MINIMUM CERTIFIED EXPLORATION FOR CONTEXTUAL BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present an algorithm that explores only as much as a regret guarantee actually requires, for general contextual bandits with a regression oracle. We introduce a one-step regret certificate: a closed-form test of whether the one-step inequality behind the regret analysis holds at the current predictions, so that every certified allocation inherits the full regret guarantee. The certified allocations have a least element, one distribution that minimizes every nongreedy probability at once, and our algorithm MinCE (Minimum Certified Exploration) computes it in a few sweeps of scalar inversions, with no additional oracle call or hyperparameter and with every intermediate allocation certified. MinCE matches the best existing first-order regret bound and the minimax rate while performing minimal exploration. In simulations MinCE attains the lowest regret, and on 130 OpenML classification datasets it achieves the lowest loss on 112 of them, outperforming the baselines.

1 INTRODUCTION

Contextual bandits (Abe & Long, 1999; Auer et al., 2002; Langford & Zhang, 2008; Li et al., 2010) model decisions that must be learned from partial feedback: only the outcome of the chosen action is observed. Learning from such feedback requires exploration, since an action that is never tried is never evaluated. In practice, however, each exploratory action is one that the current model expects to be worse. Practitioners therefore may wish to explore as little as possible. Exploring less is not automatically safe: a learner that explores too little can lock onto a wrong action, and an algorithm that deviates from its prescribed exploration loses its regret guarantee. The practical question is therefore not whether to explore, but how little exploration a regret guarantee actually requires and how to explore exactly that much.

We address this question for general contextual bandits: an arbitrary function class accessed through a regression oracle, and contexts that may be chosen adaptively. This is the setting of the regression-oracle reductions (Foster et al., 2018a; Foster & Rakhlin, 2020; Simchi-Levi & Xu, 2022; Foster & Krishnamurthy, 2021), the algorithms of choice when losses are modeled by flexible learners. These algorithms explore by fixed formulas that map predicted losses to a sampling distribution in the same way at every round, with a margin chosen so that the analysis goes through in the worst case. The exploration is provably sufficient, but it is not known how much of it is necessary, and the exploration parameter cannot tell: raising it reduces exploration but also changes the regret bound, trading one guarantee for another.

Our approach is to make the guarantee itself the constraint. Regret analyses of this kind rest on a one-step inequality that bounds the instantaneous regret of an allocation by its exploration and the prediction error of the oracle. We turn this inequality into a *one-step regret certificate*: a function of the allocation and the predictions whose sign says whether the inequality holds for every possible mean-loss vector. Any certified allocation inherits the full regret guarantee, so the learner may explore as little as the certificate allows. The recipe does not depend on a particular base algorithm, since any one-step inequality defines a certificate; we carry it out for the Bernoulli-KL inequality behind first-order regret bounds (Foster & Krishnamurthy, 2021) and find that its certificate has exactly the structure an algorithm needs. It has a closed form that takes $\mathcal{O}(A)$ time to evaluate for A actions, with no optimization over the function class, and the certified allocations have a *least element*: one distribution that minimizes every nongreedy probability at once. Our proposed method, MinCE, is the algorithm that plays this least element.

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.

Our main contributions are as follows.

1. **A minimum-exploration algorithm** (Section 3). MinCE is built on a new principle for choosing actions: instead of exploring by a fixed formula, it explores exactly as much as the regret guarantee requires at the current predictions. In every round it computes the least certified allocation, the distribution that minimizes every nongreedy probability at once, through a closed-form certificate and a few sweeps of scalar inversions. It needs no optimization over the function class, no additional oracle call, and no additional hyperparameter, and it works with any regression oracle, including neural networks.
2. **Regret matching the best existing bound** (Theorem 2, Corollary 3). MinCE achieves the first-order bound $\mathcal{O}(\sqrt{AL_T^* \operatorname{Reg}_{\mathrm{KL}}(T)} + A \operatorname{Reg}_{\mathrm{KL}}(T))$, where L_T^* is the cumulative loss of the optimal policy and $\operatorname{Reg}_{\mathrm{KL}}(T)$ is the log-loss regret of the oracle, and hence the minimax rate $\mathcal{O}(\sqrt{AT \log M})$ for classes of size M, under adaptively chosen contexts. Minimizing exploration costs nothing in achieving the near-optimal regret bounds.
3. **Minimum certified exploration** (Theorems 4 and 5). At every round the certified allocations have a least element, and MinCE computes it to within a negligible relative error in a few sweeps (e.g., three when $T = 5000$) while every intermediate allocation remains certified. At the same predictions it assigns less probability to every nongreedy action than the inverse-gap allocation it starts from, and the difference is large in practice: its mean greedy probability in the neural experiments is 99.9%, against at most 84.4% for the baselines (Table 1). MinCE thus acts far more greedily while retaining the regret guarantee.
4. **Necessary and sufficient exploration laws** (Theorem 6). No certified allocation explores less than a constant fraction of the inverse-gap allocation unless some nongreedy prediction approaches one; in that regime the minimum exploration is $\Theta(\gamma^{-1} \sum_{a \neq b_t} L_{t,a}^{-1})$, where b_t is the greedy action and $L_{t,a}$ is the log-loss cost of an alternative in which action a is lossless. A reduction beyond a constant factor is therefore possible exactly when the predictor is confident that the alternatives are bad, and MinCE realizes it whenever it is available.
5. **Empirical gains** (Section 5). In simulations with linear, generalized linear, and nonlinear rewards under Gaussian, uniform, and adversarial contexts, MinCE attains the lowest cumulative regret overall. On 130 OpenML classification datasets with three-layer neural oracles, it achieves the lowest loss on 112 datasets, against at most 9 for any baseline, and thus outperforms every baseline on the vast majority of datasets.

In addition, an epoch-based variant extends the guarantee to offline regression oracles (Appendix F).

2 Problem Setting

Let $\mathcal{X}$ be the context space and $\mathcal{A} = [A] = \{1, \ldots, A\}$ the action set, with $A \geq 2$. Fix a horizon T and a regression class $\mathcal{F} \subseteq \{f : \mathcal{X} \times \mathcal{A} \to [0, 1]\}$. Write Δ_A for the probability simplex.

At each round t, the environment provides a context $x_t \in \mathcal{X}$ and a conditional distribution over loss vectors $\ell_t : \mathcal{A} \to [0, 1]$. The learner receives x_t, forms a distribution $p_t \in \Delta_A$, draws $a_t \sim p_t$, and observes only $\ell_t(a_t)$. The context can change at every round. Both x_t and the conditional loss distribution may depend arbitrarily on previous contexts, selected actions, and observed losses.

