

Finite-sample certification of coverage regret for budgeted document selection under adaptive stratified auditing

Nora Serdyukova

Department of Statistics, University of Concepcion, Chile
nserdyukova@udec.cl

Abstract

Selecting a budget-constrained set of documentary fragments, for instance the knowledge base of a retrieval-augmented generation system, raises a statistical question: how much expected query coverage is lost relative to the best feasible selection when the query distribution is unknown? We construct finite-sample, distribution-free upper confidence bounds for this coverage regret that remain valid after data-dependent selection. A pilot stage may select the document set, screen competitors, build comparison networks and learn an annotation policy. Then an independent audit acquires relevance labels with probabilities that are constant within observable strata. The main result is a simultaneous bound for all retained coverage contrasts. The bound is made observable by empirical-Bernstein variance calibration or by a one-stage construction without calibration, and it accounts for estimated stratum proportions. For maximum-type coverage we prove variance contractions for swaps of redundant fragments, give a condition under which multiscale radii improve on a direct union bound, and show that such redundancy cannot be certified from samples alone. A simulation study confirms validity and quantifies the gains from variance calibration and certified redundancy.

MSC2020 subject classifications: Primary 62G15, 60E15; secondary 68P20, 90C27.

Keywords and phrases: Regret certification, postselection inference, empirical Bernstein inequalities, chaining, submodular coverage, retrieval-augmented generation.

1 Introduction

Constructing a compact knowledge base from a large collection of documentary fragments, for example the retrieval corpus of a retrieval-augmented generation (RAG) system, requires choosing which fragments to retain under a storage or annotation budget. The selected collection should support future queries, whose probability distribution is generally not known in advance. This leads to a stochastic set-optimization problem: maximize expected query coverage, then certify the loss in coverage relative to the best feasible collection. The distinction between a computational approximation guarantee and a statistical coverage certificate is essential. A near-optimal value for the empirical selection objective need not imply a small population regret.

For coverage defined by maxima of nonnegative relevance scores, the population objective is monotone and submodular. This connects the selection problem to classical submodular maximization (Nemhauser et al., 1978; Krause and Golovin, 2014). Yet the statistical component is different: relevance assessments may be costly, and queries may be deliberately oversampled in contexts believed to distinguish competing document selections. Such adaptive data acquisition must not be allowed to invalidate a confidence statement about the unknown query population.

Concentration for means with unknown variances is classical (Maurer and Pontil, 2009; Boucheron et al., 2013). Stratified empirical-Bernstein methods have been developed for other estimation problems, for instance by Burgess and Chapman (2021), and context-dependent evaluation and inverse-probability correction have substantial precedents (Dudík et al., 2011). Generic chaining (Talagrand, 2005; Dirksen, 2015) and local complexity theory (Bartlett et al., 2005; Koltchinskii, 2006) explain why it may be possible to replace a crude logarithm of the number of candidates by a multiscale complexity. Consequently, neither stratification nor chaining is claimed to be new here. The goal is to give a self-contained and precisely scoped concentration argument for *documentary coverage contrasts*, including all the conditioning, sampling, and calibration steps needed to turn the inequality into a valid postselection regret certificate.

Our results are the following. First, we derive a conditional, variance-sensitive Bernstein bound for increments of the empirical coverage process (Lemma 1). Second, a pre-audit multiscale representation yields a simultaneous pathwise bound for all retained competitors (Theorem 1). Third, the variance factors can be replaced by simultaneous empirical upper confidence bounds (Theorem 2). Fourth, a separate estimate of the unknown context proportions yields an unconditional confidence certificate for reduced-universe regret (Theorem 4); we also give a complementary pooled Bennett formulation (Proposition 4). Finally, a contraction lemma links the edge variances to both semantic redundancy and the frequency with which a swapped fragment is actually responsible for coverage (Lemma 5); Corollary 4 provides an explicit condition for a two-level improvement over a direct union bound. Two complementary refinements are proved: an edgewise allocation of confidence error, which controls each distinct edge only once (Proposition 3), and a one-stage empirical-Bernstein network bound that does not require preliminary variance calibration (Theorem 3). A finite-sample rare-query example (Proposition 5) explains why empirical profile agreement cannot certify uniform approximation envelopes.

A simulation study (Section 10) confirms the validity of all certificates at the nominal level and compares their sizes; it shows when variance calibration and the multiscale network pay off, and that the latter requires certified redundancy envelopes.

The reader should distinguish *validity*, which is established under specified assumptions, from *efficiency*, which depends on the geometry of the competitors and on the audit design.

2 Statistical model, selection problem, and notation

2.1 Query population and document selection

Let $(\mathcal{Q}, \mathcal{Q})$ be a measurable query space. Let Q have an arbitrary and unknown probability distribution P on this space. The set of available documentary fragments is the finite nonempty set $\mathcal{A} = \{1, \dots, K\}$, with deterministic costs $c_j > 0$. For a budget $B \geq 0$, define the feasible selection family

$$\mathcal{H}_B = \left\{ S \subseteq \mathcal{A} : \sum_{j \in S} c_j \leq B \right\}. \quad (1)$$

The empty set is feasible and is assigned zero coverage. A relevance function $r_j : \mathcal{Q} \rightarrow [0, 1]$ is associated with each fragment. Unless stated otherwise, these functions are fixed, measurable and evaluated without labeling error when a query is audited. The coverage of a selection S on query

q is

$$h_S(q) = \begin{cases} \max_{j \in S} r_j(q), & S \neq \emptyset, \\ 0, & S = \emptyset. \end{cases} \quad F(S) = \mathbb{E}_P h_S(Q). \quad (2)$$

Thus $0 \leq h_S, F(S) \leq 1$.

Since $\mathcal{H}_B$ is finite and nonempty, there exists at least one population maximizer. Fix any

$$S^* \in \arg \max_{S \in \mathcal{H}_B} F(S), \quad R_B(S) = F(S^*) - F(S) \in [0, 1]. \quad (3)$$

The definition does not presume that the maximum can be calculated computationally on a large instance. It provides the statistical target.

Proposition 1 (Monotonicity and submodularity). *For every fixed query q , the set function $S \mapsto h_S(q)$ is normalized, monotone and submodular. The population function F possesses the same properties.*

Proof. Normalization follows from $h_\emptyset(q) = 0$, while monotonicity follows because a maximum cannot decrease when additional elements are included. If $A \subseteq B$ and $j \notin B$, then

$$h_{A \cup \{j\}}(q) - h_A(q) = [r_j(q) - h_A(q)]_+ \geq [r_j(q) - h_B(q)]_+ = h_{B \cup \{j\}}(q) - h_B(q),$$

where $[x]_+ = \max(x, 0)$. This is the diminishing-returns characterization of submodularity. Expectations preserve normalization, monotonicity and the displayed pointwise inequality. $\square$

2.2 Pilot selection and honest screening

We use $\mathcal{F}_0$ to denote the sigma-field generated by all pilot observations and all randomness used before the independent audit and unlabelled-proportion samples. The pilot produces a feasible selected set $\hat{S} \in \mathcal{H}_B$ and a finite (possibly random) candidate family $\mathcal{C} \subseteq \mathcal{H}_B$ such that $\hat{S} \in \mathcal{C}$. Both are $\mathcal{F}_0$ -measurable. The *documentary contrast* against a competitor $T \in \mathcal{C}$ is

$$D_T(q) = h_T(q) - h_{\hat{S}}(q), \quad \Delta_T = \mathbb{E}_P D_T(Q) = F(T) - F(\hat{S}). \quad (4)$$

Every contrast is in $[-1, 1]$ and $D_{\hat{S}} \equiv 0$. The pilot screening guarantee, when screening is used, is

$$\mathbb{P}\{\text{some } S^* \in \mathcal{C}\} \geq 1 - \alpha_s. \quad (5)$$

If $\mathcal{C} = \mathcal{H}_B$, this property holds deterministically and $\alpha_s = 0$. Here and below “some” allows non-uniqueness of the population optimum. A sufficient, explicitly verifiable pilot rule is established in Proposition 2.

Proposition 2 (A conservative pilot screening rule). *Assume that a fully labeled iid pilot of n_0 queries from P , independent of subsequent data, is available, and let $\hat{F}_0(S) = n_0^{-1} \sum_{i=1}^{n_0} h_S(Q_i^{(0)})$ for all $S \in \mathcal{H}_B$. For $0 < \alpha_s < 1$, set*

$$r_0 = \sqrt{\frac{\log(2|\mathcal{H}_B|/\alpha_s)}{2n_0}}. \quad (6)$$

After choosing any pilot-measurable $\hat{S} \in \mathcal{H}_B$, let

$$\mathcal{C} = \{\hat{S}\} \cup \left\{ T \in \mathcal{H}_B : \hat{F}_0(T) - \hat{F}_0(\hat{S}) + 2r_0 \geq 0 \right\}. \quad (7)$$

Then (5) holds.

Proof. By Hoeffding’s inequality (Hoeffding, 1963) and a union bound, with probability at least $1 - \alpha_s$, $|\hat{F}_0(S) - F(S)| \leq r_0$ simultaneously for all $S \in \mathcal{H}_B$. On this event, if S^* is a population maximizer,

$$\hat{F}_0(S^*) - \hat{F}_0(\hat{S}) + 2r_0 \geq F(S^*) - F(\hat{S}) \geq 0.$$

Thus $S^* \in \mathcal{C}$. The argument remains valid when $\hat{S}$ is chosen adaptively from the same pilot, precisely because the pilot bound is uniform. $\square$

2.3 Observable contexts and the audit design

Let $X = \phi(Q)$ be an observable categorical feature, inexpensive to record, with values in $\mathcal{S} = \{0, \dots, H - 1\}$, where $H \geq 2$ is fixed. Let $\pi_h = \mathbb{P}(X = h) > 0$ and $P_h(\cdot) = \mathbb{P}(Q \in \cdot \mid X = h)$. Both π_h and P_h are unknown. We distinguish an unlabelled sample, used to estimate the population proportions, from the labeled audit sample.

Assumption 1 (Independent unlabelled queries). The queries $Q_1^{(u)}, \dots, Q_{N_u}^{(u)}$ are, conditionally on $\mathcal{F}_0$, iid from P and independent of the queries and relevance labels used for audit. The empirical proportions

$$\hat{\pi}_h = \frac{1}{N_u} \sum_{i=1}^{N_u} \mathbf{1}\{\phi(Q_i^{(u)}) = h\}$$

are observed, $N_u \geq 1$. The multiscale networks and acquisition policy may use this unlabelled sample, but may not use audit relevance labels.

