = F(\hat{r}_{\alpha,B}(\mu), \mu)$ and the search interval becomes $[\hat{\mu}_{\hat{k}} \pm r_{\mathrm{sim},B} s_{\hat{k}}/\sqrt{m}]$.

The next lemma bounds the probability that a masked maximum of shifted studentized errors exceeds the order statistic $\psi_B(y)$.

Lemma 12 (Finite-draw exceedance as a Beta average)**.** *Assume $B \geq \lceil 1/\alpha \rceil - 1$ and put $n = B - k_\alpha(B)$. Fix a $\mathcal{G}_{\mathrm{tr}}$-measurable $\mu \in \mathbb{R}^K$ and write $y = c(\mu)$ for the discount vector of (7), which training and the validation sample determine. Here $\psi_B(y)$ and $U_b(y)$ stand for $\psi_B(y_P)$ and $U_b(y_P)$. Let $\hat{d}$ be the studentized error vector of Appendix B.1 and, for a $\mathcal{G}_{\mathrm{tr}}$-measurable shift $\gamma \geq 0$, define*

$$U_\gamma = \max_{k \in [K]} \big(|\hat{d}_k| + \gamma\big) \mathbf{1}\{|\hat{d}_k| + \gamma > y_k\}.$$

Let $P_B \sim \mathrm{Beta}(n,\, B - n + 1)$ be independent of all data and of the multipliers. With $\hat{r}_p(\mu)$ the radius of (7) at level p, put

$$D_\gamma(p) = \mathbb{P}\Big(\exists k \in [K] :\ |\hat{d}_k| \geq \max\{\hat{r}_p(\mu), y_k\} - \gamma \;\Big|\; \mathcal{G}_{\mathrm{tr}}\Big) - p.$$

Then, with $\mathbb{E}$ taken over P_B alone,

$$\mathbb{P}\big(U_\gamma > \psi_B(y) \mid \mathcal{G}_{\mathrm{tr}}\big) \;\leq\; \frac{n}{B+1} \;+\; \mathbb{E}\big[D_\gamma(P_B)_+\big].$$

Proof. We bound the exceedance by a binomial tail, write that tail as a Beta probability, bound the resulting event at one fixed level and integrate over levels.

Condition on $\mathcal{G}_{\mathrm{tr}}$ and on the validation sample, so that y, U_γ and γ are fixed numbers. Because the multipliers are independent of the data, the B draws are then i.i.d. copies of T^*. Write $\bar{G}(u) = \mathbb{P}_e(U^* > u)$ for the survival function of the masked maximum $U^* = \max_{k \in P} |T_k^*| \mathbf{1}\{|T_k^*| > y_k\}$ under one draw, and $\chi_B(p) = \mathbb{P}(\mathrm{Bin}(B, p) \leq n - 1)$. The bounds $0 \leq k_\alpha(B) \leq B - 1$ shown before the lemma give $1 \leq n \leq B$.

Step 1. The exceedance probability is at most a binomial tail. Suppose $U_\gamma > \psi_B(y)$. Then $U_\gamma > 0$, because masked maxima are non-negative. At least $k_\alpha(B) + 1$ masked maxima are at or below $\psi_B(y)$, hence strictly below U_γ, so at most $n-1$ of the B masked maxima strictly exceed U_γ. Given the sample, the number of masked maxima strictly exceeding U_γ is $\mathrm{Bin}(B, \bar{G}(U_\gamma))$. Consequently

$$\mathbb{P}\big(U_\gamma > \psi_B(y) \mid \text{sample}\big) \leq \mathbf{1}\{U_\gamma > 0\}\, \chi_B(\bar{G}(U_\gamma)),$$
$$\mathbb{P}\big(U_\gamma > \psi_B(y) \mid \mathcal{G}_{\mathrm{tr}}\big) \leq \mathbb{E}\big[\mathbf{1}\{U_\gamma > 0\}\, \chi_B(\bar{G}(U_\gamma)) \mid \mathcal{G}_{\mathrm{tr}}\big],$$

where the second line averages the first over validation samples.

Step 2. The binomial tail is a Beta probability. Differentiating the sum defining χ_B term by term gives a telescoping sum, which equals

$$-\chi_B'(p) = B \binom{B-1}{n-1} p^{n-1} (1-p)^{B-n}.$$

The coefficient $B!/((n-1)!(B-n)!)$ is the reciprocal of $\int_0^1 p^{n-1}(1-p)^{B-n}\, dp$, so this is the density of P_B. Since $n \leq B$, also $\chi_B(1) = 0$, and integrating the derivative from x to 1 gives

$$\chi_B(x) = \mathbb{P}(P_B > x), \qquad x \in [0, 1].$$

Since P_B is independent of the sample and of the multipliers,

$$\mathbb{E}\big[\mathbf{1}\{U_\gamma > 0\}\, \chi_B(\bar{G}(U_\gamma)) \mid \mathcal{G}_{\mathrm{tr}}\big] = \mathbb{P}\big(U_\gamma > 0,\ P_B > \bar{G}(U_\gamma) \mid \mathcal{G}_{\mathrm{tr}}\big).$$

The mean $\mathbb{E}[P_B] = n/(B+1)$ is the ratio that (9) bounds by α.

Step 3. For each fixed $p \in (0, 1)$, the event $\{U_\gamma > 0,\ \bar{G}(U_\gamma) < p\}$ has probability at most $p + D_\gamma(p)$ given $\mathcal{G}_{\mathrm{tr}}$. Fix $p \in (0, 1)$ and put $v_p = \inf\{u : \bar{G}(u) \leq p\}$. By the definition of the

infimum, $\bar{G}(u) > p$ at every $u < v_p$, so $\{\bar{G}(U_\gamma) < p\}$ lies inside $\{U_\gamma \geq v_p\}$. For $r \geq 0$, the masked maximum U^* exceeds r exactly when some $k \in P$ has $|T_k^*| > \max\{r, y_k\}$, so $\bar{G}(r) = \hat{S}(F(r, \mu))$ for $r \geq 0$. Since $U^* \geq 0$, also $\bar{G}(u) = 1 > p$ for $u < 0$, so v_p equals the radius $\hat{r}_p(\mu)$ of (7) at level p.

Suppose further that $U_\gamma > 0$ and $\bar{G}(U_\gamma) < p$, and let k attain the maximum defining U_γ. Its indicator equals one, so $|\hat{d}_k| + \gamma > y_k$, and $|\hat{d}_k| + \gamma = U_\gamma \geq v_p = \hat{r}_p(\mu)$. Hence

$$|\hat{d}_k| \geq \max\{\hat{r}_p(\mu), y_k\} - \gamma,$$

which is the event defining $D_\gamma(p)$. The weak inequality there covers the case $U_\gamma = v_p$. That event has conditional probability $p + D_\gamma(p)$ given $\mathcal{G}_{\mathrm{tr}}$, so

$$\mathbb{P}\big(U_\gamma > 0,\ \bar{G}(U_\gamma) < p \mid \mathcal{G}_{\mathrm{tr}}\big) \leq p + D_\gamma(p), \qquad 0 < p < 1.$$

Step 4. Steps 1 to 3 together prove the lemma. The level P_B lies in $(0, 1)$ almost surely and is independent of the sample. Below, the first inequality combines Steps 1 and 2 and the second integrates Step 3 against the law of P_B. The equality uses $\mathbb{E}[P_B] = n/(B+1)$ from Step 2, and the last inequality uses $D_\gamma \leq (D_\gamma)_+$:

$$\begin{aligned}
\mathbb{P}\big(U_\gamma > \psi_B(y) \mid \mathcal{G}_{\mathrm{tr}}\big) &\leq \mathbb{P}\big(U_\gamma > 0,\ P_B > \bar{G}(U_\gamma) \mid \mathcal{G}_{\mathrm{tr}}\big) \\
&\leq \mathbb{E}\big[P_B + D_\gamma(P_B)\big] \ = \ \mathbb{E}\big[D_\gamma(P_B)\big] + \frac{n}{B+1} \\
&\leq \frac{n}{B+1} + \mathbb{E}\big[D_\gamma(P_B)_+\big].
\end{aligned}$$

□

B.7 Coverage at a positive floor

With the notation and conditional validation model of Appendix B.1, the coverage argument works at any positive floor chosen from training information. The grid comes from a grid rule and has finitely many points, possibly a random number of them. Assume that on every realization it satisfies the span and spacing conditions of Appendix B.5 with spacing bound $h = h(m, K)$, where $h(m, k)$ is the deterministic spacing bound of Section 3. Write $\gamma = \sqrt{m}\, h(m, K)/a$ and $j = C_K \gamma$.