Let $\mathcal{H}_t$ be the sigma-field generated by the information immediately before the current action draw: the observed history $\{(x_s, a_s, \ell_s(a_s))\}_{s<t}$, the current x_t, and the oracle and allocation state used to form p_t. The current action randomization is independent of ℓ_t conditional on $\mathcal{H}_t$; the environment cannot react to this still-unobserved draw. Our first assumption is that the conditional mean loss is realized by the regression class.

Assumption 1 (Realizability). *There exists $f^* \in \mathcal{F}$ such that, for all $t \in [T]$ and $a \in \mathcal{A}$,*

$$\mathbb{E}[\ell_t(a) \mid \mathcal{H}_t] = f^*(x_t, a). \tag{1}$$

The mean function f^* is fixed, but the values $f^*(x_t, a)$ can change with the context. Realizability is the standard assumption behind regression-oracle reductions for contextual bandits (Agarwal et al., 2012; Foster et al., 2018a; Foster & Rakhlin, 2020; Foster & Krishnamurthy, 2021; Zhu et al., 2022; Zhang et al., 2024); conditioning on $\mathcal{H}_t$ rather than on x_t alone permits adaptively chosen

contexts and loss distributions. Define the optimal policy by $\pi^*(x) = \min \operatorname{argmin}_a f^*(x, a)$. Write $f^*_{t,a} = f^*(x_t, a)$ and $f^*_t = (f^*_{t,a})_{a\in\mathcal{A}}$. We measure

$$\mathrm{Reg}_{\mathrm{CB}}(T) = \sum_{t=1}^{T} \big[\ell_t(a_t) - \ell_t(\pi^*(x_t))\big], \qquad L^*_T = \sum_{t=1}^{T} f^*(x_t, \pi^*(x_t)).$$

We bound expected regret, which equals $\mathbb{E}\sum_t (p_t^\top f^*_t - \min_a f^*_{t,a})$. The comparator uses the context sequence encountered by the learner. The variable L^*_T is the optimal cumulative conditional mean loss; its expectation equals the expected cumulative loss of π^*.

Online regression and greedy actions. The log-loss oracle $\mathsf{Alg}_{\mathrm{KL}}$ returns predictions $\widehat{y}_{t,a} \in (0, 1)$ for each action before the action draw. Write $\widehat{y}_t = (\widehat{y}_{t,a})_{a\in\mathcal{A}}$ for the prediction vector. Its designated greedy action is $b_t = \min \operatorname{argmin}_a \widehat{y}_{t,a}$. The tie rule is fixed; "nongreedy" means $a \neq b_t$, including other tied minimizers. Greedy probability p_{t,b_t} measures adherence to the predictor, and need not equal the probability of selecting a truly optimal action.

For $\ell_{\log}(u, y) = -y \log u - (1 - y) \log(1 - u)$, our second assumption is the online regression interface of Foster & Krishnamurthy (2021): the oracle has a regret guarantee for the log loss on the examples it actually receives.

Assumption 2 (Online regression oracle for log loss)**.** *The oracle* $\mathsf{Alg}_{\mathrm{KL}}$ *guarantees that, for every (possibly adaptively chosen) sequence* $(x_1, a_1, \ell_1(a_1)), \ldots, (x_T, a_T, \ell_T(a_T))$*,*

$$\sum_{t=1}^{T} \ell_{\log}(\widehat{y}_{t,a_t}, \ell_t(a_t)) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell_{\log}(f(x_t, a_t), \ell_t(a_t)) \le \mathrm{Reg}_{\mathrm{KL}}(T), \tag{2}$$

where $\mathrm{Reg}_{\mathrm{KL}}(T) \ge 0$ *is a known deterministic bound for the prescribed horizon; a bound that holds in expectation suffices.*

Predictions need not belong to $\mathcal{F}$, and boundary comparator values use the usual extended log-loss conventions. Online regression with the logarithmic loss (sequential probability assignment) is a well-studied problem with efficient algorithms for many classes. For a finite class, Vovk's aggregating algorithm gives $\mathrm{Reg}_{\mathrm{KL}}(T) \le \log|\mathcal{F}|$ (Vovk, 1995). For logistic and generalized linear classes in d dimensions, improper online algorithms achieve regret logarithmic in T (Kakade & Ng, 2004; Foster et al., 2018b; Jézéquel et al., 2020), and for general nonparametric classes the minimax rate under log loss is characterized by sequential entropy (Bilodeau et al., 2020). The same interface underlies FastCB (Foster & Krishnamurthy, 2021), and its squared-loss analogue underlies SquareCB (Foster & Rakhlin, 2020).

Define Bernoulli KL divergence by $\mathrm{KL}(u\|v) := u\log(u/v) + (1 - u)\log((1 - u)/(1 - v))$. Conditional cross-entropy differences then give $\mathbb{E}\big[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} p_{t,a}\,\mathrm{KL}(f^*_{t,a}\|\widehat{y}_{t,a})\big] \le \mathrm{Reg}_{\mathrm{KL}}(T)$. This identity holds for every $[0, 1]$-valued loss. Note that log loss only trains the predictor; the bandit loss itself is not affected.

3 Proposed Algorithm: MinCE

Our algorithm MinCE (Algorithm 1) follows the regression-oracle template of SquareCB and FastCB, which are based on inverse-gap weighting (Abe & Long, 1999): at each round it queries the oracle for the predicted losses $\widehat{y}_t$, converts them into an action distribution p_t, samples an action, and feeds the observed loss back to the oracle. Instead of simply playing the inverse-gap distribution, however, MinCE treats it as a starting point and moves probability from the nongreedy actions to the greedy action b_t for as long as a computable *certificate* confirms that the one-step regret inequality (4) still holds. Below we define the certificate, show that it can be evaluated efficiently, and describe how CertifiedAllocation (Algorithm 2) computes the least certified allocation.