Assumption 2 (Stratified, label-ignorable audit). Conditionally on the pilot, the unlabelled sample, and the labeled counts $n_h \geq 2$, the acquired audit queries $Q_{h,1}, \dots, Q_{h,n_h}$ are iid from P_h within each stratum, and the collections are mutually independent across strata. The counts and the mechanism that chooses them contain no information about the unobserved relevance values beyond the contexts. All relevance values needed to calculate $h_T(Q_{h,i})$, $T \in \mathcal{C}$, are observed for an acquired query.

An important implementation satisfying Assumption 2 starts with fresh iid queries, observes X , and acquires full relevance labels with probability $p_h \in (0, 1]$ in stratum h , where p_h is learned before audit and is constant within the stratum. Conditional on the resulting stratum counts (and the design information), the selected queries have the required conditional law. A design based on a new feature *within* an existing stratum does not generally satisfy this assumption; its selection bias must be addressed separately, for example through known-propensity estimators as in Dudík et al. (2011).

The sample size n_h may be random. Every assertion below is conditional on the pre-audit information and on the observed admissible counts, and therefore remains valid after averaging over their joint distribution. If some $n_h < 2$, our empirical-variance formulas are not used and the trivial regret upper bound 1 can be returned.

2.4 Conditional and population coverage contrasts

For every $T \in \mathcal{C}$, write

$$\mu_{h,T} = \mathbb{E}[D_T(Q) \mid X = h], \quad \bar{D}_{h,T} = \frac{1}{n_h} \sum_{i=1}^{n_h} D_T(Q_{h,i}), \quad \hat{\Delta}_T = \sum_{h \in \mathcal{S}} \hat{\pi}_h \bar{D}_{h,T}. \quad (8)$$

Table 1: Main notation. All candidate-dependent objects may be random through pre-audit pilot information.

Symbol	Meaning
$\mathcal{H}_B, \mathcal{C}, \widehat{S}, S^*$	feasible, retained, selected, optimal document sets
$h_S, F(S), R_B(S)$	query coverage, mean coverage, population regret
D_T, Δ_T	contrast against $\widehat{S}$ and its population mean
$H, \pi_h, \widehat{\pi}_h, P_h$	number, true/estimated weights, conditional query law
$n_h, \overline{D}_{h,T}, \mu_{h,T}$	audited count, sample/conditional contrast means
Z_T	weighted, centered audit error
$\mathcal{N}_\ell, \Pi_\ell, \mathcal{E}_\ell$	networks, parent maps, distinct comparison edges
d_2, d_∞	conditional variance and range semimetrics
$\overline{v}_e, \overline{b}_e$	valid observable variance/range upper bounds
$\alpha_s, \alpha_a, \alpha_v, \alpha_\pi, \alpha_g$	error budgets for screening, audit, variance, weights, means
$\alpha_B, \alpha'_v, \alpha_{\text{act}}, \alpha_{\text{rep}}$	budgets for Bennett, its variances, activity, representation

The population contrast decomposes as $\Delta_T = \sum_h \pi_h \mu_{h,T}$. The *centered weighted audit process* is

$$Z_T = \sum_{h \in \mathcal{S}} \widehat{\pi}_h (\mu_{h,T} - \overline{D}_{h,T}), \quad Z_{\widehat{S}} = 0. \quad (9)$$

The symbols $\widehat{\Delta}_T$ and Z_T will not be confused: the former is observable, while the latter contains unknown conditional means and is used only in the analysis.

3 A local Bernstein inequality for documentary increments

For document selections U, V , let

$$G_{U,V}(q) = h_U(q) - h_V(q) \in [-1, 1]. \quad (10)$$

We identify the *directed edge* $e = (U, V)$ with $G_e = G_{U,V}$. For $h \in \mathcal{S}$, denote

$$v_{h,e} = \text{Var}(G_e(Q) \mid X = h), \quad \text{osc}_h(G_e) = \text{ess sup}_{P_h} G_e - \text{ess inf}_{P_h} G_e.$$

The latter is an essential range, hence is at most 2, while $0 \leq v_{h,e} \leq 1$. Condition on $\mathcal{F}_0$, the unlabelled observations and the counts, so that all $\widehat{\pi}_h$, n_h and pre-audit edges are fixed. Define

$$d_2^2(e) = \sum_h \frac{\widehat{\pi}_h^2}{n_h} v_{h,e}, \quad (11)$$

$$d_\infty(e) = \max_h \frac{\widehat{\pi}_h}{n_h} \text{osc}_h(G_e). \quad (12)$$

These are semimetrics on coverage profiles (they may vanish for distinct selections). For clarity, the square in d_2^2 is part of the notation.

Lemma 1 (Conditional Bernstein increment). *Under Assumption 2, for every pre-audit edge $e = (U, V)$ and every $t > 0$, conditional on the pre-audit information and counts,*

$$\mathbb{P} \left\{ Z_U - Z_V > \sqrt{2d_2^2(e)t} + \frac{1}{3}d_\infty(e)t \mid \text{pre-audit information, counts} \right\} \leq e^{-t}. \quad (13)$$

No assumption of independence across different edges is needed.

Proof. Since $D_U - D_V = G_{U,V}$, we have

$$Z_U - Z_V = \sum_{h \in \mathcal{S}} \sum_{i=1}^{n_h} \frac{\hat{\pi}_h}{n_h} (\mathbb{E}_h G_e - G_e(Q_{h,i})).$$

The summands are conditionally independent and centered. Their total conditional variance equals $d_2^2(e)$. Since $\mathbb{E}_h G_e$ lies in the essential range of G_e under P_h , each centered summand is almost surely at most $(\hat{\pi}_h/n_h) \text{osc}_h(G_e) \leq d_\infty(e)$. By Theorem 2.10 of [Boucheron et al. \(2013\)](#), if $X_1, \dots, X_n$ are independent, $\sum_i \mathbb{E} X_i^2 \leq v$ and $\sum_i \mathbb{E}[(X_i)_+^q] \leq q! v c^{q-2}/2$ for all integers $q \geq 3$, then $\mathbb{P}\{\sum_i (X_i - \mathbb{E} X_i) \geq \sqrt{2vt} + ct\} \leq e^{-t}$ for all $t > 0$. As noted in [Boucheron et al. \(2013, p. 38\)](#), right after their Corollary 2.11, variables with $X_i \leq b$ almost surely satisfy these conditions with $v = \sum_i \mathbb{E} X_i^2$ and $c = b/3$. Apply this to the centered summands above, for which $\sum_i \mathbb{E} X_i^2 = d_2^2(e)$ and $b = d_\infty(e)$; this gives (13). If $d_\infty(e) = 0$ or $d_2^2(e) = 0$, the summands vanish almost surely and the claim is trivial. $\square$

Remark 1 (On the linear constant). The classical tail form

$$\mathbb{P}\{Z_U - Z_V \geq u\} \leq \exp\{-u^2/(2[d_2^2(e) + d_\infty(e)u/3])\},$$

inequality (2.10) of [Boucheron et al. \(2013\)](#), is usually inverted to the weaker radius $\sqrt{2d_2^2(e)t} + \frac{2}{3}d_\infty(e)t$. The constant $\frac{1}{3}$ in (13) comes from the sub-gamma form cited above; it is the same constant that appears in Bennett's inequality as stated in [Maurer and Pontil \(2009, Theorem 3\)](#). It is used throughout.

Remark 2 (Why the essential range still matters). Low variance alone does not eliminate the Bernstein linear term. A profile may disagree with another only on rare queries, yet a rare observation can produce a large jump. A deterministic envelope, an externally certified uniform approximation, or an explicitly valid bound on $\text{osc}_h(G_e)$ is needed to shrink the linear term; numerical proximity of embeddings is not such a certificate.

4 A simultaneous localized multiscale inequality

Definition 1 (Pre-audit comparison network). Given $\mathcal{C}$ and $\hat{S}$, fix an integer $L \geq 1$, finite nonempty sets of document selections $\mathcal{N}_0, \dots, \mathcal{N}_L \subseteq \mathcal{C}$ and mappings $\Pi_\ell : \mathcal{C} \rightarrow \mathcal{N}_\ell$ such that

$$\mathcal{N}_0 = \{\hat{S}\}, \quad \Pi_0(T) = \hat{S}, \quad \Pi_L(T) = T \quad \text{for every } T \in \mathcal{C}. \quad (14)$$

The network sets and maps must be measurable with respect to the pre-audit information (which may include independent unlabelled queries). Define $e_\ell(T) = (\Pi_\ell(T), \Pi_{\ell-1}(T))$ and let

$$\mathcal{E}_\ell = \{e_\ell(T) : T \in \mathcal{C}, \Pi_\ell(T) \neq \Pi_{\ell-1}(T)\}, \quad \mathcal{E} = \bigcup_{\ell=1}^L \mathcal{E}_\ell. \quad (15)$$

Here the sets count *distinct directed edges*, not paths: within a level, an edge shared by many candidates is counted once. An edge that appears at several levels is counted in each corresponding $\mathcal{E}_\ell$, because the union bound in Theorem 1 is allocated level by level; Proposition 3 below avoids this repeated charge by a single union bound over distinct edges. For convenience levels with $\mathcal{E}_\ell = \emptyset$ can be removed. We assume the displayed levels are nonempty.

The identity $Z_{\widehat{S}} = 0$ gives the exact telescoping representation

$$Z_T = \sum_{\ell=1}^L [Z_{\Pi_\ell(T)} - Z_{\Pi_{\ell-1}(T)}]. \quad (16)$$

No probabilistic approximation is involved in this equality.

Theorem 1 (Conditional simultaneous multiscale Bernstein bound). *Assume the model of Section 2. Choose numbers $\alpha_1, \dots, \alpha_L > 0$ with $\sum_{\ell=1}^L \alpha_\ell = \alpha_a < 1$ and let*

$$t_\ell = \log \frac{|\mathcal{E}_\ell|}{\alpha_\ell}. \quad (17)$$

Suppose that for all distinct edges $e \in \mathcal{E}$ there are nonnegative, possibly audit-data-dependent random variables $\bar{v}_e, \bar{b}_e$ such that

$$\mathbb{P} \{ \forall e \in \mathcal{E} : d_2^2(e) \leq \bar{v}_e, \quad d_\infty(e) \leq \bar{b}_e \mid \text{pre-audit information, counts} \} \geq 1 - \alpha_v. \quad (18)$$

Define the path-dependent radius

$$\mathfrak{R}(T) = \sum_{\ell=1}^L \mathbf{1}_{\{e_\ell(T) \in \mathcal{E}_\ell\}} \left\{ \sqrt{2\bar{v}_{e_\ell(T)} t_\ell} + \frac{1}{3} \bar{b}_{e_\ell(T)} t_\ell \right\}. \quad (19)$$

Then, conditionally on the pre-audit information and counts,

$$\mathbb{P} \{ \forall T \in \mathcal{C} : Z_T \leq \mathfrak{R}(T) \mid \text{pre-audit information, counts} \} \geq 1 - \alpha_a - \alpha_v. \quad (20)$$

The result remains valid when $\mathcal{C}$, $\widehat{S}$, and the network are arbitrarily data-dependent through the independent pilot and unlabelled information.