Theorem 13 (Coverage at a chosen floor)**.** *Fix $\alpha \in (0, 1)$, training-measurable $\delta, \epsilon \in (0, 1)$, and a positive training-measurable floor $a \leq \min_k \sigma_k$. With β, b, ℓ and g of (10)–(12) and γ, j as above, suppose*

$$\beta \leq \tfrac{1}{8}, \qquad \gamma < \sqrt{1 - 2\beta}\, z_{1-\alpha/2}.$$

We call the second inequality the radius condition*. Then the interval I_m of (4), computed with the exact-tail radius ψ at floor a, satisfies*

$$\mathbb{P}\Big(\{\hat{k} \in Z\} \cup \{\hat{k} \notin Z,\ \mu_{\hat{k}} \in I_m\} \mid \mathcal{G}_{\mathrm{tr}}\Big) \geq 1 - \alpha - \{b + \ell + g + j + \delta\}. \tag{20}$$

The same bound holds for the hull computed from one shared array of $B \geq \lceil 1/\alpha \rceil - 1$ bootstrap draws with the rank $k_\alpha(B)$ of (9). For this hull, the span and spacing conditions apply to the search interval with $\psi_B(0)$ in place of r_{sim}.

Proof. Steps 1 to 9 build a rectangle, fixed by training, such that, on a concentration event, $\mu_{\hat{k}} \in I_m$ whenever d lies in it. They then use Gaussian comparisons to bound the probability that d leaves it. Steps 10 to 12 extend the bound to B draws by averaging over Beta-distributed error levels. Throughout, $i^* = i^*(\mu)$ is the smallest-index population winner, and the proof uses the tail $\hat{S}^{[K]}$ and the set $\hat{C}_\alpha^{[K]} = \hat{C}_\alpha$ of Appendix B.1.

Step 1. Concentration and Gaussian comparison supply the error terms δ, b and ℓ. The event Ω of Appendix B.2 has conditional probability at least $1 - \delta$, and on Ω the set Z of constant columns is empty and

$$\max_k |s_k/\sigma_k - 1| \leq \beta, \qquad 1 - 2\beta \leq \hat{R}_{kk} \leq 1, \qquad \|\hat{R} - R\|_{\max} \leq 4\rho/a^2.$$

The Gaussian comparison (16) contributes b, and the central limit theorem (17) contributes ℓ.

Step 2. On Ω, the radius $\hat{r}_\alpha(\mu)$ exceeds γ, which gives the hypothesis $\gamma \le \hat{r}_\alpha(\mu)$ of Lemma 11. The winner i^* has $\Delta_{i^*}(\mu) = 0$, so $F(r,\mu)_{i^*} = r \le F(r,\mu)_k$ for every k and $r \ge 0$. On Ω the coordinate i^* lies in P, and $T^*_{i^*} \sim N(0, \hat{R}_{i^*i^*})$ with $\hat{R}_{i^*i^*} \ge 1 - 2\beta$. For $0 \le r < \sqrt{1-2\beta}\, z_{1-\alpha/2}$ we have $r/\hat{R}^{1/2}_{i^*i^*} < z_{1-\alpha/2}$, and the marginal tail at i^* gives

$$\hat{S}(F(r,\mu)) \ge \mathbb{P}_e(|T^*_{i^*}| > r) = 2\bar{\Phi}\big(r/\hat{R}^{1/2}_{i^*i^*}\big) > \alpha.$$

No such r lies in the set whose infimum defines $\hat{r}_\alpha(\mu)$ in (7). With the hypothesis on γ, this gives

$$\min_k F(\hat{r}_\alpha(\mu),\mu)_k = \hat{r}_\alpha(\mu) \ge \sqrt{1-2\beta}\, z_{1-\alpha/2} > \gamma. \tag{21}$$

Step 3. On Ω, the thresholds $F(r) = F(r,\mu)$ lie within factors $1 \pm 2\beta$ of the population thresholds $\bar{F}(r)_k = \max\{r, \bar{c}_k\}$, where $\bar{c}_k = \sqrt{m}(\mu_k - \mu_{i^*})/(\sigma_k + \sigma_{i^*})$. Steps 3 to 7 use this to build a training-measurable half-width vector u with $S^{\hat{Z}}(u) \le \alpha + b + g$ such that, on Ω, the event $\{d \in \mathcal{B}_u\}$ implies $\mu \in \hat{C}^{[K]}_\alpha$. The discount vector $c(\mu)$ of $F(r)$ has denominators $s_k + s_{i^*}$. On Ω, Step 1 puts each s_k/σ_k in $[1-\beta, 1+\beta]$, so

$$\frac{\sigma_k + \sigma_{i^*}}{s_k + s_{i^*}} \in \big[(1+\beta)^{-1}, (1-\beta)^{-1}\big] \subseteq [1-2\beta, 1+2\beta],$$

where the inclusion uses $\beta \le 1/8$.

Each $\bar{c}_k$ is nonnegative because i^* minimizes μ, so $c(\mu)_k$ lies between $(1-2\beta)\bar{c}_k$ and $(1+2\beta)\bar{c}_k$, and the maximum with $r \ge 0$ preserves both bounds in each coordinate. Hence, on Ω,

$$(1-2\beta)\bar{F}(r) \le F(r) \le (1+2\beta)\bar{F}(r), \qquad r \ge 0. \tag{22}$$

Step 4. The radius

$$r_* = \inf\{r \ge 0 : S^{\hat{Z}}(\bar{F}(r)) \le \alpha + b + H\}$$

is training-measurable. The level carries H to pay for the application of (18) in Step 5. Step 7 applies it once more, and since each application raises the Gaussian tail probability by at most H, the two cost $2H = g$. The set defining r_* is nonempty, since $S^{\hat{Z}}(\bar{F}(r)) \le S^{\hat{Z}}(r\mathbf{1}) \to 0$ as $r \to \infty$ and $\alpha + b + H \ge \alpha > 0$.

When $\alpha+b+H \ge 1$, attainment is automatic because every tail probability is at most one. Otherwise, Lemma 5 with $c = \bar{c} \in [0,\infty)^K$ and $\lambda = \alpha + b + H$ gives $S^{\hat{Z}}(\bar{F}(t)) \le \alpha + b + H$ for every $t \ge r_*$. Conversely, $S^{\hat{Z}}(\bar{F}(t)) \le \alpha + b + H$ implies $r_* \le t$ by the definition of r_*, so

$$\{r_* \le t\} = \{S^{\hat{Z}}(\bar{F}(t)) \le \alpha + b + H\}, \qquad t \ge 0.$$

The level $\alpha + b + H$ is training-measurable, and the proof of Lemma 4, applied to $\hat{Z}$ at the training-measurable vector $\bar{F}(t)$, shows that $S^{\hat{Z}}(\bar{F}(t))$ is training-measurable. Each event $\{r_* \le t\}$ therefore lies in $\mathcal{G}_{\mathrm{tr}}$, which makes r_* training-measurable.

Step 5. On Ω, the radius $\hat{r} = \hat{r}_\alpha(\mu)$ satisfies $\hat{r} \ge r_*$, by the chain

$$S^{\hat{Z}}(\bar{F}(\hat{r})) \overset{\text{(i)}}{\le} S^{\hat{Z}}((1+2\beta)\bar{F}(\hat{r})) + H \overset{\text{(ii)}}{\le} S^{\hat{Z}}(F(\hat{r})) + H \overset{\text{(iii)}}{\le} \hat{S}^{[K]}(F(\hat{r})) + b + H \overset{\text{(iv)}}{\le} \alpha + b + H.$$

(i) Bound (18) at $(1+2\beta)\bar{F}(\hat{r})$ with $\varrho = 1/(1+2\beta) \ge 1 - 4\beta$.

(ii) The bound $F(\hat{r}) \le (1+2\beta)\bar{F}(\hat{r})$ from (22) and monotonicity of the tail.

(iii) The comparison (16) at the realized thresholds.

(iv) Attainment of $\hat{r}$, Lemma 5.

Thus $\hat{r}$ lies in the set defining r_*, so $\hat{r} \ge r_*$.