Reference allocation. Recall that $b_t = \min \operatorname{argmin}_a \widehat{y}_{t,a}$ is the greedy action for the current predictions. The initial allocation is the reweighted inverse-gap distribution

$$q_{t,a} = \frac{\widehat{y}_{t,b_t}}{A\widehat{y}_{t,b_t} + \gamma(\widehat{y}_{t,a} - \widehat{y}_{t,b_t})} \quad (a \neq b_t), \qquad q_{t,b_t} = 1 - \sum_{a\neq b_t} q_{t,a}, \tag{3}$$

Algorithm 1 MinCE: Minimum Certified Exploration

Input: Regression oracle $\mathsf{Alg}_{\mathrm{KL}}$; exploration parameter $\gamma \geq 10A$; horizon T.

for $t = 1, \ldots, T$ **do**
  Receive the current context x_t.
  Predict $\widehat{y}_{t,a} \leftarrow \mathsf{Alg}_{\mathrm{KL}}.\mathsf{Predict}(x_t, a)$ for every $a \in \mathcal{A}$.
  Set $b_t \leftarrow \min \operatorname{argmin}_a \widehat{y}_{t,a}$ and form q_t using (3).
  $p_t \leftarrow \mathsf{CertifiedAllocation}(\widehat{y}_t, q_t, b_t; \gamma, T)$ (Algorithm 2).
  Draw $a_t \sim p_t$ and observe $\ell_t(a_t)$.
  Update $\mathsf{Alg}_{\mathrm{KL}}$ with $(x_t, a_t, \ell_t(a_t))$ using unweighted log loss.
end for

where $\gamma > 0$ is the exploration parameter: an action receives less probability the larger its predicted gap $\widehat{y}_{t,a} - \widehat{y}_{t,b_t}$, and the leftover mass goes to b_t. MinCE takes q_t as its initial allocation but recalibrate the allocation based on a one-step regret certificate (Definition 1). Throughout, $\gamma \geq 10A$; Section 4 shows that this range makes q_t certified and makes the iteration below converge at a uniform rate.

The one-step regret certificate. Regret analyses of inverse-gap weighting (Abe & Long, 1999; Foster & Rakhlin, 2020; Foster & Krishnamurthy, 2021) rest on a one-step inequality that relates the instantaneous regret of an allocation to its exploration and to the prediction error. For FastCB this inequality reads, for every mean-loss vector $f = (f_a)_{a \in \mathcal{A}} \in [0,1]^A$,

$$p^\top f - \min_a f_a \leq \frac{5A}{\gamma}\, p^\top f + 14\gamma \sum_a p_a \,\mathrm{KL}(f_a \| \widehat{y}_{t,a}). \tag{4}$$

When f is the true mean-loss vector, the left side is the instantaneous regret of p; on the right, the first term is a small multiple of the expected loss, and the second is a prediction-error term whose sum over rounds is controlled by the oracle regret (Section 2). Foster & Krishnamurthy (2021) establish an inequality of this form for $p = q_t$, with the triangular discrimination in place of the KL term. Any allocation that satisfies (4) for every f inherits the same regret analysis, so we make this the requirement.

Definition 1 (One-step regret certificate)**.** For predictions $\widehat{y}_t \in (0,1)^A$ and $p \in \Delta_A$, the *one-step regret certificate* of p is its largest violation of (4) over all mean vectors (negative when (4) holds with room to spare),

$$C_t(p) := \sup_{f \in [0,1]^A} \left\{ \Big(1 - \frac{5A}{\gamma}\Big) p^\top f - \min_a f_a - 14\gamma \sum_{a \in \mathcal{A}} p_a \,\mathrm{KL}(f_a \| \widehat{y}_{t,a}) \right\}. \tag{5}$$

We call p *certified at round* t if $C_t(p) \leq 0$, that is, if p satisfies (4) for every $f \in [0,1]^A$.

The definition involves a supremum over all mean vectors, but the algorithm never has to evaluate it, because the supremum has a closed form (Lemma 1). Set $r = (1 - 5A/\gamma)/(14\gamma)$ and, for each action, define the coefficient $U_{t,a}$ and the scalar function $D_{t,a}$,

$$\begin{aligned} U_{t,a} &:= \log(1 - \widehat{y}_{t,a} + \widehat{y}_{t,a} e^{r}), \\ D_{t,a}(s) &:= s\left[U_{t,a} - \log(1 - \widehat{y}_{t,a} + \widehat{y}_{t,a} e^{r - 1/(14\gamma s)})\right] \quad (s > 0), \qquad D_{t,a}(0) := 0. \end{aligned} \tag{6}$$

Both are built from the cumulant generating function $\theta \mapsto \log(1 - \widehat{y}_{t,a} + \widehat{y}_{t,a} e^{\theta})$ of a Bernoulli($\widehat{y}_{t,a}$) variable: $U_{t,a}$ is its value at r, and $D_{t,a}(s)$ compares that value with the value at the shifted argument $r - 1/(14\gamma s)$; the convention $D_{t,a}(0) = 0$ is the continuous extension. Lemma 1 in Section 4.1 shows that, with $U_t := (U_{t,a})_{a \in \mathcal{A}}$,

$$C_t(p) = 14\gamma\Big[p^\top U_t - \min_{a \in \mathcal{A}} D_{t,a}(p_a)\Big]. \tag{7}$$

The closed form is easy to read. The inner minimum in (4) is a choice of comparator action; in (7) each action contributes one constraint, and p is certified exactly when

$$D_{t,a}(p_a) \;\geq\; p^\top U_t \qquad \text{for every } a \in \mathcal{A}. \tag{8}$$

The right side is a common *level*, a weighted average of the coefficients $U_{t,a}$, which increase with the predicted loss $\widehat{y}_{t,a}$. The left side is a *correction* that increases with p_a, so giving an action more probability makes its own constraint easier. Evaluating $C_t(p)$ takes only $O(A)$ time and involves no optimization over $\mathcal{F}$; $U_{t,a}$ and $D_{t,a}$ are computed from the predictions, not tuned.

Algorithm 2 CertifiedAllocation: one bracketed inverse per nongreedy action per sweep

Input: Predictions $\widehat{y}_t$, reference q_t, greedy action b_t; parameters γ, T.
Output: Certified p_t with $p^{\mathrm{opt}}_{t,a} \le p_{t,a} \le (1+1/T)p^{\mathrm{opt}}_{t,a}$ for $a \neq b_t$ (Theorem 5).