Proof. For every level ℓ and every $e = (U, V) \in \mathcal{E}_\ell$, apply Lemma 1 with $t = t_\ell$. Its one-sided failure probability is bounded by $e^{-t_\ell} = \alpha_\ell / |\mathcal{E}_\ell|$. A union bound over the distinct edges in $\mathcal{E}_\ell$ gives failure probability at most α_ℓ ; a second union bound over levels gives a joint event Ω_a of probability at least $1 - \alpha_a$ on which

$$Z_U - Z_V \leq \sqrt{2d_2^2(e)t_\ell} + \frac{1}{3}d_\infty(e)t_\ell \quad \text{for every } e = (U, V) \in \mathcal{E}_\ell, \ell = 1, \dots, L.$$

On Ω_a , use the telescoping identity (16) for each $T \in \mathcal{C}$ and add the corresponding edge inequalities. Intersect with the calibration event in (18). The monotonicity of the square root and of the linear term permits replacement of each unknown $d_2^2(e), d_\infty(e)$ by $\bar{v}_e, \bar{b}_e$ without any independence assumption between Ω_a and the calibration event. A final union bound proves (20). Conditioning is legitimate because every edge and its allocated failure probability were fixed without examining audit relevance values. $\square$

Proposition 3 (A weighted, unique-edge simultaneous inequality). *Let $\mathcal{E} = \bigcup_{\ell=1}^L \mathcal{E}_\ell$ be the set of distinct directed edges of a pre-audit network as in Definition 1, and let $0 < \alpha_a < 1$. Conditionally on the pre-audit information and counts, choose weights $\omega_e > 0$ satisfying $\sum_{e \in \mathcal{E}} \omega_e \leq 1$. The weights may depend on the pilot, the unlabelled observations and the counts, but not on audit relevance labels. Put*

$$t_e = \log \frac{1}{\alpha_a \omega_e}, \quad e \in \mathcal{E}. \quad (21)$$

Suppose that the simultaneous edge calibration event (18) holds with conditional failure probability at most α_v , and define

$$\mathfrak{R}_\omega(T) = \sum_{\ell=1}^L \mathbf{1}\{e_\ell(T) \in \mathcal{E}_\ell\} \left\{ \sqrt{2\bar{v}_{e_\ell(T)} t_{e_\ell(T)}} + \frac{1}{3} \bar{b}_{e_\ell(T)} t_{e_\ell(T)} \right\}. \quad (22)$$

Then, conditionally on the pre-audit information and counts,

$$\mathbb{P}\{\forall T \in \mathcal{C} : Z_T \leq \mathfrak{R}_\omega(T) \mid \text{pre-audit information, counts}\} \geq 1 - \alpha_a - \alpha_v.$$

The same event bounds all candidate paths, even if a directed edge appears at more than one level.

Proof. Since $\alpha_a \omega_e < 1$, each t_e is positive. For each distinct edge $e = (U, V) \in \mathcal{E}$, apply Lemma 1 at $t = t_e$. Its conditional failure probability is at most $e^{-t_e} = \alpha_a \omega_e$. A single union bound over $e \in \mathcal{E}$ therefore has failure probability at most $\alpha_a \sum_e \omega_e \leq \alpha_a$. On its complement every edge inequality holds, regardless of how many times that edge occurs in the telescoping paths. Intersect with the simultaneous calibration event and sum along each path exactly as in Theorem 1. No independence between these events is needed. $\square$

Remark 3 (Uniform levels, repeated edges, and pre-audit tuning). If every edge appears at only one level, the allocation $\omega_e = \alpha_\ell / (\alpha_a |\mathcal{E}_\ell|)$ for $e \in \mathcal{E}_\ell$ gives $t_e = t_\ell$ and recovers Theorem 1. For repeated edges, let

$$\omega_e = \sum_{\ell: e \in \mathcal{E}_\ell} \frac{\alpha_\ell}{\alpha_a |\mathcal{E}_\ell|}.$$

These weights sum to 1, and $t_e \leq t_\ell$ whenever $e \in \mathcal{E}_\ell$. Consequently, the weighted radius is *never* larger than the levelwise radius (19) for this allocation, and it can be strictly smaller on paths with repeated edges. Other weights may prioritize difficult or frequently used edges, but they must be fixed without the audit relevance labels: optimizing them retrospectively from the audit is not covered by this proposition.

Remark 4 (The role of localization). The radius in (19) is *pathwise localized*: each T pays only for increments along its own approximation path, and an edge with negligible variance or range contributes little. Moreover, the pilot-screened family $\mathcal{C}$ can be much smaller than $\mathcal{H}_B$. This does not establish a universal entropy integral or a new generic chaining inequality; the generic mixed-tail background is developed, for example, by Dirksen (2015).

5 Observable calibration of all edge variances

Theorem 1 must not rely on oracle variances. We now provide explicit, fully observable choices of $\bar{v}_e$ with separate statistical error control. Recall the unbiased sample variance

$$s_n^2(y_1, \dots, y_n) = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

Lemma 2 (Empirical Bernstein inequality for a single mean). *Let $Y_1, \dots, Y_n$ be iid in $[a, b]$ for $n \geq 2$ and let s_n^2 be their unbiased sample variance. For $0 < \delta < 1$,*

$$\mathbb{P} \left\{ \mathbb{E}Y - \bar{Y} > \sqrt{\frac{2s_n^2 \log(2/\delta)}{n}} + \frac{7(b-a) \log(2/\delta)}{3(n-1)} \right\} \leq \delta. \quad (23)$$

The same assertion holds with $\bar{Y} - \mathbb{E}Y$ in place of $\mathbb{E}Y - \bar{Y}$.

Proof. For $a < b$, apply Theorem 4 of [Maurer and Pontil \(2009\)](#) to the rescaled observations $(Y_i - a)/(b - a) \in [0, 1]$, then multiply the resulting inequality by $b - a$. The sample variance $V_n = \{n(n-1)\}^{-1} \sum_{i < j} (Z_i - Z_j)^2$ used by [Maurer and Pontil \(2009\)](#) coincides with the unbiased sample variance s_n^2 , and it rescales by $(b - a)^2$. Apply the same theorem to $(b - Y_i)/(b - a)$ for the opposite one-sided deviation. If $a = b$ the deviations vanish identically. $\square$

Assumption 3 (Pre-audit residual predictors). For every stratum h and every distinct edge $e \in \mathcal{E}$, there is an $\mathcal{F}_0$ - and/or unlabelled-sample-measurable predictor $m_{h,e} \in [-1, 1]$. It is fixed without using the audit relevance labels.

Let

$$Y_{h,i,e} = (G_e(Q_{h,i}) - m_{h,e})^2, \quad M_{h,e} = (1 + |m_{h,e}|)^2, \quad \bar{Y}_{h,e} = \frac{1}{n_h} \sum_i Y_{h,i,e}, \quad (24)$$

and write $s_{Y,h,e}^2$ for the sample variance of the Y 's. Since $G_e \in [-1, 1]$, $0 \leq Y_{h,i,e} \leq M_{h,e} \leq 4$.

Lemma 3 (Uniform residual-second-moment upper bounds). *Let $E = |\mathcal{E}| \geq 1$. Set*

$$t_v = \log \frac{2HE}{\alpha_v}, \quad v_{h,e}^+ = \min \left\{ 1, \bar{Y}_{h,e} + \sqrt{\frac{2s_{Y,h,e}^2 t_v}{n_h}} + \frac{7M_{h,e} t_v}{3(n_h - 1)} \right\}.$$

Under Assumptions 2 and 3,

$$\mathbb{P} \left\{ \forall h, e: \quad v_{h,e} \leq v_{h,e}^+ \mid \text{pre-audit information, counts} \right\} \geq 1 - \alpha_v. \quad (25)$$

Proof. For each fixed h, e , the bias-variance decomposition gives

$$\mathbb{E}_h(G_e - m_{h,e})^2 = \text{Var}_h(G_e) + (\mathbb{E}_h G_e - m_{h,e})^2 \geq v_{h,e}.$$

The predictor is fixed under the conditioning, so the residual squares are iid and bounded by $M_{h,e}$. Apply the one-sided empirical Bernstein inequality of Lemma 2 (equivalently, Corollary 5 of [Maurer and Pontil \(2009\)](#) after rescaling) with $\delta = \alpha_v/(HE)$. Since $\log(2/\delta) = t_v$, the displayed empirical upper expression bounds $\mathbb{E}_h Y_{h,i,e}$ except on an event of probability at most $\alpha_v/(HE)$. Union over all HE pairs. Finally, the separate Popoviciu bound $v_{h,e} \leq (2)^2/4 = 1$ justifies the minimum with 1. $\square$

Theorem 2 (Observable simultaneous chaining radius). *Under Assumptions 2 and 3, let $\bar{b}_{h,e} \in [0, 2]$ be known pre-audit constants satisfying $\text{osc}_h(G_e) \leq \bar{b}_{h,e}$. The universally valid choice is $\bar{b}_{h,e} = 2$. Put*

$$\bar{v}_e = \sum_h \frac{\hat{\pi}_h^2}{n_h} v_{h,e}^+, \quad \bar{b}_e = \max_h \frac{\hat{\pi}_h}{n_h} \bar{b}_{h,e}. \quad (26)$$

Then the radius (19) is observable from the pilot, unlabelled sample and acquired audit labels, and conclusion (20) holds with probability at least $1 - \alpha_a - \alpha_v$.

Proof. By Lemma 3, the collection of variance bounds in (26) dominates $d_2^2(e)$ simultaneously with conditional probability at least $1 - \alpha_v$. The deterministic oscillation bounds imply $d_\infty(e) \leq \bar{b}_e$ without additional probability loss. All requirements of Theorem 1 are satisfied. $\square$

Remark 5 (No hidden additional labeled sample). The $v_{h,e}^+$ may be estimated on the same audited labels used to form $\bar{D}_{h,T}$. The argument does not condition on the estimated variances; instead it intersects a Bernstein event involving the true variances with an empirical calibration event that holds simultaneously. This avoids an incorrect independence claim and does not double the annotation budget.