Step 6. Put $u = (1-2\beta)^2\bar{F}(r_*)$, a vector fixed by training. On Ω, the event $\{d \in \mathcal{B}_u\}$ implies $\mu \in \hat{C}^{[K]}_\alpha$. One factor $1-2\beta$ pays for the denominators of $c(\mu)$ and the other for studentization. On

Ω, $\sigma_k/s_k \le 1/(1-\beta) \le 1+2\beta$, where the last step uses $\beta \le \frac{1}{2}$. Since $\hat{d}_k = (\sigma_k/s_k)d_k$, this gives $|\hat{d}_k| \le (1+2\beta)|d_k|$ on Ω. The sandwich (22), monotonicity of $\bar{F}$ in r and $\hat{r} \ge r_*$ give, on Ω,

$$(1+2\beta)u \le (1-2\beta)\bar{F}(r_*) \le (1-2\beta)\bar{F}(\hat{r}) \le F(\hat{r}),$$

where the first inequality uses $(1+2\beta)(1-2\beta) \le 1$. Hence $|d_k| \le u_k$ forces $|\hat{d}_k| \le (1+2\beta)u_k \le F(\hat{r})_k$ on Ω, and

$$\{d \in \mathcal{B}_u\} \cap \Omega \subseteq \{\hat{d} \in \mathcal{B}_{F(\hat{r})}\},$$

which is the event $\{\mu \in \hat{C}_\alpha^{[K]}\}$ in (15).

Step 7. The Gaussian tail satisfies $S^{\hat{Z}}(u) \le \alpha + b + g$, with no restriction to Ω, because every quantity in the bound is fixed by training. Since $1 - 4\beta \le (1-2\beta)^2 \le 1$, bound (18) applies at $\bar{F}(r_*)$ with $\varrho = (1-2\beta)^2$. Attainment at r_* bounds $S^{\hat{Z}}(\bar{F}(r_*))$ by $\alpha + b + H$. Together these give

$$S^{\hat{Z}}(u) \le S^{\hat{Z}}(\bar{F}(r_*)) + H \le \alpha + b + 2H = \alpha + b + g.$$

Step 8. Shrinking the rectangle by γ leaves room for the grid. With $u^{\mathrm{g}} = (u - \gamma\mathbf{1})_+$, the inclusion $\mathcal{B}_u \subseteq \mathcal{B}_{u^{\mathrm{g}}+\gamma\mathbf{1}}$ and Nazarov's inequality (Chernozhukov et al., 2017b, Theorem 1) for the $2K$ unit-variance coordinates $\pm\hat{Z}_k$ give

$$S^{\hat{Z}}(u^{\mathrm{g}}) \le S^{\hat{Z}}(u) + C_K\gamma \le \alpha + b + g + j.$$

On $\Omega \cap \{d \in \mathcal{B}_{u^{\mathrm{g}}}\}$, every $|\hat{d}_k|$ lies at least γ below $F(\hat{r})_k$. If $u_k \ge \gamma$, then $|\hat{d}_k| \le (1+2\beta)(u_k - \gamma) \le F(\hat{r})_k - \gamma$. If $u_k < \gamma$, the shrunk rectangle forces $d_k = 0$, so $|\hat{d}_k| = 0 < F(\hat{r})_k - \gamma$ by (21). Hence, in either case,

$$\Omega \cap \{d \in \mathcal{B}_{u^{\mathrm{g}}}\} \subseteq \{\hat{d} \in \mathcal{B}_{F(\hat{r})-\gamma\mathbf{1}}\}. \tag{23}$$

On Ω no column is constant, so the procedure does not abstain, and Lemma 11, whose hypothesis $\gamma \le \hat{r}_\alpha(\mu)$ is (21), places $\mu_{\hat{k}}$ in I_m.

Step 9. A union bound proves (20) for the exact-tail interval. Because u^{g} is training-measurable, Lemma 4 lets us apply the central limit theorem (17) at u^{g}, so

$$\mathbb{P}(d \notin \mathcal{B}_{u^{\mathrm{g}}} \mid \mathcal{G}_{\mathrm{tr}}) \le S^{\hat{Z}}(u^{\mathrm{g}}) + \ell \le \alpha + b + g + j + \ell.$$

By (23) and Lemma 11, the event in (20) contains $\Omega \cap \{d \in \mathcal{B}_{u^{\mathrm{g}}}\}$. Since $\mathbb{P}(\Omega^c \mid \mathcal{G}_{\mathrm{tr}}) \le \delta$, a union bound, which needs no independence between Ω and d, proves (20).

Step 10. With B draws, a nonabstaining realization that misses $\mu_{\hat{k}}$ lies in an event that Lemma 12 controls. Appendix B.6 shows that Theorem 10 and Lemma 11 also hold for the empirical tail. In Lemma 11, with $y = c(\mu)$ and $\psi_B(y)$ standing for $\psi_B(y_P)$ as in Lemma 12, the threshold vector becomes $v = F(\psi_B(y), \mu)$, with coordinates $\max\{\psi_B(y), y_k\}$ and minimum $\psi_B(y)$ since $y_{i^*} = 0$. On a nonabstaining realization, Lemma 11 therefore places $\mu_{\hat{k}}$ in I_m whenever $\psi_B(y) \ge \gamma$ and $\hat{d} \in \mathcal{B}_{v-\gamma\mathbf{1}}$. If $\psi_B(y) < \gamma$, then $v_{i^*} - \gamma < 0 \le |\hat{d}_{i^*}|$, so $\hat{d} \notin \mathcal{B}_{v-\gamma\mathbf{1}}$. Hence, on such a realization, $\mu_{\hat{k}} \notin I_m$ implies $\hat{d} \notin \mathcal{B}_{v-\gamma\mathbf{1}}$, which is the event

$$\{\exists k : |\hat{d}_k| > \max\{\psi_B(y), y_k\} - \gamma\} = \{U_\gamma > \psi_B(y)\}.$$

Lemma 12, applied with this γ, bounds the probability of this event by $(B - k_\alpha(B))/(B+1) + \mathbb{E}[D_\gamma(P_B)_+]$, with P_B and $D_\gamma(p)$ as in that lemma. The term $D_\gamma(p)$ is the probability that some $|\hat{d}_k|$ reaches its threshold $\max\{\hat{r}_p(\mu), y_k\}$ lowered by γ, minus p. With $\eta_0 = b + \ell + g + \delta$ and $p_0 = 2\bar{\Phi}(\gamma/\sqrt{1-2\beta})$, Steps 11 and 12 show that $D_\gamma(p)_+ \le \eta_0 + j$ for $p < p_0$ and for $p \ge p_0$ respectively, hence for every level p.

Step 11. The bound $D_\gamma(p)_+ \le \eta_0 + j$ holds for $p < p_0$. For such p we have $\gamma < \sqrt{1-2\beta}\, z_{1-p/2}$, so we may fix ς with $0 < \varsigma < \sqrt{1-2\beta}\, z_{1-p/2} - \gamma$, which is the radius condition at level p with γ replaced by $\gamma + \varsigma$. Repeat the argument from (21) to (23) with p in place of α and $\gamma + \varsigma$ in place of γ. Its version of (23) places $\hat{d}$ in $\mathcal{B}_{F(\hat{r}_p(\mu),\mu)-(\gamma+\varsigma)\mathbf{1}}$ outside the event N_p on which Ω fails or d leaves that version's shrunk rectangle. The union bound of Step 9 gives

$$\mathbb{P}(N_p \mid \mathcal{G}_{\mathrm{tr}}) \le p + \eta_0 + C_K(\gamma + \varsigma) = p + \eta_0 + j + C_K\varsigma.$$

Off N_p, every k has $|\hat{d}_k| \le \max\{\hat{r}_p(\mu), y_k\} - \gamma - \varsigma$, which is strictly below the threshold in $D_\gamma(p)$. The event defining $D_\gamma(p)$ therefore lies in N_p, so $D_\gamma(p) \le \eta_0 + j + C_K\varsigma$. Letting $\varsigma \downarrow 0$ gives $D_\gamma(p) \le \eta_0 + j$, and hence $D_\gamma(p)_+ \le \eta_0 + j$.