Compute $r = (1-5A/\gamma)/(14\gamma)$ and $U_{t,a}, D_{t,a}$ in (6); cache $w_{t,a} \leftarrow \widehat{y}_{t,a} e^{r-U_{t,a}}$.
Set $\tau \leftarrow 1/(16T)$, $\lambda \leftarrow (A-1)/(112A)$; initialize $p_t \leftarrow q_t$ and $\underline{z}_t \leftarrow U_{t,b_t}$.
while true **do**
  $\overline{z}_t \leftarrow p_t^\top U_t$. // level of the current allocation
  for each $a \neq b_t$ **do**
    $(\underline{u}_{t,a}, \overline{u}_{t,a}) \leftarrow$ InverseBracket$(D_{t,a}, \overline{z}_t, p_{t,a}, \tau)$ (Algorithm 3). // bracket of $D^{-1}_{t,a}(\overline{z}_t)$
  end for
  $g_t^- \leftarrow U_{t,b_t} + \sum_{a \neq b_t}(U_{t,a} - U_{t,b_t})\underline{u}_{t,a}$. // lower estimate of the next level
  $\underline{z}_t \leftarrow \max\{\underline{z}_t, (g_t^- - \lambda\overline{z}_t)/(1-\lambda)\}$. // lower bound on $z_t^\star$
  $\underline{s}_{t,a} \leftarrow \underline{u}_{t,a} \dfrac{\underline{z}_t}{\overline{z}_t} \dfrac{w_{t,a} - 14\gamma\overline{z}_t}{w_{t,a} - 14\gamma\underline{z}_t}$ for $a \neq b_t$. // lower bound on $p^{\mathrm{opt}}_{t,a}$
  $p'_{t,a} \leftarrow \overline{u}_{t,a}$ for $a \neq b_t$; $p'_{t,b_t} \leftarrow 1 - \sum_{a \neq b_t} \overline{u}_{t,a}$. // conservative sweep
  if $\overline{u}_{t,a} \le (1+1/T)\underline{s}_{t,a}$ for every $a \neq b_t$ **then**
    return p'_t.
  end if
  $p_t \leftarrow p'_t$.
end while

Computing the least certified allocation. Two properties of (8) suggest a simple iteration. First, $D_{t,a}$ is strictly increasing, so for a given level z the smallest probability of action a that meets its constraint is the scalar inverse $D^{-1}_{t,a}(z)$. Second, the greedy action has the smallest coefficient U_{t,b_t}, so moving probability from any other action to b_t lowers the level and makes every constraint easier. Starting from a certified p_t (initially q_t), a *sweep* reads off the current level $\overline{z}_t = p_t^\top U_t$, shrinks every nongreedy probability to the smallest value that still meets this level, and gives the freed mass to b_t:

$$p'_{t,a} = D^{-1}_{t,a}(\overline{z}_t) \le p_{t,a} \quad (a \neq b_t), \qquad p'_{t,b_t} = 1 - \sum_{a \neq b_t} p'_{t,a}. \tag{9}$$

The new allocation is again certified: the nongreedy constraints hold at the old level, the greedy constraint only became easier, and the new level is no larger than the old one. Repeating the sweep decreases every nongreedy probability monotonically, and the levels converge to a limit $z_t^\star$. The limit allocation, $p^{\mathrm{opt}}_{t,a} = D^{-1}_{t,a}(z_t^\star)$ for $a \neq b_t$, is what MinCE aims to play; Section 4.2 proves that it is the *least* certified allocation, in the sense that every certified p satisfies $p_a \ge p^{\mathrm{opt}}_{t,a}$ for all $a \neq b_t$.

Algorithm 2 turns (9) into a finite procedure with a stopping rule, using three ingredients. (i) *Bracketed inverses.* Each inverse is replaced by a bracket $\underline{u}_{t,a} \le D^{-1}_{t,a}(\overline{z}_t) \le \overline{u}_{t,a} \le p_{t,a}$ with $\overline{u}_{t,a} \le (1+\tau)\underline{u}_{t,a}$, computed by the safeguarded Newton subroutine InverseBracket (Algorithm 3, Appendix D.4); the internal accuracy $\tau = 1/(16T)$ is set from the horizon, so no tolerance or iteration count is exposed. The sweep uses the *upper* endpoint $\overline{u}_{t,a}$ as the new probability: it never undershoots the exact inverse, so the argument above still applies and every iterate is certified. (ii) *A lower bound on the limit level.* The sweep contracts the level toward $z_t^\star$ at a rate at most $\lambda = (A-1)/(112A)$ (Theorem 5), so the lower estimate g_t^- of the next level, computed from the lower endpoints, yields $\underline{z}_t = (g_t^- - \lambda\overline{z}_t)/(1-\lambda) \le z_t^\star$. (iii) *A lower bound on the target.* A sensitivity bound for $D^{-1}_{t,a}$ (Lemma 13) transports the lower endpoint from the level $\overline{z}_t$ down to the level $\underline{z}_t$, using only the cached tilted mean $w_{t,a}$, and gives $\underline{s}_{t,a} \le D^{-1}_{t,a}(\underline{z}_t) \le p^{\mathrm{opt}}_{t,a}$ without a further inverse. The sweep stops as soon as $\overline{u}_{t,a} \le (1+1/T)\underline{s}_{t,a}$ for every $a \neq b_t$, which sandwiches the returned allocation between $p^{\mathrm{opt}}_{t,a}$ and $(1+1/T)p^{\mathrm{opt}}_{t,a}$. Each sweep costs $\mathcal{O}(A)$ time plus one bracketed inverse per nongreedy action, and no additional calls to the regression oracle are made.

4 Theoretical Guarantees

This section answers the core question: does playing the least certified allocation cost anything in regret? Section 4.1 shows that it does not, since every certified allocation inherits the first-order regret

analysis. Section 4.2 shows that the allocation MinCE targets is the least certified one, Section 4.3 that Algorithm 2 reaches it within $\mathcal{O}(\log T)$ sweeps, and Section 4.4 quantifies how much exploration the certificate lets us remove.

4.1 Certified allocations attain the optimal regret guarantee

We first prove that the certificate of Definition 1 is computable and that the reference allocation is certified, which is what Algorithm 2 needs to start.

Lemma 1 (Closed form of the one-step regret certificate)**.** *For every round t and $p \in \Delta_A$, the supremum in* (5) *is attained and equals* (7)*; hence p is certified if and only if* (8) *holds. The function C_t is continuous and convex in p, and $C_t(q_t) < 0$.*

A certified allocation satisfies (4) for the true means in particular. Summing (4) over rounds and controlling the prediction-error term by Assumption 2 gives the following bound, in which nothing depends on how much exploration was removed.