Remark 6 (A predictor-free alternative). Assumption 3 can be dropped at the price of a different constant. For iid observations in $[a, b]$, inequality (3) in Theorem 10 of [Maurer and Pontil \(2009\)](#), applied after rescaling to $[0, 1]$, gives

$$\mathbb{P}\left\{\sigma > s_n + (b - a)\sqrt{2\log(1/\delta)/(n - 1)}\right\} \leq \delta,$$

where σ^2 is the variance and s_n the unbiased sample standard deviation. (Their theorem is stated for $\sqrt{\mathbb{E}V_n}$; in the iid case the expected unbiased sample variance $\mathbb{E}V_n$ equals σ^2 .) With $G_e \in [-1, 1]$ and $\delta = \alpha_v/(HE)$, the choice

$$v_{h,e}^+ = \min\left\{1, \left(s_{h,e} + 2\sqrt{2\log(HE/\alpha_v)/(n_h - 1)}\right)^2\right\},$$

where $s_{h,e}$ is the sample standard deviation of the $G_e(Q_{h,i})$, satisfies (25) without any pre-audit predictor. Both forms have the same orders in n_h , namely terms of order $n_h^{-1/2}$ scaled by an empirical standard deviation and terms of order n_h^{-1} . They differ in their constants and in the squared bias $(\mathbb{E}_h G_e - m_{h,e})^2$, which the residual form pays and the predictor-free form avoids; which one is tighter depends on the data and on the quality of the predictor.

Theorem 3 (One-stage empirical-Bernstein multiscale certificate). *Assume the stratified, label-ignorable audit design (Assumption 2) and a pre-audit network as in Definition 1. For every h and every distinct directed edge $e \in \mathcal{E}$, let $\bar{G}_{h,e} = n_h^{-1} \sum_{i=1}^{n_h} G_e(Q_{h,i})$, and let $s_{h,e}^2$ be the unbiased sample variance of $G_e(Q_{h,1}), \dots, G_e(Q_{h,n_h})$. Let $\bar{b}_{h,e} \in [0, 2]$ be the known pre-audit constants of Theorem 2, satisfying $\text{osc}_h(G_e) \leq \bar{b}_{h,e}$. For $\sum_\ell \alpha_\ell = \alpha_a < 1$ set*

$$t_\ell^{\text{EB}} = \log \frac{2H|\mathcal{E}_\ell|}{\alpha_\ell} \quad (27)$$

and define

$$\mathfrak{R}_{\text{EB}}(T) = \sum_{\ell=1}^L \mathbf{1}\{e_\ell(T) \in \mathcal{E}_\ell\} \sum_{h \in \mathcal{S}} \hat{\pi}_h \left\{ \sqrt{\frac{2s_{h,e_\ell(T)}^2 t_\ell^{\text{EB}}}{n_h}} + \frac{7\bar{b}_{h,e_\ell(T)} t_\ell^{\text{EB}}}{3(n_h - 1)} \right\}. \quad (28)$$

Then, conditionally on the pre-audit information and counts,

$$\mathbb{P}\{\forall T \in \mathcal{C} : Z_T \leq \mathfrak{R}_{\text{EB}}(T) \mid \text{pre-audit information, counts}\} \geq 1 - \alpha_a. \quad (29)$$

In particular, this certificate requires neither pre-audit residual predictors (Assumption 3) nor the separate calibration budget α_v .

Proof. Condition on the pre-audit information and counts, so that the $\hat{\pi}_h$ are fixed and, for each h and e , the values $G_e(Q_{h,i})$ are iid. They lie almost surely in the interval between the essential infimum and supremum of G_e under P_h , whose length is $\text{osc}_h(G_e) \leq \bar{b}_{h,e}$. The endpoints need not be known, since the bound of Lemma 2 depends on the interval only through its length

and is nondecreasing in it. For each level ℓ , edge $e \in \mathcal{E}_\ell$ and stratum h , apply Lemma 2 with $\delta = \alpha_\ell/(H|\mathcal{E}_\ell|)$, so that $\log(2/\delta) = t_\ell^{\text{EB}}$. This yields

$$\mathbb{E}_h G_e - \bar{G}_{h,e} \leq \sqrt{\frac{2s_{h,e}^2 t_\ell^{\text{EB}}}{n_h}} + \frac{7\bar{b}_{h,e} t_\ell^{\text{EB}}}{3(n_h - 1)}$$

outside an event whose conditional probability does not exceed $\alpha_\ell/(H|\mathcal{E}_\ell|)$. A union bound over the $H|\mathcal{E}_\ell|$ stratum-edge pairs at level ℓ , and then over levels, gives a single event with probability at least $1 - \alpha_a$. For $e = (U, V)$ we have

$$Z_U - Z_V = \sum_h \hat{\pi}_h (\mathbb{E}_h G_e - \bar{G}_{h,e}).$$

On the event just described, multiply each stratum-specific bound by $\hat{\pi}_h \geq 0$, sum over strata, and telescope along every candidate path as in (16). The result is (29). $\square$

Remark 7 (No unconditional dominance). The single-stage radius (28) dispenses with α_v , but it sums stratumwise deviations and pays the larger empirical-Bernstein linear constant $7/3$. The calibrated Bernstein radius of Theorem 2 instead pools the variances before applying the square root, but it pays the predictor bias and the split of α_v . Neither is uniformly smaller. Both should be computed as baselines under comparable confidence and annotation budgets. If the final reported certificate is selected as the *minimum* of two post-audit certificates, their failure probabilities must first be jointly budgeted (or the choice between them must have been fixed without audit labels); taking their minimum at their individual nominal failure levels does not generally preserve those levels.

6 Unknown stratum probabilities and certified regret

Fix stratum 0 as a reference and write, for any $T \in \mathcal{C}$,

$$\begin{aligned} \Delta_T - \hat{\Delta}_T &= Z_T + \sum_h (\pi_h - \hat{\pi}_h) \mu_{h,T} \\ &= Z_T + \sum_{h=1}^{H-1} (\pi_h - \hat{\pi}_h) (\mu_{h,T} - \mu_{0,T}). \end{aligned} \tag{30}$$

The second identity exploits $\sum_h (\pi_h - \hat{\pi}_h) = 0$ and is useful: the cost of estimating the population weights is small when the *contrasts*, not necessarily the coverage levels, are similar between strata.

Let $J = |\mathcal{C}| - 1$ be the number of nontrivial competitors, and suppose $J \geq 1$ (the case $J = 0$ is handled below). Let

$$e_\pi = \sqrt{\frac{\log(2(H-1)/\alpha_\pi)}{2N_u}}. \tag{31}$$

For each h, T denote by $s_{h,T}^2$ the sample variance of the $D_T(Q_{h,i})$, and define

$$t_g = \log \frac{4HJ}{\alpha_g}, \quad r_{h,T} = \sqrt{\frac{2s_{h,T}^2 t_g}{n_h}} + \frac{14t_g}{3(n_h - 1)}, \tag{32}$$

and, for $h = 1, \dots, H - 1$, the observable factor

$$\rho_{h,T} = \min \{2, |\overline{D}_{h,T} - \overline{D}_{0,T}| + r_{h,T} + r_{0,T}\}. \quad (33)$$

The level $t_g = \log(4HJ/\alpha_g)$ is calibrated exactly: each of the $2HJ$ one-sided statements used below is applied with $\delta = \alpha_g/(2HJ)$, so that $\log(2/\delta) = t_g$. No independence between these events and the chaining event is assumed. For $T = \widehat{S}$ one has $\mu_{h,\widehat{S}} = \overline{D}_{h,\widehat{S}} = 0$ deterministically, so no probability is spent on $\widehat{S}$.

Lemma 4 (Uncertainty in unknown context proportions). *Under Assumptions 1 and 2, with probability at least $1 - \alpha_\pi - \alpha_g$, simultaneously for all $T \in \mathcal{C}$,*

$$\left| \sum_h (\pi_h - \widehat{\pi}_h) \mu_{h,T} \right| \leq e_\pi \sum_{h=1}^{H-1} \rho_{h,T}. \quad (34)$$

Proof. For each $h = 1, \dots, H - 1$, Hoeffding's inequality, applied to the Bernoulli indicators $\mathbf{1}\{\phi(Q_i^{(u)}) = h\}$, gives

$$\mathbb{P}\{|\widehat{\pi}_h - \pi_h| > e_\pi\} \leq 2 \exp(-2N_u e_\pi^2) = \alpha_\pi/(H - 1).$$

Union over the $H - 1$ nonreference strata yields the simultaneous proportion event with probability at least $1 - \alpha_\pi$.

Every D_T takes values in $[-1, 1]$. Apply Lemma 2 in both directions, with range 2 and $\delta = \alpha_g/(2HJ)$ for each tail, and take a union bound over the $2HJ$ one-sided statements for the J nontrivial competitors. This gives $|\mu_{h,T} - \overline{D}_{h,T}| \leq r_{h,T}$ for every h and every $T \in \mathcal{C} \setminus \{\widehat{S}\}$, except with probability at most α_g ; for $T = \widehat{S}$ the inequality is trivial. Now use (30), bound $|\pi_h - \widehat{\pi}_h| \leq e_\pi$, and insert

$$|\mu_{h,T} - \mu_{0,T}| \leq \min \{2, |\overline{D}_{h,T} - \overline{D}_{0,T}| + r_{h,T} + r_{0,T}\} = \rho_{h,T},$$

where the bound 2 holds deterministically because $\mu_{h,T} \in [-1, 1]$. Intersection of the two events and a union bound prove the assertion. $\square$

Remark 8 (Where the cardinality of $\mathcal{C}$ re-enters). Because $\rho_{h,T} \leq 2$, the weight correction never exceeds $2(H - 1)e_\pi$. This crude level holds on the proportion event alone, does not depend on $|\mathcal{C}|$, and requires no mean intervals: if one sets $\rho_{h,T} \equiv 2$, Lemma 4 holds with probability at least $1 - \alpha_\pi$ and $\alpha_g = 0$ may be used in Theorem 4. The empirical refinement through $r_{h,T}$, in contrast, relies on a flat union bound over the J competitors, so a factor $\log J$ re-enters the certificate through t_g , multiplied by e_π . The localization of Theorem 1 therefore concerns the audit fluctuation Z_T ; the weight-estimation term is controlled only through the factor e_π , of order $N_u^{-1/2}$, which is small when the unlabelled sample is large.