Step 12. The bound also holds for $p \ge p_0$, which proves the theorem for B draws. For $p \ge p_0$, the bounds $1 - 2\bar{\Phi}(x) \le x\sqrt{2/\pi}$, $\sqrt{1-2\beta} \ge \sqrt{3/4}$ and $C_K > 1$ give

$$D_\gamma(p)_+ \le 1 - p \le 1 - p_0 \le \sqrt{\frac{2}{\pi}} \frac{\gamma}{\sqrt{1-2\beta}} \le 0.93\gamma \le j.$$

The bound $\eta_0 + j$ does not depend on p, so $\mathbb{E}[D_\gamma(P_B)_+] \le \eta_0 + j$. The rank bound in (9) gives $(B - k_\alpha(B))/(B+1) \le \alpha$. Adding the two bounds proves (20) for B draws without a union bound over candidate grid values. □

The bound (20) also holds, trivially, where $\beta > 1/8$ or the radius condition fails. The five error terms b, ℓ, g, j, δ are non-negative and training-measurable, and the bound is trivial wherever their sum $\eta = b + \ell + g + j + \delta$ is at least $1 - \alpha$. Where $\beta > 1/8$, the term g exceeds $2C_K\sqrt{2\log 4} > 1$. Where $\beta \le 1/8$ and $\gamma \ge \sqrt{1-2\beta}\, z_{1-\alpha/2}$, the term j is at least $C_K\sqrt{3/4}\, z_{1-\alpha/2}$. This exceeds $1 - \alpha$ because convexity of $q \mapsto z_q$ on $[1/2, 1)$ gives $z_{1-\alpha/2} \ge (1-\alpha)\sqrt{\pi/2}$. Hence (20) holds almost surely when the floor and grid hypotheses hold almost surely.

If η is also integrable, Lemma 3 averages (20) over training, for I_m computed with the exact tail or from the B bootstrap draws of Theorem 13:

$$\mathbb{P}\Big(\{\hat{k} \in Z\} \cup \{\hat{k} \notin Z,\ \mu_{\hat{k}} \in I_m\}\Big) \ge 1 - \alpha - \mathbb{E}[\eta].$$

B.8 Proof of Theorem 1

Proof of Theorem 1. Steps 1 to 4 apply Theorem 13 at the floor $a = cM$ with $c = \xi/2$ and $\delta = \epsilon = 1/m$. Steps 5 to 7 transfer the bound to zero floor by a coupling, and Step 8 averages over training. The floor a is deterministic and satisfies $a \le \xi M \le \min_k \sigma_k$, as Theorem 13 requires.

Step 1. The bound β is at most $1/8$ for all sufficiently large m, uniformly over $K \le K_{\max}(m)$. For $K \ge 2$ and $m \ge 3$, the definition $\Lambda = \log(K \vee m)$ gives

$$\log(2K) \le 2\Lambda, \qquad \log(2Km) \le 3\Lambda, \qquad \log(4K^2m) \le 5\Lambda, \qquad C_K \le 4\sqrt{\Lambda}.$$

Since $1 < \log 3 \le \Lambda \le \Lambda_{\max}$, the ratio Λ/m is at most $\Lambda_{\max}^5/m = q_{\max}$, which tends to zero. With $\delta = 1/m$, the bound on $\log(4K^2m)$ turns (10) into $\rho = O(M^2\sqrt{\Lambda/m})$. Hence $\beta = \rho/(2a^2) = O(c^{-2}\sqrt{\Lambda/m})$ is at most $1/8$ for all sufficiently large m. The threshold depends only on ξ and $K_{\max}$, because β depends on training only through $K \le K_{\max}(m)$.

Step 2. We bound the error terms b, ℓ and g. Substituting ρ, β and $\epsilon = 1/m$ into (11) and (12) gives

$$b = O(c^{-1}\Lambda^{5/4}m^{-1/4}), \quad \ell = O(c^{-1/2}\Lambda^{5/4}m^{-1/4}), \quad g = O(c^{-2}\Lambda^{3/2}m^{-1/2}) + 1/m.$$

The bound on b uses $\log(2K) \le 2\Lambda$, and the bound on g uses $C_K \le 4\sqrt{\Lambda}$ and $\log(2Km) \le 3\Lambda$. The factor M cancels from each bound, so the constants depend only on $c = \xi/2$. The first term of g is of smaller order than the bound on b, uniformly, because the ratio of the two rates is $(\Lambda/m)^{1/4} \to 0$. The terms $\epsilon = 1/m$ in g and $\delta = 1/m$ are at most $m^{-1/4} \le q_{\max}^{1/4}$.

Step 3. The grid and radius conditions hold for all sufficiently large m. The grid condition (5) gives $j = o(q_{\max}^{1/4})$ uniformly. Since $C_K \ge 2$ and $q_{\max} \to 0$, also $\gamma = j/C_K \to 0$ uniformly. Once $\beta \le 1/8$, we have $\sqrt{1-2\beta} \ge \sqrt{3/4}$, so for all sufficiently large m

$$\gamma < \sqrt{3/4}\, z_{1-\alpha/2} \le \sqrt{1-2\beta}\, z_{1-\alpha/2},$$

which is the radius condition of Theorem 13. This threshold depends also on α and h.

Step 4. The theorem holds at the floor $\sigma_{\min} = a$. Since $\Lambda \le \Lambda_{\max}$ and $q_{\max}^{1/4} = \Lambda_{\max}^{5/4}m^{-1/4}$, Steps 2 and 3 bound the error sum $b + \ell + g + j + \delta$ by $O_\xi(\Lambda_{\max}^{5/4}m^{-1/4})$ for all sufficiently large m.

Theorem 13 bounds the coverage deficit by this sum for the exact-tail interval and for its computed version with B draws.

Step 5. An auxiliary procedure at floor a satisfies the bound of Step 4. The auxiliary procedure uses the same data and bootstrap draws as the zero-floor procedure. On the event $\min_k \hat{\sigma}_k \geq a$, both use the scales $s_k = \hat{\sigma}_k$, so their tails and search intervals agree. There the auxiliary procedure takes the zero-floor grid, which spans the common search interval. Otherwise it takes an equally spaced grid spanning its search interval (3), with spacing at most $h(m, K)$. This grid rule is measurable because the zero-floor grid rule is. On every realization it meets the span and spacing conditions of Appendix B.5 with spacing bound $h(m, K)$, so Step 4 applies.

Step 6. The two procedures return the same interval on a concentration event. Let Ω be the concentration event of Appendix B.2 with $\delta = 1/m$, which does not depend on the floor. On Ω, the bound $\|\hat{V} - \Sigma\|_{\max} \leq \rho$ gives, for every k,

$$\hat{\sigma}_k^2 \geq \sigma_k^2 - \rho \geq 4a^2 - 2\beta a^2 \geq a^2,$$

because $\sigma_k \geq \xi M = 2a$, $\rho = 2\beta a^2$ and $\beta \leq 1/8$. Hence Ω lies inside the event $\min_k \hat{\sigma}_k \geq a$. On Ω, both procedures have $Z = \emptyset$, so neither abstains. They also share $\hat{k}$, the scales, the bootstrap draws, the tie rule and the grid, so they return the same interval I_m and have the same coverage event.

Step 7. The theorem holds at the floor $\sigma_{\min} = 0$. Since $\mathbb{P}(\Omega^c \mid \mathcal{G}_{\mathrm{tr}}) \leq 1/m$, the zero-floor coverage probability is at least the auxiliary bound minus $1/m$. The extra $1/m \leq q_{\max}^{1/4}$ leaves the rate unchanged.

Step 8. The rate survives averaging over training. Each conditional bound holds with a deterministic error η once m exceeds a deterministic threshold. Lemma 3 with this η gives the same rate after averaging over training. □

Fixed grids. An equally spaced grid with endpoints at the ends of a search interval of width w meets the spacing bound $h(m, K)$ exactly when it has at least $1 + \lceil w/h(m, K) \rceil$ points. With the radius ψ_B, no fixed count suffices on every realization, because $\psi_B(0)$ is unbounded over the bootstrap draws. With the exact tail, take m large enough that $\beta \leq 1/8$ (Step 1). On Ω, the variance $\hat{R}_{\hat{k}\hat{k}}$ of $T_{\hat{k}}^*$ is at least $1 - 2\beta$, and every scale is at least a at either floor (Step 6). The radius r_{sim} is at least the $(1-\alpha)$ quantile of $|T_{\hat{k}}^*|$, hence at least $\sqrt{1-2\beta}\, z_{1-\alpha/2}$. The width $w = 2r_{\mathrm{sim}} s_{\hat{k}}/\sqrt{m}$ of (3) and $h(m, K) = j\, a/(C_K\sqrt{m})$ then give

$$\frac{w}{h(m, K)} \geq \frac{2\sqrt{1-2\beta}\, z_{1-\alpha/2}\, C_K}{j} \to \infty,$$

since $j \to 0$ by (5), so the number of grid points must grow with m.