Theorem 2 (Regret guarantee)**.** *Under Assumptions 1 and 2, MinCE with any fixed $\gamma \geq 10A$ satisfies, in exact real arithmetic,*

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{CB}}(T)] \leq \frac{(5A/\gamma)\mathbb{E}L_T^* + 14\gamma\,\mathrm{Reg}_{\mathrm{KL}}(T)}{1 - 5A/\gamma}. \tag{10}$$

If $\mathrm{Reg}_{\mathrm{KL}}(T) > 0$ and $\bar{L} \geq \mathbb{E}L_T^*$ is known, the choice $\gamma = \max\left\{10A, \sqrt{\frac{A\bar{L}}{3\,\mathrm{Reg}_{\mathrm{KL}}(T)}}\right\}$ gives

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{CB}}(T)] \leq 10\sqrt{3}\sqrt{A\bar{L}\,\mathrm{Reg}_{\mathrm{KL}}(T)} + 280A\,\mathrm{Reg}_{\mathrm{KL}}(T). \tag{11}$$

Two points deserve emphasis. First, the bound is *first-order*: the regret scales with $\sqrt{A\bar{L}\ \mathrm{Reg}_{\mathrm{KL}}(T)}$, matching the regret guarantee in Foster & Krishnamurthy (2021), so minimizing exploration costs nothing in the regret bound. Second, the guarantee applies to the allocation that Algorithm 2 actually returns, not only to the exact minimizer: the finite computation preserves the certificate at every sweep (Theorem 5), so its accuracy parameter does not appear in (10). The bound needs no upper bound on the optimal loss; $\bar{L}$ enters only through the choice of γ.

Corollary 3 (Worst-case regret)**.** *Suppose $\mathrm{Reg}_{\mathrm{KL}}(T) > 0$, and choose γ as in Theorem 2 with $\bar{L} = T$. Under the same assumptions,*

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{CB}}(T)] \leq \min\left\{T, 10\sqrt{3}\sqrt{AT\,\mathrm{Reg}_{\mathrm{KL}}(T)} + 280A\,\mathrm{Reg}_{\mathrm{KL}}(T)\right\}. \tag{12}$$

For a realizable class of $M \geq 2$ interior-valued predictors, the uniform log-loss mixture satisfies Assumption 2 with $\mathrm{Reg}_{\mathrm{KL}}(T) = \log M$, yielding

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{CB}}(T)] = \mathcal{O}\Big(\min\{T, \sqrt{AT\log M}\}\Big). \tag{13}$$

For finite classes, (13) is minimax optimal for fixed A and matches the realizable lower bound $\Omega(\sqrt{AT\log M/\log A})$ of Agarwal et al. (2012, Theorem 5.1), valid for $T \geq \log M/\log A$, up to a $\sqrt{\log A}$ factor. The least certified allocation therefore attains the optimal worst-case rate.

4.2 The allocation MinCE targets is the least certified one

We now show that the sweep iteration of Section 3 has a well-defined limit and that the limit is optimal among *all* certified allocations, not only among those the iteration can reach. Each $D_{t,a}$ is strictly increasing and concave. Write its inverse, and the map that one exact sweep (9) induces on the level $p^\top U_t$, as

$$z_t^q = q_t^\top U_t, \qquad H_{t,a}(z) = D_{t,a}^{-1}(z) \quad (0 \leq z \leq z_t^q),$$
$$G_t(z) = U_{t,b_t} + \sum_{a \neq b_t} (U_{t,a} - U_{t,b_t}) H_{t,a}(z).$$

Strict reference feasibility, $C_t(q_t) < 0$, ensures $z_t^q < D_{t,a}(q_{t,a})$ for every action, so the inverses are defined on $[0, z_t^q]$.

Theorem 4 (The simultaneous exploration minimum)**.** *At every round t, there is a unique fixed point $z_t^\star = G_t(z_t^\star)$ in $[U_{t,b_t}, z_t^q]$. The allocation*

$$p_{t,a}^{\mathrm{opt}} = H_{t,a}(z_t^\star) \quad (a \neq b_t), \qquad p_{t,b_t}^{\mathrm{opt}} = 1 - \sum_{a \neq b_t} H_{t,a}(z_t^\star)$$

is certified. Every $p \in \Delta_A$ with $C_t(p) \leq 0$ satisfies

$$p_a \geq p_{t,a}^{\mathrm{opt}}, \qquad p_{t,a}^{\mathrm{opt}} < q_{t,a} \qquad (a \neq b_t).$$

Thus p_t^{opt} uniquely minimizes $1 - p_{b_t}$ and every positively weighted sum of nongreedy probabilities over the certified simplex.

The theorem is stronger than the existence of a minimizer of one objective: the certified set has a *least element* coordinatewise, so there is no tradeoff among nongreedy actions, and every positively weighted exploration objective is minimized by the same allocation. The reason is that all constraints in (8) share one level, which the greedy action, having the smallest coefficient U_{t,b_t}, lowers whenever it gains probability. On the learner's own trajectory, $p_{t,a} \leq q_{t,a}$ gives $\mathbb{E}\sum_t \mathbf{1}\{a_t \neq b_t\} = \mathbb{E}\sum_t (1 - p_{t,b_t}) \leq \mathbb{E}\sum_t (1 - q_{t,b_t})$, where the right side is the exploration the IGW allocation (e.g., FastCB) would prescribe at MinCE's own predictions.

4.3 Finite computation preserves the certificate

Theorem 4 describes p_t^{opt} through exact inverses. The next result shows that the bracketed iteration of Algorithm 2 reaches it, with the number of sweeps controlled uniformly over prediction vectors; the full statement, including bounds on the inner bracket refinements, is Theorem 14 in Appendix D.

Theorem 5 (Finite computation of the exploration minimum)**.** *At every round t, G_t is nondecreasing on $[U_{t,b_t}, z_t^q]$ with $G_t' \leq (A-1)/(112A) < 1/112$. Consequently, in exact real arithmetic, Algorithm 2 terminates within $N_T = 1 + \lceil \log_{100} T \rceil$ sweeps, each using $A - 1$ bracketed scalar inverses and no regression calls, and returns a certified allocation with*

$$p_{t,a}^{\mathrm{opt}} \leq p_{t,a} \leq \min\{q_{t,a}, (1 + 1/T) p_{t,a}^{\mathrm{opt}}\} \qquad (a \neq b_t).$$

In particular, $N_{5000} = 3$.

The contraction constant is the one Algorithm 2 uses to lower-bound $z_t^\star$. Two consequences matter. Every sweep produces a certified allocation, so the finite computation adds no term to (10); and the relative accuracy gives $\sum_t [(1 - p_{t,b_t}) - (1 - p_{t,b_t}^{\mathrm{opt}})] \leq 1$ on the realized prediction sequence, so over the whole horizon the computation costs at most one expected nongreedy draw beyond the exact minimum. The statement concerns exact arithmetic; the implementation uses safeguarded refinements and records numerical certificate checks (Appendix D).