Define for every $T \in \mathcal{C} \setminus \{\widehat{S}\}$

$$U_T = \widehat{\Delta}_T + \mathfrak{R}(T) + e_\pi \sum_{h=1}^{H-1} \rho_{h,T}, \quad (35)$$

and set $U_{\widehat{S}} := 0$. The latter convention does not follow from (35), whose terms need not vanish at $T = \widehat{S}$ (the path of $\widehat{S}$ in the network may pass through other nodes, and $\rho_{h,\widehat{S}} > 0$); it is legitimate because $\Delta_{\widehat{S}} = 0$ deterministically, so no confidence statement is required for $\widehat{S}$. Let

$$\overline{R}_{\text{red}} = \min \left\{ 1, \max \left\{ 0, \max_{T \in \mathcal{C}} U_T \right\} \right\}. \quad (36)$$

Theorem 4 (Honest postselection regret certificate). *Let Assumptions 1, 2, and 3 hold. Suppose the pilot screening event (5) has probability at least $1 - \alpha_s$. Construct the pre-audit networks as in Definition 1, use the observable edge radius from Theorem 2, and form (35)–(36). If*

$$\alpha_s + \alpha_a + \alpha_v + \alpha_\pi + \alpha_g \leq \alpha < 1,$$

then

$$\mathbb{P}\{R_B(\hat{S}) \leq \bar{R}_{\text{red}}\} \geq 1 - \alpha. \quad (37)$$

If $\mathcal{C} = \{\hat{S}\}$, define $\bar{R}_{\text{red}} = 0$; the statement remains valid. The distribution P and the proportions π_h need not be known or parametrically specified.

Proof. By Theorem 2 and Lemma 4, with probability at least $1 - \alpha_a - \alpha_v - \alpha_\pi - \alpha_g$, we have for every $T \in \mathcal{C} \setminus \{\hat{S}\}$,

$$\Delta_T = \hat{\Delta}_T + Z_T + \sum_h (\pi_h - \hat{\pi}_h) \mu_{h,T} \leq U_T,$$

while $\Delta_{\hat{S}} = 0 = U_{\hat{S}}$ holds deterministically. This simultaneous event is valid even though $\hat{S}, \mathcal{C}$ and the network were learned from the pilot, since the relevant concentration statements hold conditionally on the pre-audit information and counts. On the screening event, some population maximizer $S^* \in \mathcal{C}$. Consequently,

$$R_B(\hat{S}) = F(S^*) - F(\hat{S}) = \Delta_{S^*} \leq U_{S^*} \leq \max_{T \in \mathcal{C}} U_T.$$

Because $0 \leq R_B(\hat{S}) \leq 1$, truncation at 0 and 1 is harmless. The union bound with the screening failure probability proves the result. If $\mathcal{C} = \{\hat{S}\}$, the screening event implies that $\hat{S}$ itself is optimal, hence its regret is zero; off the screening event no assertion is made. $\square$

Corollary 1 (Postselection certification without variance calibration). *Let Assumptions 1 and 2 hold, let the pilot screening event (5) has probability at least $1 - \alpha_s$, and let the pre-audit networks be constructed as in Definition 1. For $T \in \mathcal{C} \setminus \{\hat{S}\}$ put*

$$U_T^{\text{EB}} = \hat{\Delta}_T + \mathfrak{R}_{\text{EB}}(T) + e_\pi \sum_{h=1}^{H-1} \rho_{h,T},$$

with $\mathfrak{R}_{\text{EB}}$ from (28), and set $U_{\hat{S}}^{\text{EB}} := 0$, which is legitimate because $\Delta_{\hat{S}} = 0$. Define $\bar{R}_{\text{EB}} = \min\{1, \max(0, \max_{T \in \mathcal{C}} U_T^{\text{EB}})\}$, with the same singleton convention as in Theorem 4. If $\alpha_s + \alpha_a + \alpha_\pi + \alpha_g \leq \alpha < 1$, then

$$\mathbb{P}\{R_B(\hat{S}) \leq \bar{R}_{\text{EB}}\} \geq 1 - \alpha.$$

No assumption on pre-audit residual predictors is needed.

Proof. Intersect the simultaneous event of Theorem 3 with the proportion and mean calibration events of Lemma 4, which uses only Assumptions 1 and 2, and with the screening event. On this intersection, identity (30) gives $\Delta_T \leq U_T^{\text{EB}}$ for every $T \in \mathcal{C}$; then argue over a screened population optimum exactly as in the proof of Theorem 4. The total failure probability is bounded by $\alpha_s + \alpha_a + \alpha_\pi + \alpha_g$. $\square$

Corollary 2 (Reduced universe versus original corpus). *Suppose the original corpus contains a larger finite feasible family $\mathcal{H}_B^{\text{full}} \supseteq \mathcal{H}_B$ with optimal expected coverage $V_B^{\text{full}} = \max_{S \in \mathcal{H}_B^{\text{full}}} F(S)$. If, with probability at least $1 - \alpha_{\text{rep}}$,*

$$0 \leq V_B^{\text{full}} - \max_{S \in \mathcal{H}_B} F(S) \leq \bar{\rho}_{\text{rep}},$$

then

$$\mathbb{P} \left\{ V_B^{\text{full}} - F(\hat{S}) \leq \bar{\rho}_{\text{rep}} + \bar{R}_{\text{red}} \right\} \geq 1 - \alpha - \alpha_{\text{rep}}.$$

Proof. Decompose $V_B^{\text{full}} - F(\hat{S}) = [V_B^{\text{full}} - \max_{S \in \mathcal{H}_B} F(S)] + R_B(\hat{S})$ and intersect the two coverage events. $\square$

Corollary 3 (Uniformly bounded relevance errors). *Suppose the audit measures functions $\tilde{r}_j : \mathcal{Q} \rightarrow [0, 1]$ satisfying the deterministic uniform calibration condition $\sup_{q,j} |\tilde{r}_j(q) - r_j(q)| \leq \eta$, and all hypotheses of Theorem 4 hold for the induced measured coverages $\tilde{h}_S$ on the same pre-audit family $\mathcal{C}$. Let $\tilde{U}_T$ be their simultaneous contrast bounds and assume the true-population screening event (5) has probability at least $1 - \alpha_s$ (Remark 9 gives a sufficient rule). Then the true reduced regret is bounded by*

$$R_B(\hat{S}) \leq \min\{1, \max(0, \max_{T \in \mathcal{C}} \tilde{U}_T + 2\eta)\}$$

with the same joint confidence level as in Theorem 4.

Proof. The map $(x_j)_{j \in S} \mapsto \max_{j \in S} x_j$ is 1-Lipschitz under the supremum norm. Thus $|h_S(q) - \tilde{h}_S(q)| \leq \eta$ for every S and q , so $|\Delta_T - \tilde{\Delta}_T| \leq 2\eta$ for all T . On the simultaneous measured-contrast event and the true screening event, $R_B(\hat{S}) = \Delta_{S^*} \leq \tilde{U}_{S^*} + 2\eta$. Truncation to $[0, 1]$ completes the proof. $\square$

Remark 9 (Screening with measured pilot labels). Corollary 3 requires the screening event for the true population. If the pilot of Proposition 2 is itself labeled with the measured functions $\tilde{r}_j$, its uniform Hoeffding event controls $|\hat{F}_0(S) - \tilde{F}(S)|$, where $\tilde{F}(S) = \mathbb{E}_P \tilde{h}_S(Q)$ and $|\tilde{F}(S) - F(S)| \leq \eta$. On that event, $\hat{F}_0(S^*) - \hat{F}_0(\hat{S}) + 2r_0 \geq F(S^*) - F(\hat{S}) - 2\eta \geq -2\eta$. Hence the rule (7) must be widened to $\hat{F}_0(T) - \hat{F}_0(\hat{S}) + 2r_0 + 2\eta \geq 0$; with this change the true-population screening event again has probability at least $1 - \alpha_s$.

7 A complementary pooled stratified Bennett certificate

Before using the full multiscale construction, one may combine the conditional stratum fluctuations using a single exponential tilt. The result is included for completeness and as a benchmark, not claimed to be an original concentration inequality.

For each candidate $T \in \mathcal{C}$, denote $\sigma_{h,T}^2 = \text{Var}(D_T(Q) \mid X = h)$. In the formulas below take $J = |\mathcal{C}| - 1 \geq 1$. For every h and every $T \in \mathcal{C} \setminus \{\hat{S}\}$, use a pre-audit predictor $m_{h,T} \in [-1, 1]$ satisfying the analogue of Assumption 3, and replace G_e by D_T in the construction of Lemma 3. The union bound now runs over the HJ pairs (h, T) , so t_v is replaced by $t'_v = \log(2HJ/\alpha'_v)$. This gives simultaneous empirical upper bounds $v_{h,T}^+ \geq \sigma_{h,T}^2$ with failure probability at most α'_v . The budget α'_v is distinct from the budget α_v of Lemma 3; both must be paid if both certificates are reported. Set

$$w_h = \hat{\pi}_h, \quad \beta_h = \frac{2w_h}{n_h}, \quad \mathcal{K}_T^+(\lambda) = \sum_{h \in \mathcal{S}} \frac{n_h v_{h,T}^+}{4} \left(e^{\lambda \beta_h} - 1 - \lambda \beta_h \right). \quad (38)$$

If $w_h = 0$, the corresponding summand is defined to be zero.

Proposition 4 (Pooled Bennett bound with certified residual moments). *Assume the model of Section 2 and let every $\lambda_T > 0$ be fixed conditionally on the pre-audit information and the audit stratum counts, without using audit relevance values. Then, with probability at least $1 - \alpha_B - \alpha'_v$, simultaneously for all $T \in \mathcal{C} \setminus \{\hat{S}\}$,*

$$Z_T \leq \mathcal{B}_T := \frac{\log(J/\alpha_B) + \mathcal{K}_T^+(\lambda_T)}{\lambda_T}. \quad (39)$$

Combining this with Lemma 4 and pilot screening provides an alternative population-regret certificate

$$\bar{R}_{\text{Ben}} = \min \left\{ 1, \max \left\{ 0, \max_{T \in \mathcal{C}} \left[\hat{\Delta}_T + \mathcal{B}_T + e_\pi \sum_{h=1}^{H-1} \rho_{h,T} \right] \right\} \right\}, \quad (40)$$

with coverage at least $1 - \alpha_s - \alpha_B - \alpha'_v - \alpha_\pi - \alpha_g$. The bracket for $T = \hat{S}$ is replaced by 0, which is legitimate because $\Delta_{\hat{S}} = 0$; if $J = 0$, the certificate is zero on the screening event.