Remark 14 (A shrinking floor). Theorem 13 also allows population standard deviations that shrink with m. Let $q = q(m, K) = \Lambda^5/m$ and $a = Mq^{1/8}$. In Theorem 1, replace $\min_k \sigma_k \geq \xi M$ by $a \leq \min_k \sigma_k$ almost surely, and use the floor $\sigma_{\min} = a$. In (5), replace a by $Mq(m, k)^{1/8}$ and $q_{\max}^{1/4}$ by $q_{\max}^{1/8}$.

Apply Theorem 13 with $\delta = q^{1/8}$ and $\epsilon = (\Lambda/m)^{1/4}$. As $\log(4K^2/\delta) = O(\Lambda)$, (10) gives $\rho = O(M^2\sqrt{\Lambda/m})$ and $\beta = O(\Lambda^{-3/4}m^{-1/4})$. Hence $\beta \leq 1/8$ and the radius condition hold for all sufficiently large m, as in Steps 1 and 3. The terms of the error sum of Step 4 are $b = O(q^{1/8})$, $\ell = C_\diamond q^{3/16}$, $g = O((\Lambda/m)^{1/4})$, $\delta = q^{1/8}$ and $j = o(q_{\max}^{1/8})$. For all sufficiently large m, the coverage deficit is therefore $O(q_{\max}^{1/8}) = O(\Lambda_{\max}^{5/8}m^{-1/8})$ with an absolute constant, uniformly over $K \leq K_{\max}(m)$. The same holds with B draws when the grid condition is imposed on the search interval built from ψ_B, and averaging over training preserves the rate as in Step 8.

C Additional experimental results

These results support Sections 4 and 5.

C.1 WIDTH DECOMPOSITION

With w_{Bonf}, w_{band}, w_{grid} and $w_{\text{B-Zoom}}$ the widths of Bonferroni, the band, the grid band and B-Zoom, B-Zoom's saving over Bonferroni decomposes as

$$w_{\mathrm{Bonf}} - w_{\text{B-Zoom}} = (w_{\mathrm{Bonf}} - w_{\mathrm{band}}) + (w_{\mathrm{band}} - w_{\mathrm{grid}}) + (w_{\mathrm{grid}} - w_{\text{B-Zoom}}).$$

With the draws and the grid held fixed, the three terms measure dependence, grid rounding and localization. As in the studentized multiple tests of Romano & Wolf (2005), the band's bootstrap threshold captures the joint dependence of the statistics, which the first term measures. Table 4 reports localization as $100(1 - \bar{w}_{\text{B-Zoom}}/\bar{w}_{\mathrm{grid}})$, a ratio of mean widths. In all $84{,}000$ cases, twice the grid spacing is at most 0.9% of B-Zoom's width and equals 0.51% of the band's width.

Table 4: All 84 cells of the seven arms at nominal 0.90, each from 100 training seeds with ten validation samples per seed. Each entry is B-Zoom coverage / grid-band coverage / localization saving in percent of the grid-band width. B-Zoom coverage is at least 0.899 in every cell, and the saving grows with m for every arm and every K. The grid band shares B-Zoom's draws and grid, so the saving comes from the gap discounts alone.

Arm	m	$K = 30$	$K = 150$	$K = 600$	$K = 3000$
CIFAR-10 main	100	0.969/0.974/4.0	0.972/0.973/3.0	0.971/0.972/2.7	0.977/0.977/2.9
	1000	0.964/0.991/17.8	0.986/0.997/16.7	0.994/1.000/16.3	0.995/1.000/17.0
	5000	0.949/0.991/26.6	0.971/0.999/27.3	0.984/1.000/28.4	0.987/1.000/29.9
Interior optimum	100	0.946/0.948/1.2	0.956/0.956/0.7	0.962/0.962/0.6	0.957/0.957/0.6
	1000	0.982/0.989/8.0	0.992/0.993/7.9	0.996/0.997/7.7	0.998/0.998/6.8
	5000	0.952/0.990/21.2	0.973/0.998/23.0	0.985/1.000/23.3	0.990/1.000/20.7
Label noise, CE	100	0.979/0.984/5.8	0.984/0.987/4.7	0.984/0.986/4.5	0.985/0.987/4.8
	1000	0.962/0.992/20.3	0.979/0.998/19.8	0.989/0.999/20.4	0.992/0.999/21.4
	5000	0.934/0.992/27.8	0.957/0.999/29.5	0.968/1.000/31.6	0.971/1.000/33.1
Label noise, error	100	0.965/0.970/2.5	0.968/0.968/1.9	0.965/0.968/1.7	0.966/0.966/1.6
	1000	0.987/0.995/12.2	0.991/0.993/10.8	0.993/0.995/10.2	0.992/0.998/10.2
	5000	0.971/0.995/20.7	0.982/0.998/19.6	0.992/1.000/19.4	0.997/1.000/20.0
CIFAR-100, CE	100	0.974/0.982/9.7	0.986/0.990/8.8	0.989/0.992/8.7	0.990/0.993/10.0
	1000	0.951/0.990/22.7	0.974/0.994/22.8	0.984/0.996/23.7	0.987/0.996/25.7
	5000	0.939/0.992/28.8	0.954/0.996/30.7	0.966/0.999/33.0	0.971/1.000/35.3
CIFAR-100, error	100	0.986/0.987/5.7	0.993/0.993/4.9	0.991/0.991/4.5	0.995/0.995/4.4
	1000	0.970/0.987/17.3	0.984/0.995/16.1	0.988/0.997/15.8	0.992/0.997/16.3
	5000	0.968/0.993/24.2	0.986/0.998/23.7	0.990/1.000/24.0	0.993/1.000/25.2
Fine-tuning	100	0.916/0.922/3.3	0.904/0.905/2.1	0.904/0.908/2.3	0.899/0.903/2.5
	1000	0.971/0.978/5.8	0.983/0.986/4.7	0.983/0.985/4.7	0.982/0.987/4.8
	5000	0.980/0.989/11.1	0.991/0.997/10.0	0.996/0.998/10.0	0.997/0.998/9.9

Storing cross-entropy losses in half precision shifts their pool mean by at most a recorded bound. Under a stricter score that asks the interval to contain every value within this bound of the target, no B-Zoom cell has an upper coverage limit below nominal. These limits are computed across training seeds, as in Section 4.

Coverage-and-return is the fraction of cases in which a method returns an interval containing its target. Table 5 compares all seven arms, including zoom-baseline declines and split cost. On the CIFAR-100 error arm, 2822 of $12{,}000$ cases contain a constant column, and at $m = 100$ their share rises from 35.2% at $K = 30$ to 100.0% at $K = 3000$. B-Zoom uses only the nonconstant columns and returns an interval in every case, because the selected column is never constant. At $m = 100$, a checkpoint with error rate 0.01 has an all-zero loss column with probability $0.99^{100} \approx 0.366$, and B-Zoom abstains at any floor when the selected column is constant. Capping cross-entropy at $M = 10$ also produces constant columns, in 271 cases of the CIFAR-100 cross-entropy arm.

The zoom baseline declines every case with a constant column. Of the 9000 cases at $K \leq 600$, it returns an interval in 7180 on the error arm and 8927 on the cross-entropy arm. The 1820 declines on the error arm lower the zoom baseline's coverage-and-return there to 0.784.

Table 5: Seven arms at nominal 0.90, with coverage over all 12,000 cases per arm, where B-Zoom exceeds nominal on every arm and is lowest on fine-tuning at 0.959. Brackets give 95% Student-t limits over seed averages, and band and Bonferroni coverage limits have half-width at most 0.0062. Split cost (Table 1) is in loss units and positive on all seven arms. Splitting is scored at its own selected checkpoint, and its coverage limits straddle nominal on six arms and are $[0.869, 0.890]$ on fine-tuning. The two width columns give each method's mean width relative to B-Zoom's in percent, negative when narrower. Zoom-baseline columns use the 9000 cases at $K \leq 600$ and count declines as misses for coverage-and-return. Zoom width uses the cases where both methods return, on which the two coverages differ by at most 0.003.