4.4 How much exploration can be removed?

Theorem 4 shows how to explore least; we now ask how small the minimum is compared with the reference. Write $\mathcal{E}_t^{\mathrm{opt}} = 1 - p_{t,b_t}^{\mathrm{opt}}$ and $Q_t = 1 - q_{t,b_t}$ for the total nongreedy probability of the two allocations.

Theorem 6 (A necessary boundary condition and matching exploration laws)**.** *For any round t, every certified $p \in \Delta_A$ satisfies*

$$p_a \geq \frac{1 - \widehat{y}_{t,a}}{281}\, q_{t,a} \qquad (a \neq b_t).$$

If $\widehat{y}_{t,a} \leq 1 - d$ for all $a \neq b_t$, with $d \in (0,1)$, then $\mathcal{E}_t^{\mathrm{opt}} \geq dQ_t/281$. Conversely, suppose $\widehat{y}_{t,b_t} = h \in [1/10, 1/2]$ and $\widehat{y}_{t,a} = 1 - e^{-L_{t,a}}$, $L_{t,a} \geq 4$, for $a \neq b_t$. Then

$$\frac{1}{560\gamma} \sum_{a \neq b_t} \frac{1}{L_{t,a}} \leq \mathcal{E}_t^{\mathrm{opt}} \leq \frac{3}{56\gamma} \sum_{a \neq b_t} \frac{1}{L_{t,a}}.$$

These constants are uniform over the stated ranges.

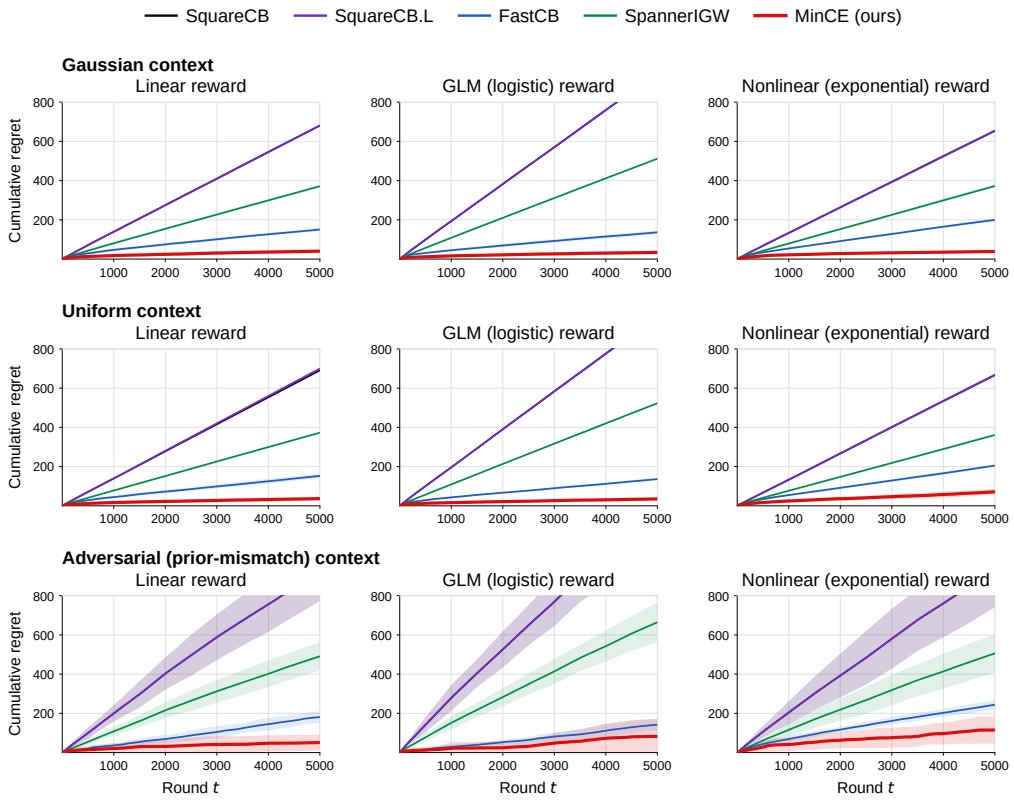

Figure 1: **Synthetic cumulative pseudo-regret.** Rows vary the context process; columns vary the mean reward. Curves show means over ten seeds and bands show one sample SD. MinCE is red. Appendix H reports the full results in a table and the corresponding greediness curves.

The first part limits what any certificate-preserving allocation can achieve: unless some nongreedy prediction approaches one, exploration shrinks by at most a constant factor $(281/d)$ relative to FastCB. The second part shows that this boundary is exactly where the reduction becomes unbounded. When the nongreedy actions are predicted to be confidently bad, $\widehat{y}_{t,a} = 1 - e^{-L_{t,a}}$, the required exploration is $\Theta(\gamma^{-1} \sum_{a \neq b_t} L_{t,a}^{-1})$, whereas $Q_t = \Theta((A-1)/\gamma)$ in the same regime, so the saving grows like $L_{t,a} = \mathrm{KL}(0 \| \widehat{y}_{t,a})$, the log-loss cost of an alternative in which action a is lossless. Such alternatives are the mean vectors that force exploration in the certificate, and explicit feasible allocations match them up to the stated constants; Theorem 20 in Appendix E.1 identifies the exact leading constant, $\mathcal{E}_t^{\mathrm{opt}} \sim U_{t,b_t} \sum_{a \neq b_t} L_{t,a}^{-1}$, as the $L_{t,a}$ grow proportionally. All statements compare allocations at the same prediction vector and make no assumption on how contexts evolve. Appendices E and E.1 give the proofs; Appendix D.5 adds an actionwise description, $H_{t,a}(U_{t,b_t}) \leq p_{t,a}^{\mathrm{opt}} \leq 1.285\, H_{t,a}(U_{t,b_t})$, of each optimal nongreedy probability by a single scalar inverse.

What the exploration results do and do not say. When b_t is truly optimal, every reduction of a nongreedy probability reduces instantaneous regret, since each term of $\sum_{a \neq b_t} p_{t,a}(f_{t,a}^* - f_{t,b_t}^*)$ is nonnegative; when the predictor is wrong this need not hold, so the results above guarantee an exploration minimum and a regret bound, not cumulative dominance over another learner. The minimum is never deterministic: if $p_a = 0$ for some a, then $\min_a D_{t,a}(p_a) = 0 < p^\top U_t$.