Proof. Condition on the pilot, the independent unlabelled sample and the stratum counts. For each T , define the centered summands

$$Y_{h,i,T} = \frac{w_h}{n_h} (\mu_{h,T} - D_T(Q_{h,i})).$$

Their sum equals Z_T , their variances are $(w_h/n_h)^2 \sigma_{h,T}^2$, and, since $\mu_{h,T} - D_T \leq 2$, they are almost surely at most β_h . Bennett's moment-generating-function bound (Boucheron et al., 2013, Theorem 2.9), applied with $n = 1$ to a single centered random variable $Y \leq \beta$ (so that $\mathbb{E}Y^2$ is its variance v), gives

$$\log \mathbb{E} e^{\lambda Y} \leq \frac{v}{\beta^2} (e^{\lambda \beta} - 1 - \lambda \beta) \quad (\beta > 0, \lambda > 0).$$

Summing conditional cumulant bounds across the independent queries, we obtain

$$\log \mathbb{E}[e^{\lambda Z_T} \mid \text{pre-audit information, counts}] \leq \sum_h \frac{n_h \sigma_{h,T}^2}{4} (e^{\lambda \beta_h} - 1 - \lambda \beta_h).$$

When $\beta_h = 0$ the variables vanish, so their contribution is zero. Applying Markov's inequality with the fixed λ_T to the event

$$Z_T > \frac{\log(J/\alpha_B) + \sum_h \frac{n_h \sigma_{h,T}^2}{4} (e^{\lambda_T \beta_h} - 1 - \lambda_T \beta_h)}{\lambda_T}.$$

The conditional probability is at most α_B/J . A union bound over the J nontrivial competitors yields an event of probability at least $1 - \alpha_B$ on which all those inequalities hold. Intersect this event with the simultaneous residual-moment event of probability at least $1 - \alpha'_v$, and replace $\sigma_{h,T}^2$ by $v_{h,T}^+$ using the monotonicity of the exponential cumulant for $\lambda_T > 0$. This gives (39). The decomposition (30), Lemma 4, and the screening argument used in Theorem 4 complete the proof of (40) and its stated coverage probability. $\square$

Remark 10 (Data-dependent exponential tilting). Optimizing λ_T after seeing the audit labels is not justified by Proposition 4: a fixed- λ Markov inequality is not automatically uniform in λ . A grid with an explicitly allocated union bound or a time-uniform mixture construction would be needed for such an optimization. The same caution applies to choosing a multiscale network on the audit labels without additional sample splitting or a uniform complexity argument.

8 Documentary activity and structural contraction

The concentration theory above is valid for arbitrary bounded coverage functions. We now exploit the particular maximum structure in (2). For a nonempty set R , write

$$a_R(q) = \max_{g \in R} r_g(q), \quad a_\emptyset(q) = 0.$$

Lemma 5 (Single-fragment swap and activity). *Let $j, k \notin R$, let $S = R \cup \{j\}$, $T = R \cup \{k\}$, and suppose $\sup_{q \in \mathcal{Q}} |r_j(q) - r_k(q)| \leq \varepsilon$. Define the activity event*

$$A_{j,k,R}(q) = \mathbf{1}\{\max(r_j(q), r_k(q)) > a_R(q)\}, \quad (41)$$

and $a_h = \mathbb{P}\{A_{j,k,R}(Q) = 1 \mid X = h\}$. Then for every q ,

$$|h_S(q) - h_T(q)| \leq \varepsilon A_{j,k,R}(q), \quad (42)$$

and, for every stratum h ,

$$\text{Var}(h_S(Q) - h_T(Q) \mid X = h) \leq \varepsilon^2 a_h, \quad \text{osc}_h(h_S - h_T) \leq 2\varepsilon. \quad (43)$$

Consequently the edge $e = (S, T)$ obeys

$$d_2^2(e) \leq \varepsilon^2 \sum_h \frac{\hat{\pi}_h^2}{n_h} a_h, \quad d_\infty(e) \leq 2\varepsilon \max_h \frac{\hat{\pi}_h}{n_h}. \quad (44)$$

Proof. For every $a \geq 0$, the scalar map $u \mapsto \max(a, u)$ is 1-Lipschitz. Therefore

$$|h_S(q) - h_T(q)| = |\max(a_R(q), r_j(q)) - \max(a_R(q), r_k(q))| \leq |r_j(q) - r_k(q)| \leq \varepsilon.$$

When $A_{j,k,R}(q) = 0$, both $r_j(q)$ and $r_k(q)$ are at most $a_R(q)$, and the two coverages are identical. Combining the observations proves (42). Squaring and taking conditional expectations gives

$$\text{Var}_h(h_S - h_T) \leq \mathbb{E}_h(h_S - h_T)^2 \leq \varepsilon^2 \mathbb{E}_h A_{j,k,R} = \varepsilon^2 a_h.$$

The range inequality follows because the difference is in $[-\varepsilon, \varepsilon]$. Substitute these bounds into (11)–(12) to obtain (44). $\square$

Proposition 5 (Rare-query obstruction to sample-only uniform envelopes). *For every finite pilot size $n_0 \geq 1$ and every $\delta \in (0, 1)$, there exist a two-point query distribution and two relevance functions with values in $[0, 1]$ such that, with probability exceeding $1 - \delta$, the pilot observes identical relevance scores for both fragments on all n_0 queries, although $\sup_q |r_j(q) - r_k(q)| = 1$. Consequently, an empirical supremum of observed relevance differences, without additional structural assumptions, is not a valid distribution-free upper confidence bound for the population uniform envelope.*

Proof. Let $\mathcal{Q} = \{0, 1\}$ with $P(Q = 1) = \theta > 0$. Set $r_j(0) = r_k(0) = 0$, $r_j(1) = 1$, and $r_k(1) = 0$. The true uniform difference equals 1. Yet with probability $(1 - \theta)^{n_0}$ the entire iid pilot consists of queries equal to 0, on which the scores agree exactly. Choosing $0 < \theta < 1 - (1 - \delta)^{1/n_0}$ makes this probability greater than $1 - \delta$. Thus a pilot-only empirical supremum of zero is compatible with a population uniform difference of one, and it covers the true envelope with probability less than δ . More generally, any procedure whose output on this all-zero pilot sample is some value $c < 1$ covers the true envelope 1 with probability at most $1 - (1 - \theta)^{n_0} < \delta$ under this distribution. $\square$

Proposition 6 (Simultaneously estimated activity factors). *Let $\mathcal{W}$ be a finite set of pre-audit swap triples (j, k, R) with known deterministic envelopes $\varepsilon_{j,k,R}$ as in Lemma 5, and let $W = |\mathcal{W}| \geq 1$. Assume that, for every acquired audit query $Q_{h,i}$ and every $w = (j, k, R) \in \mathcal{W}$, the relevance values $r_j(Q_{h,i})$, $r_k(Q_{h,i})$ and $r_g(Q_{h,i})$, $g \in R$, are observed, so that the indicators $A_w(Q_{h,i})$ are observed. This does not follow from Assumption 2, which guarantees only the coverages $h_T(Q_{h,i})$, $T \in \mathcal{C}$: these determine neither A_w when $h_S = h_T$, nor any quantity involving fragments outside the retained competitor family. For each h and triple w , define $\hat{a}_{h,w} = n_h^{-1} \sum_i A_w(Q_{h,i})$ and*

$$a_{h,w}^+ = \min \left\{ 1, \hat{a}_{h,w} + \sqrt{\frac{\log(HW/\alpha_{\text{act}})}{2n_h}} \right\}.$$

Then with probability at least $1 - \alpha_{\text{act}}$, simultaneously for all h, w , $a_{h,w} \leq a_{h,w}^+$. In particular,

$$\bar{v}_{(S,T)}^{\text{act}} = \varepsilon_{j,k,R}^2 \sum_h \frac{\hat{\pi}_h^2}{n_h} a_{h,w}^+$$

is a simultaneous variance bound for all such swap edges. It can be combined by taking the minimum with the residual bound (26), at the additional error cost α_{act} .

Proof. For each h, w , the random variables $A_w(Q_{h,i})$ are iid Bernoulli with conditional mean $a_{h,w}$. The one-sided Hoeffding inequality gives failure probability at most $\alpha_{\text{act}}/(HW)$ for the displayed upper bound. Union over all HW pairs, followed by Lemma 5, proves the variance assertion. To take the minimum with the residual upper bounds, intersect their two simultaneous validity events and use another union bound. $\square$

Proposition 7 (Pathwise extension to several swaps). *Suppose two selections S, T can be connected by a finite path $S = S_0, S_1, \dots, S_m = T$ in which each consecutive pair differs by a single fragment swap covered by Lemma 5, with parameters $\varepsilon_k, a_{h,k}$. Then for each h ,*

$$\sqrt{\mathbb{E}[(h_S - h_T)^2 \mid X = h]} \leq \sum_{k=1}^m \varepsilon_k \sqrt{a_{h,k}}. \quad (45)$$

In particular,

$$\text{Var}(h_S - h_T \mid X = h) \leq \left(\sum_{k=1}^m \varepsilon_k \sqrt{a_{h,k}} \right)^2.$$

Proof. Pointwise,

$$h_S - h_T = \sum_{k=1}^m (h_{S_{k-1}} - h_{S_k}).$$

Apply Minkowski's inequality in $L_2(P_h)$, then Lemma 5 to each summand. Variance is at most the uncentered second moment. $\square$

9 A two-level sufficient condition for a sharper bound

The multiscale construction is not automatically better than a direct union bound, because it splits the error probability across levels. Here is a sufficient condition for improvement. In this section condition on the pre-audit information and counts, and consider only the audit fluctuation Z_T , not the common error from estimating population stratum proportions.