Arm	B-Zoom [95%]	Band	Bonferroni	Split
CIFAR-10 main	0.977 [0.972, 0.981]	0.990	0.995	0.892
Interior optimum	0.974 [0.970, 0.978]	0.982	0.993	0.896
Label noise, CE	0.974 [0.969, 0.979]	0.994	0.998	0.892
Label noise, error	0.981 [0.977, 0.985]	0.987	0.993	0.891
CIFAR-100, CE	0.972 [0.966, 0.978]	0.993	0.997	0.898
CIFAR-100, error	0.986 [0.983, 0.990]	0.994	0.997	0.898
Fine-tuning	0.959 [0.952, 0.965]	0.963	0.983	0.879

Arm	Split cost [95%]	Split width vs B-Zoom	Zoom returns / 9000	Zoom coverage and return	Zoom width vs B-Zoom
CIFAR-10 main	0.0080 [0.0072, 0.0088]	−11.7%	9,000	0.971	−0.40%
Interior optimum	0.0219 [0.0199, 0.0239]	−8.4%	9,000	0.970	−0.56%
Label noise, CE	0.0042 [0.0037, 0.0047]	−10.6%	9,000	0.970	−0.34%
Label noise, error	0.0063 [0.0059, 0.0067]	−20.3%	9,000	0.979	−0.03%
CIFAR-100, CE	0.0088 [0.0079, 0.0098]	−7.6%	8,927	0.959	−0.32%
CIFAR-100, error	0.0057 [0.0052, 0.0061]	−21.5%	7,180	0.784	−0.12%
Fine-tuning	0.0059 [0.0053, 0.0065]	+0.2%	9,000	0.958	−0.10%

C.2 Joint and selected coverage

Table 6: Joint and selected coverage of the band on 800 cases at $K = 600$ and $\alpha = 0.10$, with 20 training seeds and ten validation samples per seed in each cell. Joint coverage is the fraction of cases in which the band covers every checkpoint's population risk at once, with 95% Student-t limits from seed averages. Selected coverage is the fraction in which it covers the selected risk. Added slack is the fraction in which it covers the selected risk but not every risk. B-Zoom's coverage equals the band's selected coverage in each cell.

Arm, m	Joint coverage	Selected	Added slack
Main, 100	0.760 [0.697, 0.823]	0.980	0.220
Main, 1000	0.890 [0.847, 0.933]	0.990	0.100
Interior, 100	0.745 [0.682, 0.808]	0.920	0.175
Interior, 1000	0.885 [0.844, 0.926]	1.000	0.115

The band covers the selected checkpoint's population risk more often than it covers every checkpoint's population risk at once. Benjamini & Yekutieli (2005) show that intervals reported only for parameters selected after viewing the data lose their nominal coverage. The 20 seeds are the first 20, in numerical order, of 50 training seeds drawn at random on each arm (Appendix E.3), and every case uses $B = 2000$ and 400 grid points.

Let J be the event that the band covers every checkpoint's population risk and S the event that it covers the selected checkpoint's. Let R be the event that it returns an interval for the selected checkpoint. Shared scales and a shared radius give $J \subseteq S$, so

$$\mathbf{1}_{R \cap S} = \mathbf{1}_J + \mathbf{1}_{R \cap S \cap J^c} - \mathbf{1}_{R^c \cap J}.$$

The middle term is the added slack of Table 6. Because no case abstains, the added slack equals selected coverage minus joint coverage. The event J also requires coverage of μ_k for each constant column $k \in Z$, which holds only when $\hat{\mu}_k = \mu_k$, and none of the 800 cases has a constant column.

C.3 THE MAXIMUM OVER REFERENCES

Theorem 10 applies to the maximum over references and gives no guarantee for the single-reference ablation of Section 4.2.

The real comparison uses the main and interior-optimum arms at $K = 600$ and $m \in \{100, 1000\}$, with the first 20 sorted training seeds per arm and ten validation samples per seed. Both methods share $B = 2000$ Gaussian draws and 400 grid points. The maximum changes endpoints in 82.0 to 97.5% of cases, raises mean width by 0.39 to 0.78% and changes coverage by at most 0.005.

With the known Gaussian tail in place of $\hat{S}$, a search over 448 configurations of $K \in \{3, 4, 5, 6\}$ independent Gaussian coordinates reruns its four most adverse configurations with 200,000 replications each. In these four, the ablation covers at least 0.921 and the maximum adds at most 1.06 percentage points of coverage.

A VALUE THAT NEEDS A SECOND REFERENCE

The single-reference ablation can reject a selected-risk value that the confidence set $\hat{C}_\alpha$ of Appendix A.1 permits, with both using the known Gaussian tail in place of $\hat{S}$. Take three independent Gaussian coordinates with observed means $(0, 0.01, 22)$, known standard errors $(1, 10, 1)$ and $\alpha = 0.1$, so the first coordinate is selected. The set $\hat{C}_\alpha$ contains the point $\mu = (2, 1.999999, 19.885946)$, where $\hat{r}_\alpha(\mu) = 2.11405447$. At $t = 2$, the ablation's radius is 1.94882186, too small to accept t. Adding the plausible second reference raises the radius to 2.11405447, enough to accept t.

C.4 RARE LOSSES

The rare-loss design (6) of Section 4.3 has $m = 100$ and $M = 1$. Both columns have mean $\mu = 0.01(1 - \epsilon) + \epsilon/2 = 0.010049$. At $\alpha = 0.10$ and every floor $\sigma_{\min} \geq 0$, the exact-tail interval covers μ with probability at most $1 - 0.99^{100} \approx 0.634$.

Proof. The proof computes the accepted set $\hat{C}_\alpha^W$ and then finds an event on which this set misses μ.

Step 1. The procedure never abstains. The two sample means are equal, so validation selects the smaller index, $\hat{k} = 1$. The U_j are distinct almost surely, so both empirical variances are positive, Z is empty and $P = \{1, 2\}$.

Step 2. At every floor, $\hat{C}_\alpha^W = [\hat{\mu} \pm z_{0.95}\hat{\sigma}/\sqrt{100}]$. Write $\hat{\mu}$ and $\hat{\sigma}$ for the common column mean and standard deviation, and $s = \max(\hat{\sigma}, \sigma_{\min}) > 0$ for the common scale. Conditional on L, the multiplier coordinates satisfy $T_1^* = T_2^* \sim N(0, \hat{\sigma}^2/s^2)$. Fix t and a reference $i \in W(t)$. Since $\hat{k} = 1$, the discount (1) has $g_1^{(i)}(t) = 0$, so for $r \geq 0$ the vector $v = r\mathbf{1} \vee g^{(i)}(t)$ has $v_1 = r \leq v_2$. Because $T_2^* = T_1^*$, the exceedance event in $\hat{S}(v)$ reduces to $\{|T_1^*| > r\}$, and

$$\hat{S}(v) = \mathbb{P}_e\{|T_1^*| > r\} \leq 0.10 \quad \Longleftrightarrow \quad r \geq z_{0.95}\hat{\sigma}/s.$$

The radius $\psi(g^{(i)}(t))$ therefore equals $z_{0.95}\hat{\sigma}/s$ for every t and every $i \in W(t)$, and the maximum over $W(t)$ gives $\hat{r}_\alpha^W(t) = z_{0.95}\hat{\sigma}/s$. The accepted set multiplies this radius by $s/\sqrt{100}$, which gives the claimed interval.

Step 3. The set misses μ whenever no rare loss occurs. All A_j equal zero with probability 0.99^{100}. On that event every loss lies in $[0, \epsilon]$, so $\hat{\mu} \leq \epsilon$. The sample mean minimizes the mean squared deviation, so

$$\hat{\sigma}^2 \leq \frac{1}{100}\sum_{j=1}^{100}(L_{j1} - \epsilon/2)^2 \leq \frac{\epsilon^2}{4}.$$

The upper endpoint therefore satisfies

$$\hat{\mu} + z_{0.95}\hat{\sigma}/\sqrt{100} \leq \epsilon(1 + z_{0.95}/20) < 1.1 \times 10^{-4} < \mu,$$

and $\hat{C}_\alpha^W$ covers μ with probability at most $1 - 0.99^{100}$.