5 Experiments

We evaluate whether reducing certified exploration improves incurred performance while keeping the regression interface and computational cost practical. We compare MinCE with SquareCB,

Table 1: **OpenML dataset results.** Loss and greedy probability give equal weight to each of the 130 datasets. Loss SD is across ten seed-specific dataset averages. For the full results, see Table 4.

Method	Best / 130 datasets	Loss (%)	Greedy action (%)
SquareCB (Foster & Rakhlin, 2020)	1	54.47 ± 0.38	51.428
SquareCB.L (Foster & Rakhlin, 2020)	3	54.15 ± 0.21	51.085
FastCB (Foster & Krishnamurthy, 2021)	9	37.35 ± 0.23	84.363
SpannerIGW (Zhu et al., 2022)	5	47.42 ± 0.57	73.438
MinCE	112	35.34 ± 0.31	99.943

SquareCB.L, FastCB, and SpannerIGW (Foster & Rakhlin, 2020; Foster & Krishnamurthy, 2021; Zhu et al., 2022). All methods use neural regressors with three hidden $\tanh$ layers of width 32, selected-action updates, and ten matched evaluation seeds. SquareCB.L uses the SquareCB allocation with a log-loss oracle. Allocation parameters are constant and follow the theoretical formulas with explicit bounded-loss oracle budgets; FastCB and MinCE share $\gamma = 10A$. Only optimizer learning rates are selected empirically, using a separate tuning seed and the same candidate grid. Appendix G gives the details.

5.1 Experiments on simulated data

We cross three reward models (linear, logistic, and exponential) with three context processes: Gaussian IID contexts, uniform IID contexts, and an oblivious prior-mismatch sequence that changes in blocks and is chosen before learning, without regard to any learner's actions, to be unfavorable to a reference predictor. Each of the nine conditions uses $A = 5$ actions and $T = 5{,}000$ rounds. Figure 1 reports cumulative pseudo-regret. MinCE has the lowest mean final regret in all nine conditions, improving on FastCB by 41.8–80.8%, while exploring far less: its mean probability of choosing its own predicted-greedy action is 99.858%, compared with 80.131% for FastCB. This is the comparison that the theory cannot settle by itself. The exploration results of Section 4 hold at the same predictions, whereas here each learner is trained on its own feedback, so a greedier learner could in principle learn more slowly and fall behind. It does not: the reduction in exploration accompanies lower regret in every condition, including the prior-mismatch sequence. MinCE remains randomized; near-unit greedy probability is not a deterministic greedy schedule, and the residual exploration is what the certificate requires. The cost of computing the least certified allocation is modest: mean online-loop runtime is 0.187 s per run, versus 0.171 s for FastCB and 0.169 s for SquareCB, an overhead of less than ten percent. Appendix H documents the construction and timing protocol.

5.2 Experiments on OpenML classification tasks

On 130 OpenML datasets, each run makes one randomly ordered pass and observes only the chosen class's binary loss. We report incurred classification loss, including exploration costs, rather than regret against an unknown best neural comparator. MinCE attains the uniquely lowest mean loss on 112 datasets and has lower mean loss than FastCB on 116 of 130 (Table 1). Its equal-dataset mean loss is 35.34%, versus 37.35% for FastCB, while its mean greedy probability rises from 84.363% to 99.943%. The median runtime overhead relative to FastCB is 5.60%.

6 Conclusion

We asked how little exploration a regret guarantee actually requires in general contextual bandits with a regression oracle, and answered it with a one-step regret certificate: a closed-form test that any allocation must pass to inherit the guarantee. The certified allocations have a least element, and MinCE computes it in a few sweeps of scalar inversions without additional oracle calls or hyperparameters. The result is an algorithm that matches the best existing first-order regret bound and the minimax rate while exploring far less, with tight laws describing when the reduction exceeds a constant factor, and that outperforms the baselines on synthetic tasks and on 130 OpenML datasets. Since any one-step inequality defines a certificate, applying the core principle to other methods is a natural next step.

AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks requiring disclosure, such as research ideation, literature review, experimental design, data analysis, or generating original text or results. We used generative AI tools only minimally, for grammar checks, proofreading, and debugging code. We have reviewed all AI-assisted work. The authors manually checked every grammar and proofreading suggestion to make sure it did not change the intended meaning, and reviewed and tested all AI-assisted code fixes for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The appendix contains the complete proofs (Appendices B–F) and full experimental details, including the fixed configurations, seeds, and numerical summaries. The neural protocol and per-dataset results are given in Appendix I; the synthetic construction, oracle guarantees, and runtime measurements are given in Appendix H. The supplementary material contains the source code for the algorithms and the experiments.

REFERENCES

Naoki Abe and Philip M. Long. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 3–11, 1999.

Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert E. Schapire. Contextual bandit learning with predictable rewards. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pp. 19–26. PMLR, 2012.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. *Journal of Machine Learning Research*, 22(133):1–49, 2021.

Blair Bilodeau, Dylan J. Foster, and Daniel M. Roy. Tight bounds on minimax regret under logarithmic loss via self-concordance. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 919–929. PMLR, 2020.

Rohan Deb, Yikun Ban, Shiliang Zuo, Jingrui He, and Arindam Banerjee. Contextual bandits with online neural regression. In *The Twelfth International Conference on Learning Representations*, 2024.

Dylan J. Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Dylan J. Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3199–3210. PMLR, 2020.

Dylan J. Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert E. Schapire. Practical contextual bandits with regression oracles. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1539–1548. PMLR, 2018a.

Dylan J. Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 167–208. PMLR, 2018b.

Dylan J. Foster, Noah Golowich, and Yanjun Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 3969–4043. PMLR, 2023.

Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Efficient improper learning for online logistic regression. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2085–2108. PMLR, 2020.

Sham M. Kakade and Andrew Y. Ng. Online bounds for Bayesian algorithms. In *Advances in Neural Information Processing Systems*, volume 17, 2004.

Sampath Kannan, Jamie Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems*, 2018.

Seok-Jin Kim and Min-hwan Oh. Local anti-concentration class: Logarithmic regret for greedy linear contextual bandit. In *Advances in Neural Information Processing Systems*, volume 37, pp. 77525–77592, 2024.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, volume 20, pp. 817–824, 2008.

Harin Lee and Min-hwan Oh. Infrequent exploration in linear bandits. In *Advances in Neural Information Processing Systems*, 2025.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670, 2010.

Gergely Neu, Iuliia Olkhovskaia, Matteo Papini, and Ludovic Schwartz. Lifting the information ratio: An information-theoretic analysis of Thompson sampling for contextual bandits. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9486–9498, 2022.