Corollary 4 (Two-level improvement over a flat Bernstein union bound). *Suppose the $J = |\mathcal{C}| - 1 \geq 2$ nontrivial competitors admit K_0 distinct representatives, $1 \leq K_0 < J$, and a map $c : \mathcal{C} \setminus \{\hat{S}\} \rightarrow \{1, \dots, K_0\}$ such that every candidate $T \neq \hat{S}$ is associated with a representative $V_{c(T)}$ and*

$$\sup_{q \in \mathcal{Q}} |h_T(q) - h_{V_{c(T)}}(q)| \leq \varepsilon, \quad 0 \leq \varepsilon \leq 1. \quad (46)$$

Assume each representative belongs to $\mathcal{C}$, and use the two-level network $\hat{S} \rightarrow V_{c(T)} \rightarrow T$ for $T \neq \hat{S}$, while the selected set follows the constant path $\hat{S} \rightarrow \hat{S} \rightarrow \hat{S}$. The first-level network includes $\hat{S}$ and all representatives. Define

$$q_n = \sum_h \frac{\hat{\pi}_h^2}{n_h}, \quad b_n = 2 \max_h \frac{\hat{\pi}_h}{n_h}, \quad L_n(t) = \sqrt{2q_n t} + \frac{1}{3} b_n t. \quad (47)$$

For $\alpha_1 + \alpha_2 = \alpha_a$, the uniform two-level radius is

$$C_{\text{chain}} = L_n \left(\log \frac{K_0}{\alpha_1} \right) + \varepsilon L_n \left(\log \frac{J}{\alpha_2} \right). \quad (48)$$

The flat cardinality-based direct Bernstein radius is

$$C_{\text{flat}} = L_n \left(\log \frac{J}{\alpha_a} \right). \quad (49)$$

Whenever

$$0 \leq \varepsilon < \frac{L_n(\log(J/\alpha_a)) - L_n(\log(K_0/\alpha_1))}{L_n(\log(J/\alpha_2))}, \quad (50)$$

with positive numerator, one has $C_{\text{chain}} < C_{\text{flat}}$. Both bounds apply simultaneously to all candidates with conditional failure probability at most α_a .

Proof. For a coarse edge $\hat{S} \rightarrow V_k$, its coverage difference lies in $[-1, 1]$, so its variance is at most 1 and its range at most 2 in each stratum. Lemma 1 therefore bounds the coarse fluctuation by $L_n(\log(K_0/\alpha_1))$ simultaneously over the K_0 distinct edges. For a fine edge $V_{c(T)} \rightarrow T$, (46) implies that the difference is in $[-\varepsilon, \varepsilon]$, with variance at most ε^2 and range at most 2ε in every stratum. Its Bernstein radius is therefore at most $\varepsilon L_n(\log(J/\alpha_2))$, simultaneously over at most J fine edges (at most $J - K_0$ when the representatives are themselves nontrivial competitors, since their fine edges are trivial; the cruder count J is kept for simplicity). The telescoping identity yields (48). Because all variance and range bounds used here are deterministic, no calibration budget α_v is needed. Direct union over the J edges from $\hat{S}$ to each candidate yields (49). Subtracting the two deterministic expressions, and dividing by the strictly positive term $L_n(\log(J/\alpha_2))$, gives the sufficient condition (50). $\square$

Remark 11 (Precise scope of the comparison). Corollary 4 compares two conservative Bernstein radii built from variance and range bounds, under common nominal confidence and the same audit design.

Numerical illustration (exact arithmetic on the formulas). Take $J = 5000$, $K_0 = 5$, $H = 2$, an audit count of $n_h = 500$ in each of the two strata, and $\hat{\pi}_0 = \hat{\pi}_1 = 1/2$. Choose $\alpha_a = 0.05$, $\alpha_1 = \alpha_2 = 0.025$, and $\varepsilon = 0.05$. Then $q_n = 0.001$ and $b_n = 0.002$, so

$$C_{\text{flat}} \approx 0.15942, \quad C_{\text{chain}} \approx 0.11469, \quad \frac{C_{\text{flat}} - C_{\text{chain}}}{C_{\text{flat}}} \approx 0.2806. \quad (51)$$

The threshold on ε in (50) is about 0.3221. This is a deterministic illustration of the algebraic sufficient condition, not a simulation and not an experimentally measured gain in total documentary regret.

10 Simulation study

This section examines three questions: whether the certificates attain their nominal level in finite samples; how the observable radii of Sections 4–7 compare in size; and when the two-level network of Corollary 4 improves on a flat union bound.

10.1 Design

Query population and fragments. There are $H = 3$ observable strata, with probabilities $\pi = (0.5, 0.3, 0.2)$. Given $X = h$, a latent topic $Z \in \{1, \dots, 6\}$ is drawn with probabilities θ_h ,

$$\begin{aligned}\theta_0 &= (.35, .25, .15, .10, .10, .05), \\ \theta_1 &= (.05, .10, .15, .20, .20, .30), \\ \theta_2 &= (.10, .15, .30, .25, .10, .10).\end{aligned}$$

Each topic c has a base relevance $\tilde{r}_c(q) = 0.75 \mathbf{1}\{Z = c\} + 0.25 W_c$, with $W_c \sim \text{Beta}(1, 3)$ independent. The corpus has $K = 30$ fragments organized in six clusters of five. The first fragment of cluster c has relevance $\tilde{r}_c$; each other fragment j of the cluster has $r_j = \min\{1, \max\{0, \tilde{r}_c + \varepsilon U_j\}\}$, with $U_j \sim \text{U}[-1, 1]$ independent. Hence $|r_j - \tilde{r}_c| \leq \varepsilon$ deterministically: ε controls the functional redundancy within clusters, and $\varepsilon = 0$ gives exact duplicates.

Selection, competitors and samples. All costs equal 1 and $B = 3$, so $\mathcal{H}_B$ consists of the 4526 subsets of size at most 3. We take $\mathcal{C} = \mathcal{H}_B$, so that $\alpha_s = 0$ and $J = 4525$. In each replication, a fully labelled pilot of $n_0 = 1000$ iid queries selects $\hat{S}$ by exhaustive maximization of $\hat{F}_0$, supplies the residual predictors $m_{h,e}$ (pilot stratum means of G_e) and fixes the Bennett tilts; an independent unlabelled sample of $N_u = 50\,000$ queries gives $\hat{\pi}_h$; and the audit draws n_h queries from each stratum. The population quantities $\mu_{h,T}$, $F(T)$ and $R_B(\hat{S})$ are computed from an independent Monte Carlo sample of 150 000 queries per stratum, with standard errors below 1.3×10^{-3} .

Networks. For $T \in \mathcal{H}_B$ let $V(T)$ be the set of first fragments of the clusters met by T . Then $V(T) \in \mathcal{H}_B$, and the 1-Lipschitz property of the maximum, used as in the proof of Corollary 3, gives $\sup_q |h_T(q) - h_{V(T)}(q)| \leq \varepsilon$. The two-level network is $\hat{S} \rightarrow V(T) \rightarrow T$, with 42 coarse and 4483 fine edges in every replication; the flat network joins every competitor directly to $\hat{S}$. In this design the envelope ε is known by construction; in applications it would have to be certified externally (Proposition 5).

Certificates and error budgets. All certificates have total level $\alpha = 0.05$. The weight correction of Lemma 4 uses $\alpha_\pi = \alpha_g = 0.005$ and is common to all of them; the remaining 0.04 is allocated to the audit fluctuation. We compare seven certificates of the form $\bar{R} = \min\{1, \max(0, \max_T U_T)\}$:

- (i) *Hoeffding, flat*: a union of one-sided Hoeffding bounds over the HJ pairs (h, T) , which gives $Z_T \leq \sum_h \hat{\pi}_h 2\{\log(HJ/0.04)/(2n_h)\}^{1/2}$;

Table 2: Mean regret certificate $\bar{R}$ over 200 replications, $\varepsilon = 0.05$, as a function of the number n_h of audited queries per stratum. Standard errors are below 0.003 in all cells. The last rows give the mean true regret $R_B(\hat{S})$ and the mean largest weight-correction term $\max_T e_\pi \sum_{h \geq 1} \rho_{h,T}$, common to all certificates.

Certificate	$n_h = 100$	$n_h = 200$	$n_h = 400$	$n_h = 800$	$n_h = 1600$
Hoeffding, flat	0.550	0.381	0.265	0.187	0.134
One-stage EB, flat (Thm. 3)	0.930	0.507	0.276	0.156	0.092
One-stage EB, two-level	0.744	0.409	0.224	0.129	0.077
One-stage EB, two-level, $\bar{b}_{h,e} = 2$	1.000	0.723	0.381	0.208	0.117
Calibrated Bernstein, flat (Thm. 2)	0.340	0.186	0.101	0.061	0.037
Calibrated Bernstein, two-level	0.293	0.161	0.088	0.053	0.033
Pooled Bennett (Prop. 4)	0.332	0.182	0.099	0.060	0.037
True regret	0.0037	0.0027	0.0029	0.0035	0.0044
Weight correction	0.033	0.028	0.019	0.014	0.012

(ii)–(iv) *one-stage empirical Bernstein* (Theorem 3, Corollary 1) with $\alpha_a = 0.04$: flat; two-level with $\alpha_1 = \alpha_2 = 0.02$ and the known envelope $\bar{b}_{h,e} = 2\varepsilon$ on fine edges; and two-level with the universal envelope $\bar{b}_{h,e} = 2$;

(v)–(vi) *calibrated Bernstein* (Lemma 3, Theorem 2, Theorem 4) with $\alpha_a = 0.03$ and $\alpha_v = 0.01$: flat, and two-level with $\alpha_1 = \alpha_2 = 0.015$. On fine edges, where $|G_e| \leq \varepsilon$, the residual range is taken as $(\varepsilon + |m_{h,e}|)^2$ and the variance bound is truncated at ε^2 ; both are direct consequences of Lemma 2 and of Popoviciu’s inequality for a variable with range 2ε ;

(vii) *pooled Bennett* (Proposition 4) with $\alpha_B = 0.03$ and $\alpha'_v = 0.01$. Each λ_T minimizes, over a fixed grid, the bound (39) in which $v_{h,T}^+$ is replaced by the same residual formula evaluated on the pilot with the known audit counts; thus λ_T uses no audit label.

We vary $n_h \in \{100, 200, 400, 800, 1600\}$ at $\varepsilon = 0.05$, and $\varepsilon \in \{0, 0.02, 0.05, 0.1, 0.2, 0.4\}$ at $n_h = 400$, with 200 replications per setting.

10.2 Results

Validity. In all 10 settings and for all seven certificates, the certificate covered the true regret in every replication, and the simultaneous audit event $\{Z_T \leq \text{radius}(T) \ \forall T\}$ held in every replication. With 200 replications, the one-sided 95% Clopper–Pearson lower bound for each coverage probability is 0.985. As expected from union bounds with conservative constants, the certificates are valid with room to spare.