Step 4. The bound transfers to the exact-tail interval. The exact-tail interval I_m of (4) lies inside $\hat{C}_\alpha^W$, because $\hat{C}_\alpha^W$ is an interval. The exact-tail interval therefore obeys the same bound, with a grid failure counted as noncoverage. □

Table 7 reports coverage with $B = 2000$ bootstrap draws under this design and under a control with two identical $\mathrm{Uniform}[0,1]$ columns. B-Zoom and the band share draws and radii, so grid rounding accounts for their width difference. The Hoeffding union-bound interval $[\hat{\mu}_{\hat{k}} \pm r_{\mathrm{H}}] \cap [0, M]$ of Section 4.3, with $r_{\mathrm{H}} = M\sqrt{\log(2K/\alpha)/(2m)}$, covers $\mu_{\hat{k}}$ with probability at least $1-\alpha$ conditionally on $\mathcal{G}_{\mathrm{tr}}$, at every m and for every loss law on $[0, M]$. At $K = 2$, $m = 100$ and $M = 1$, its unclipped width is $2r_{\mathrm{H}} = 0.2716$.

Proof. Step 1. Each column. Given $\mathcal{G}_{\mathrm{tr}}$, column k has m i.i.d. entries in $[0, M]$ with mean μ_k. Hoeffding's two-sided inequality (Hoeffding, 1963) therefore gives, for each k,

$$\mathbb{P}\big(|\hat{\mu}_k - \mu_k| > r_{\mathrm{H}} \mid \mathcal{G}_{\mathrm{tr}}\big) \le 2\exp(-2mr_{\mathrm{H}}^2/M^2) = \alpha/K.$$

Step 2. Union bound. A union bound over the K columns, which needs no independence between them, gives $\max_k |\hat{\mu}_k - \mu_k| \le r_{\mathrm{H}}$ with conditional probability at least $1-\alpha$. On this event $[\hat{\mu}_k \pm r_{\mathrm{H}}]$ contains μ_k for every k, so coverage holds at $\hat{k}$ under any selection rule.

Step 3. Clipping. Intersecting with $[0, M]$ keeps $\mu_{\hat{k}}$, because $\mu_{\hat{k}} \in [0, M]$. □

Variance-sensitive intervals for bounded losses (Maurer & Pontil, 2009) offer an alternative to the Hoeffding union-bound interval. A rule that chooses between the Hoeffding and a Gaussian interval after reading the validation data inherits neither interval's guarantee. It needs a guarantee for the data-dependent choice or a separate sample reserved for choosing the interval.

Table 7: Coverage under nonconstant rare losses and under a uniform control. The three Gaussian-approximation methods cover 0.630 under rare losses against nominal coverage 0.90. The Hoeffding union-bound interval covers every replication in both designs. Each design has 5000 independent replications at $m = 100$, $K = 2$ and $B = 2000$, with identical loss columns for the two checkpoints. The exact Gaussian tail row is the exact-tail interval without a grid, $\hat{\mu} \pm z_{0.95}\hat{\sigma}/\sqrt{100}$. Entries give coverage with its Clopper–Pearson 95% interval, then mean width. B-Zoom and the band share bootstrap draws.

Method	Rare events	Uniform control
B-Zoom (zero floor)	0.630 [0.617, 0.644]; 0.0254	0.894 [0.886, 0.903]; 0.0945
Band	0.630 [0.617, 0.644]; 0.0254	0.894 [0.886, 0.903]; 0.0946
Exact Gaussian tail	0.630 [0.617, 0.644]; 0.0254	0.895 [0.886, 0.903]; 0.0945
Hoeffding union-bound interval	1.000 [0.999, 1.000]; 0.1460	1.000 [0.999, 1.000]; 0.2716

C.5 The near-tie design

Figure 3 summarizes coverage across all fifteen near-tie cells at nominal 0.90 and $M = 1$, each over 500 replications. B-Zoom covers between 0.916 and 0.996, and its mean width is 0.80 to 0.93 times Bonferroni's in every cell. The uncorrected interval covers between 0.000 and 0.818, and its Clopper–Pearson 95% interval lies below nominal in every cell.

C.6 The hybrid interval

The *scaling design* of Appendix E.2 starts with Gaussian losses at $m = 500$. Before clipping, one checkpoint's mean is two standard errors below those of $K-1$ tied, correlated rivals. In this design, the hybrid interval of Andrews et al. (2024), which combines conditioning on the selected checkpoint with a simultaneous band, is narrower than B-Zoom. It treats the estimates as Gaussian with a known covariance, for which we substitute the *Ledoit–Wolf shrinkage* estimate, a weighted average of the sample covariance and a scaled identity matrix (Ledoit & Wolf, 2004). At $K = 150$, 500 and

1500, the hybrid covers 0.898, 0.928 and 0.934, with 95% limits $[0.868, 0.923]$, $[0.902, 0.949]$ and $[0.909, 0.954]$. Its limits straddle nominal at $K = 150$ and lie above it at $K = 500$ and 1500.

The hybrid's mean width is 0.0187 in every cell, against B-Zoom's 0.0266, 0.0280 and 0.0290 at coverage 0.982, 0.992 and 0.996. The uncorrected interval covers 0.890, 0.920 and 0.924 at widths 0.0182, 0.0182 and 0.0181, so it covers near nominal in this design and the hybrid is 3% wider than it.

The hybrid eigendecomposes the dense $K \times K$ covariance, so we run it only at $K \leq 2000$. At $K = 3000$ and $10{,}000$, where no hybrid result exists, B-Zoom covers 0.992 and 0.996 at widths 0.0297 and 0.0308.

C.7 Endpoint search and timing

The exact endpoint search returns the endpoints of the hull (4) that the full sweep of Section 2.4 reports, without assuming that the accepted grid values are contiguous. Its search for the lower endpoint rests on one monotone property. Raising t raises no discount and removes no reference from $W(t)$, so the radius $\hat{r}^{W}_{\alpha,B}$ is non-decreasing in t. If $t < t' \leq \hat{\mu}_{\hat{k}}$ are grid values and t is accepted, then

$$\sqrt{m}\,(\hat{\mu}_{\hat{k}} - t')/s_{\hat{k}} \leq \sqrt{m}\,(\hat{\mu}_{\hat{k}} - t)/s_{\hat{k}} \leq \hat{r}^{W}_{\alpha,B}(t) \leq \hat{r}^{W}_{\alpha,B}(t'),$$

and t' is accepted.

Any accepted grid values at or below $\hat{\mu}_{\hat{k}}$ therefore form a block ending at t_ℓ, the largest grid value at or below $\hat{\mu}_{\hat{k}}$. The search first queries t_ℓ. If t_ℓ is accepted, bisection on the grid values up to t_ℓ returns the smallest accepted one. If t_ℓ is rejected, the contrapositive of the display rejects every grid value below it, and the search scans upward from t_ℓ to the first accepted value. A grid with no value at or below $\hat{\mu}_{\hat{k}}$ starts this scan at its smallest value.

Above $\hat{\mu}_{\hat{k}}$, acceptance can fail and resume, so bisection is unavailable. A downward scan from the largest grid value stops at the largest accepted value, whatever rejected values lie below it. If no grid value is accepted, both the search and the full sweep report a grid failure. Each query computes the maximum over all of $W(t)$ afresh, with the same array A and rank $k_\alpha(B)$ as the full sweep (Algorithm 1). Queried values are cached, so the two scans together query each grid value above $\hat{\mu}_{\hat{k}}$ at most once.

A grid of n_{grid} values with $n_- \geq 1$ of them at or below $\hat{\mu}_{\hat{k}}$ therefore needs at most $n_{\mathrm{grid}} - n_- + 1 + \lceil \log_2 n_- \rceil$ queries, counting the upper grid values, the query at t_ℓ and the bisection steps. The experimental grid has 400 equally spaced values on $[\hat{\mu}_{\hat{k}} \pm 1.02\,\psi_B(0) s_{\hat{k}}/\sqrt{m}]$, so $n_- = 200$ and the bound is 209.

Only the full sweep records the decision at every grid value, so only it can check contiguity. In all 800 cases of Table 6, the accepted grid values form one contiguous block, which does not exclude rejected values between adjacent grid points. The search returns the full-sweep endpoints in every case, with a mean of 13.9 to 36.3 queries across the four cells of arm and sample size.

The timing in Section 2.4 is the median over three single-thread repetitions, after an untimed warmup, on the main-arm path of seed 20260729 with $B = 2000$ and $\alpha = 0.10$. The method order changes across repetitions, and totals exclude input loading. There the endpoint search queries 42 of the 400 grid values and matches the full sweep in every repetition.

D Estimator ablations

Varying the number of draws, the multiplier law, the floor and α on the near-tie design keeps B-Zoom's coverage above nominal.

On the 800 matrices of Table 8, B-Zoom misses 3, 5, 13 and 26 times at $\alpha = 0.01$, 0.05, 0.10 and 0.20, against 8, 40, 80 and 160 expected at exact nominal coverage.

Table 8: Estimator ablations at nominal 0.90 on the four near-tie cells with $K \in \{150, 1000\}$ and $m \in \{50, 200\}$, 200 matrices each (Table 10), with exact Clopper–Pearson (CP) limits. The default uses Gaussian multipliers, floor 0 and $B = 2000$ (third row), and each other row changes one of these settings. The floor $0.2M$ exceeds every $\hat{\sigma}_k$ and so gives every column the common scale $0.2M$. The coverage proof assumes Gaussian multipliers, so it does not cover the Rademacher row, whose multipliers are independent signs ± 1 of equal probability.

Multipliers	Floor $\sigma_{\min}/M$	B	Coverage	CP 95%	Mean width
Gaussian	0	500	0.9838	$[0.9724, 0.9913]$	0.0765
	0	1000	0.9838	$[0.9724, 0.9913]$	0.0764
	0	2000	0.9838	$[0.9724, 0.9913]$	0.0764
	0	5000	0.9838	$[0.9724, 0.9913]$	0.0764
	10^{-4}	2000	0.9838	$[0.9724, 0.9913]$	0.0760
	10^{-3}	2000	0.9800	$[0.9677, 0.9885]$	0.0750
	10^{-2}	2000	0.9800	$[0.9677, 0.9885]$	0.0754
	5×10^{-2}	2000	0.9800	$[0.9677, 0.9885]$	0.0774
	0.2 (common scale)	2000	0.9888	$[0.9788, 0.9948]$	0.0835
Rademacher	0	2000	0.9788	$[0.9662, 0.9876]$	0.0742

E Reproducibility

This appendix gives the settings that the protocol of Section 4 leaves out.

E.1 Training runs and held-out data

Table 9 gives the training setting and the scored loss of each arm.

Table 9: Training settings. All arms use 100 seeds, 3000 SGD steps, batch size 128, momentum 0.9, weight decay 5×10^{-4}, and a 25,000-example held-out pool. The cosine learning-rate schedule starts at 0.1, or at 0.01 for fine-tuning.

Setting / data	Network	Train size	Augmentation	Scored loss
Main / CIFAR-10	ResNet-8	25,000	on	cross-entropy
Interior optimum / CIFAR-10	ResNet-8	5000	off	cross-entropy
Label noise / CIFAR-10	ResNet-8	25,000	on	cross-entropy and 0–1
CIFAR-100	ResNet-18	25,000	on	cross-entropy and 0–1
Fine-tuning / CIFAR-10	ResNet-18	25,000	on	cross-entropy

The ResNet-8 has 78,042 parameters and the CIFAR-100 ResNet-18 has 11,220,132. Every fine-tuning seed starts from the final iterate of CIFAR-100 seed 20260729 with a reinitialized classification head. The label-noise arms replace each training label independently with probability 0.2 by a class drawn uniformly from the other classes. All seeds share one corrupted label set, and the pool keeps its clean labels.

The full path stores the iterate after each of the 3000 updates. Selecting a stored iterate after training ends differs from validation-based early stopping (Prechelt, 1998), which uses validation losses to decide when training stops. The *checkpoint grid* of size K keeps every $(3000/K)$th stored iterate, starting with the first, so checkpoint $k \in [K]$ is the iterate after $1 + 3000(k-1)/K$ updates. A checkpoint grid therefore never includes the initialization and includes the final iterate only at $K = 3000$.

The position of checkpoint k on a checkpoint grid is $(k-1)/(K-1)$, and its position on the full path is its position on the grid with $K = 3000$. Over the 12,000 cases per arm, the selected checkpoint's median position on the full path is 0.165 on the interior-optimum arm and 0.967 on the main arm. On the interior-optimum arm, the population winner on the full path has median position 0.163 across training seeds. Averaged over those seeds, the population risk is 1.211 at its minimum along the path and 1.597 at the last checkpoint.

A fixed seeded permutation separates the training images from the pool. Each population risk μ_k is computed in double precision from the stored pool losses and is used only for scoring, never in the interval computation. Network evaluation uses batch-normalization running statistics and mixed precision, with both losses computed from the same logits. Stored cross-entropy losses are capped at $M = 10$ and kept in half precision, whose rounding moves a column mean by at most $2^{-8} \approx 3.9 \times 10^{-3}$. The 0–1 entries are exact.

E.2 Synthetic designs

Table 10 gives the generators of the near-tie and scaling designs. Near-tie results are in Section 4.3 and Appendices C.5 and D. Appendix C.6 holds the scaling-design results, and Appendix C.4 describes the rare-loss designs.

Table 10: Synthetic designs. Clipped-Gaussian column means are evaluated in closed form.

Study	Generator and design
Near tie	Independent Gaussian columns clipped to $[0, 1]$. Three columns have pre-clipping mean 0.5 and standard deviation 10^{-4}. For $j \in \{1, \ldots, K-3\}$, the jth remaining column has pre-clipping mean $0.5 + 0.25j/(K-3)$, with standard deviation 0.12 for odd j and 10^{-4} for even j. $K \in \{150, 1000, 10{,}000\}$, $m \in \{25, 50, 100, 200, 500\}$, 500 replications. Ablations use the first 200 replications at $K \in \{150, 1000\}$, $m \in \{50, 200\}$.
Scaling	Gaussian columns clipped to $[0, 1]$. The first column has pre-clipping mean $(1-\Delta)/2$. The other $K-1$ columns have pre-clipping mean $(1+\Delta)/2$, so Δ is the pre-clipping gap. All columns share standard deviation $(1-\Delta)/8$. A common Gaussian component gives pre-clipping correlation $\rho = 0.9$. The gap Δ equals two pre-clipping standard errors $(1-\Delta)/(8\sqrt{m})$ of a column mean. $K \in \{150, 500, 1500, 3000, 10{,}000\}$, $m = 500$, 500 replications.

E.3 Comparators, seeds and code

Cox (1975) analyzed sample splitting, which selects on one part of the data and tests on the other, and found its efficiency high in a problem with normal means. In exponential-family models, Fithian et al. (2014) construct selective tests that improve power over sample splitting. Randomizing the selection can improve power (Tian & Taylor, 2018) or yield simple stability-based corrections (Zrnic & Jordan, 2023). These selective methods randomize the selection or assume a parametric model for the losses, so sample splitting is the only one of these alternatives that we compare.

The zoom baseline takes its band radius at zero-based rank $\lceil (B-1)(1-\alpha) \rceil$, which equals $k_\alpha(B)$ at $B = 2000$ and $\alpha \in \{0.01, 0.05, 0.10, 0.20\}$, and inverts at rank $\lceil B(1-\alpha) \rceil - 1$. It divides variances by $m-1$, so its empirical standard deviations are $\sqrt{m/(m-1)}$ times B-Zoom's. Because its draws have unit variance, this factor widens its intervals.

The zoom baseline factorizes a dense covariance, so on the real arms we ran it only on checkpoint grids with $K \leq 600$. It estimates the covariance from the same m validation rows as B-Zoom. It uses $B = 2000$ and 200 grid points.

Data seeds depend only on the cell and the replication. B-Zoom, the band and the single-reference ablation share the bootstrap draws on each loss matrix, while the other randomized comparators make their own. Training seeds are the integers 20260729 through 20260828. The joint-coverage study of Appendix C.2 indexes the main arm by $f = 1$ and the interior-optimum arm by $f = 2$. It shuffles each arm's sorted training seeds with NumPy's `SeedSequence` seeded by $(2026091601, f)$ and keeps the first 50. Its validation and multiplier seeds derive from $(20260916, f, \text{training seed}, m, \text{replication})$.

A training seed does not fix the training path across devices. Dense-covariance samplers can also return different draws across numerical builds, because eigenvectors are unique only up to sign and, for repeated eigenvalues, up to rotation.

For every method and real arm, the approximate Student-t coverage limits of Section 4 are 1.3–2.0 times wider than Clopper–Pearson limits that treat the arm's 12,000 cases as independent.

The release contains the estimator and generator code, the training configurations, the seeds, and code for the full sweep and the exact endpoint search. Its per-replication rows record targets, timings and zoom-baseline decline reasons, and they reproduce the statistical summaries. For the 800 cases of Table 6, the release records the acceptance decision at every grid value. It omits the per-example loss matrices of the real arms.