Size of the certificates. Table 2 and Figure 1 report the mean certificates at $\varepsilon = 0.05$. The calibrated Bernstein certificate of Theorem 4 and the pooled Bennett certificate are nearly indistinguishable, and both are about 2.5–2.7 times smaller than the one-stage empirical-Bernstein certificate at every n_h . The one-stage certificate pays the linear term $7\bar{b}_{h,e}t/\{3(n_h - 1)\}$ in every stratum; for $n_h \leq 400$ it is even larger than the flat Hoeffding certificate, which it overtakes only for $n_h \geq 800$. In this design, therefore, the separate calibration budget α_v is well spent, in line with Remark 7. All certificates remain far above the true regret, which is about 0.003: at $n_h = 1600$ the smallest certificate is 0.033, of which about 0.012 is the weight-correction term. This gap is the price of a simultaneous, honest guarantee over 4525 data-dependent competitors.

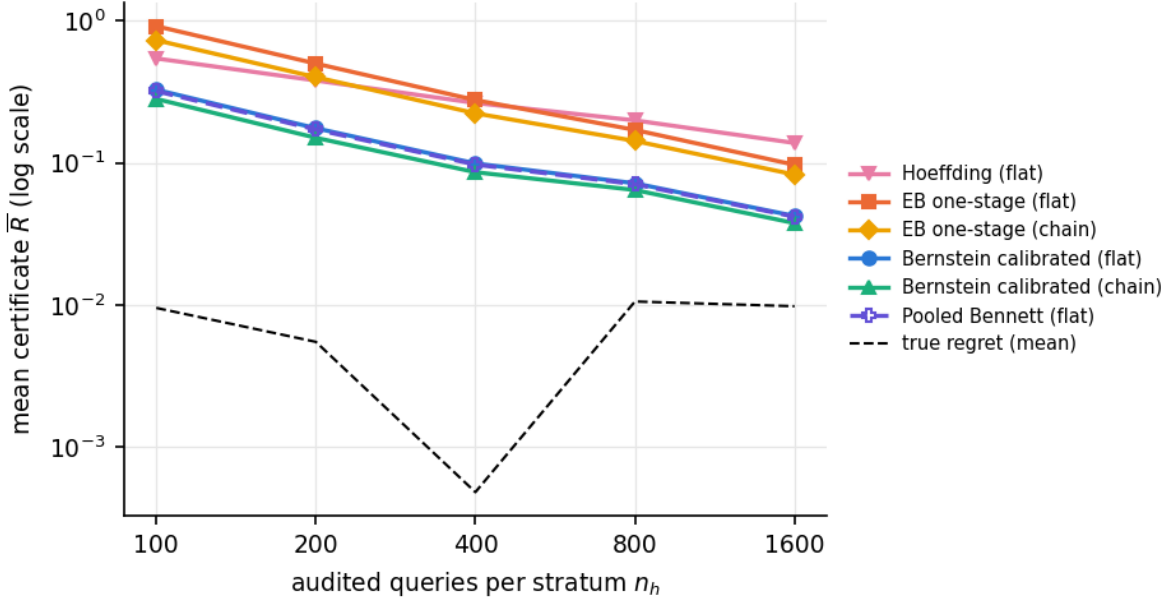


Figure 1: Mean regret certificate $\bar{R}$ (logarithmic scales) over 200 replications at $\varepsilon = 0.05$, as a function of the number of audited queries per stratum. The pooled Bennett certificate (dashed, open marks) nearly coincides with the flat calibrated Bernstein certificate. The dashed black line is the mean true regret.

When does the multiscale network help? Table 3 and Figure 2 vary the redundancy envelope at $n_h = 400$. With a known envelope, the two-level certificate is smaller than the flat one for small ε : by 19% ($\varepsilon = 0$), 13% ($\varepsilon = 0.05$) and 7% ($\varepsilon = 0.1$) for the calibrated Bernstein version, and by 25%, 19% and 13% for the one-stage version. The advantage vanishes between $\varepsilon = 0.1$ and $\varepsilon = 0.2$ for the calibrated version and at about $\varepsilon = 0.2$ for the one-stage version, and it turns into a loss at $\varepsilon = 0.4$. This crossover agrees with the deterministic threshold of Corollary 4, which equals about 0.19 for $K_0 = 42$, $J = 4525$, $n_h = 400$, $\alpha_1 = \alpha_2$ and the present weights. The reduction of the radius at the population optimum can be much larger than the reduction of the certificate (from 0.061 to 0.013 at $\varepsilon = 0$ for the calibrated version), because the certificate is a maximum over all competitors. Finally, without a certified envelope ($\bar{b}_{h,e} = 2$ on fine edges) the two-level certificate is 35–58% *larger* than the flat one: splitting the error budget across levels is then not compensated by any reduction of the linear term. This illustrates why the gains of Section 9 require certified, not merely empirical, envelopes (Proposition 5).

Summary. All certificates are valid. Among the fully observable ones, variance calibration (Theorem 4) or the pooled Bennett tilt (Proposition 4) gives the smallest flat certificates, and the multiscale network yields an additional reduction of up to about 20% when fragments are strongly redundant and the redundancy is certified, in agreement with Corollary 4.

We have specified a complete finite-sample probability argument: pilot screening, conditional concentration of every pre-audit comparison edge, telescoping over finite multiscale networks, simultaneous empirical variance calibration, correction for unknown context proportions, and conversion to postselection regret. The redundancy lemmas give explicit variance contractions

Table 3: Mean regret certificate $\bar{R}$ over 200 replications at $n_h = 400$, as a function of the redundancy envelope ε . Rows marked “ratio” divide the two-level certificate by the corresponding flat one.

Certificate	$\varepsilon = 0$	$\varepsilon = 0.02$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$	$\varepsilon = 0.4$
Calibrated Bernstein, flat	0.101	0.104	0.101	0.104	0.107	0.109
Calibrated Bernstein, two-level	0.082	0.087	0.088	0.097	0.113	0.139
ratio	0.81	0.84	0.87	0.93	1.05	1.27
One-stage EB, flat	0.278	0.280	0.276	0.278	0.280	0.276
One-stage EB, two-level	0.209	0.218	0.224	0.243	0.277	0.337
One-stage EB, two-level, $\bar{b}_{h,e} = 2$	0.374	0.380	0.381	0.391	0.409	0.436
ratio (known envelope)	0.75	0.78	0.81	0.87	0.99	1.22
ratio ($\bar{b}_{h,e} = 2$)	1.35	1.36	1.38	1.41	1.46	1.58
Pooled Bennett, flat	0.099	0.102	0.099	0.102	0.105	0.107

for swapping documentary fragments. The confidence level in Theorem 4 is unconditional: it averages over the pilot, whose only contribution is the screening error α_s , and it requires no model for P . Conditionally on $\mathcal{F}_0$, only the post-screening part of the statement holds, with level $1 - \alpha_a - \alpha_v - \alpha_\pi - \alpha_g$.

The sets $\mathcal{N}_\ell$ may be constructed using an independent pilot, including estimated similarities. Such a construction does not need to be accurate for the base theorem: validity follows from the edgewise simultaneous concentration with conservative envelopes.

References

- Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local Rademacher complexities. *The Annals of Statistics*, 33(4), 1497–1537. [doi:10.1214/009053605000000282](https://doi.org/10.1214/009053605000000282).
- Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press. [doi:10.1093/acprof:oso/9780199535255.001.0001](https://doi.org/10.1093/acprof:oso/9780199535255.001.0001).
- Burgess, M. A. and Chapman, A. C. (2021). Approximating the Shapley value using stratified empirical Bernstein sampling. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pp. 73–81. [doi:10.24963/ijcai.2021/11](https://doi.org/10.24963/ijcai.2021/11).
- Dirksen, S. (2015). Tail bounds via generic chaining. *Electronic Journal of Probability*, 20(53), 1–29. [doi:10.1214/EJP.v20-3760](https://doi.org/10.1214/EJP.v20-3760).
- Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104. https://icml.cc/2011/papers/554_icmlpaper.pdf.
- Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301), 13–30. [doi:10.1080/01621459.1963.10500830](https://doi.org/10.1080/01621459.1963.10500830).
- Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), 1055–1080. [doi:10.1214/20-AOS1991](https://doi.org/10.1214/20-AOS1991).

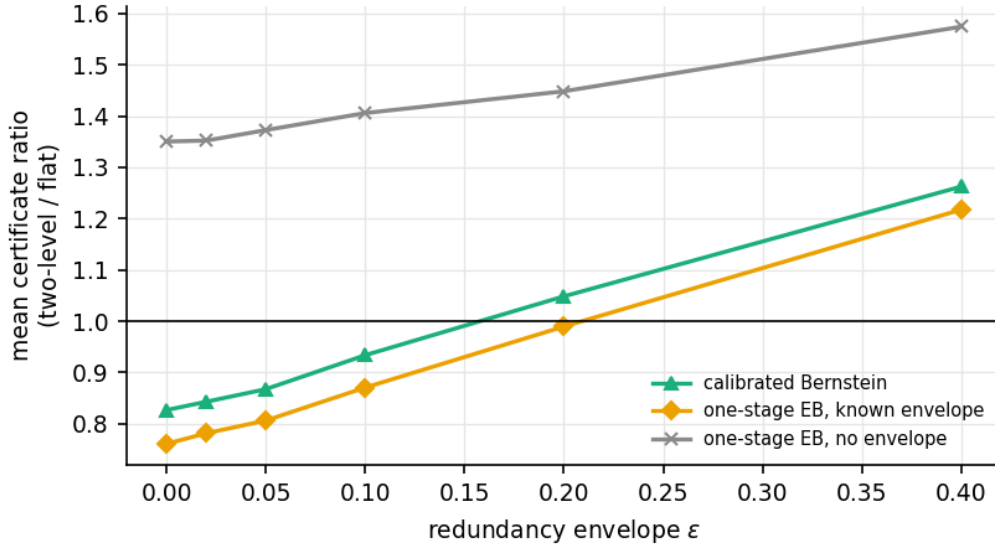


Figure 2: Ratio of the mean two-level certificate to the mean flat certificate at $n_h = 400$, as a function of the redundancy envelope ε ; values below 1 indicate a gain from the multiscale network.

Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6), 2593–2656. doi:10.1214/009053606000001019.

Krause, A. and Golovin, D. (2014). Submodular function maximization. In Bordeaux, L., Hamadi, Y., and Kohli, P. (eds.), *Tractability: Practical Approaches to Hard Problems*, Cambridge University Press, pp. 71–104. doi:10.1017/CBO9781139177801.004.

Maurer, A. and Pontil, M. (2009). Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*. <https://arxiv.org/abs/0907.3740>.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14, 265–294. doi:10.1007/BF01588971.

Talagrand, M. (2005). *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*. Springer. doi:10.1007/3-540-27499-5.