Gergely Neu, Matteo Papini, and Ludovic Schwartz. Optimistic information directed sampling. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 3970–4006. PMLR, 2024.

Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. The externalities of exploration and how data diversity helps exploitation. In *Proceedings of the 31st Conference On Learning Theory*, 2018.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022. doi: 10.1287/moor.2021.1193.

Vladimir Vovk. A game of prediction with expert advice. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pp. 51–60. ACM, 1995.

Yunbei Xu and Assaf Zeevi. Bayesian design principles for frequentist sequential learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38768–38800. PMLR, 2023.

Mengxiao Zhang, Yuheng Zhang, Olga Vrousgou, Haipeng Luo, and Paul Mineiro. Practical contextual bandits with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 30592–30617, 2023.

Mengxiao Zhang, Yuheng Zhang, Haipeng Luo, and Paul Mineiro. Efficient contextual bandits with uninformed feedback graphs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60336–60358. PMLR, 2024.

Yinglun Zhu, Dylan J. Foster, John Langford, and Paul Mineiro. Contextual bandits with large action spaces: Made practical. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27428–27453. PMLR, 2022.

A Additional discussion on Related work and limitations

Regression-oracle reductions. SquareCB and FastCB establish the regression interface and derive their allocations from one-step regret inequalities (Foster & Rakhlin, 2020; Foster & Krishnamurthy, 2021); Simchi-Levi & Xu (2022) obtain the same rate with an offline oracle under IID contexts, and Appendix F gives the corresponding epoch-based variant of MinCE. MinCE starts from the FastCB allocation and its one-step inequality; its contribution is the least certified allocation, a finite computation that preserves the certificate, and matching exploration laws. Neural online regression oracles for this interface are analyzed by Deb et al. (2024). SpannerIGW extends inverse-gap exploration to large action spaces through optimal design (Zhu et al., 2022); our experiments use its exact finite-action specialization. Information-directed sampling, the decision-estimation coefficient, and the algorithmic information ratio offer broader design principles with tractable specializations (Neu et al., 2024; Foster et al., 2023; Xu & Zeevi, 2023). They select actions by optimizing a per-round objective that penalizes estimated regret against information or estimation error, whereas the certificate treats the one-step inequality as a hard constraint and minimizes exploration subject to it.

Analytic certificates. Bernoulli conjugacy appears in information-ratio analyses (Neu et al., 2022). The closest computational precedent is SquareCB.G: Zhang et al. (2023, Section 3.3) eliminate the adversarial mean optimization analytically and obtain A convex constraints on the allocation. Analytic elimination is therefore not by itself new. What is new here is the structure of the resulting certified set, namely its least element, the uniformly contractive iteration that computes this element while every iterate remains certified, and the exploration laws that bound the achievable reduction from both sides.

Greedy and infrequent exploration. A separate line of work asks when exploration can be dispensed with altogether. Under smoothed or diverse contexts, greedy or mostly greedy policies attain near-optimal regret in linear contextual bandits (Kannan et al., 2018; Raghavan et al., 2018; Bastani et al., 2021; Kim & Oh, 2024), and empirical studies report that greedy baselines are hard to beat (Bietti et al., 2021). Most recently, Lee & Oh (2025) show that in stochastic linear bandits with a fixed arm set, running any base algorithm on a schedule of $\omega(\log T)$ exploratory rounds and acting greedily otherwise preserves the base algorithm's asymptotic regret, and that this frequency is necessary. These results reduce exploration by restricting the environment: they require linear rewards, and either distributional assumptions on the contexts or a fixed feature set with no contexts at all. MinCE addresses the complementary question. It makes no assumption on the contexts, which may be chosen adaptively, and it applies to any function class with a regression oracle; in exchange, exploration is minimized relative to the one-step certificate rather than eliminated. The two viewpoints are also different in kind: infrequent exploration controls in how many rounds a learner explores, whereas the certificate controls how much it explores in each round. Combining them is a natural direction.

B Certificate and regret analysis

This section proves Lemma 1, Theorem 2, and Corollary 3. The sampling computation is analyzed in the subsequent appendices. All statements about that computation use exact real arithmetic.

B.1 The certificate identity and strict reference feasibility

Fix a round and write $y = \widehat{y}_t \in (0,1)^A$, $b = b_t$, and $q = q_t$. In this proof only, omit the round index and set $r = (1 - 5A/\gamma)/(14\gamma)$. The quantities U_i and D_i are those in (6), so

$$C_y(p) = 14\gamma \left[\sum_i p_i U_i - \min_i D_i(p_i) \right], \qquad D_i(0) = 0.$$

Proof of the identity and regularity in Lemma 1. The Bernoulli conjugacy identity is

$$\max_{z \in [0,1]} \{sz - \mathrm{KL}(z \| y)\} = \log(1 - y + ye^s), \qquad y \in (0,1),\ s \in \mathbb{R}. \tag{14}$$

Differentiation gives the unique maximizer $z = ye^s/(1 - y + ye^s)$; substitution proves the formula.

In the variational expression (5), write $-\min_j f_j = \max_j(-f_j)$ and interchange the two maxima. For a fixed comparator j, maximization separates across coordinates. If $p_i > 0$, apply (14) with

$$s_i = \frac{(1 - 5A/\gamma)p_i - \mathbf{1}\{i = j\}}{14\gamma p_i}.$$

The resulting value for comparator j is

$$14\gamma \sum_i p_i U_i - 14\gamma D_j(p_j).$$

If $p_j = 0$, the comparator coordinate contributes $\max_{f_j \in [0,1]}(-f_j) = 0$, as specified by $D_j(0) = 0$. Every other zero-mass coordinate also contributes zero. Maximizing over j proves (5) on the entire closed simplex.

For each fixed f, the expression maximized in (5) is affine in p, so its maximum is convex. Because the prediction vector is interior, the objective is jointly continuous on the compact set $\Delta_A \times [0, 1]^A$. Its maximum is therefore continuous in p. In particular, the convention at $p_i = 0$ is the continuous extension of the computable formula. This proves all claims of the lemma except strict feasibility, proved next. □

Strict feasibility of the reference allocation. The reference distribution is strictly feasible, not merely on the boundary of the certificate set. This fact supplies the initial bracket used by the allocation computation.

Proposition 7 (Strict reference feasibility)**.** *For every $y \in (0, 1)^A$ and $\gamma \geq 10A$, the reference distribution q in* (3) *satisfies $C_y(q) < 0$.*

Proof. Theorem 4 of Foster & Krishnamurthy (2021) gives, for every $f \in [0, 1]^A$,

$$q^\top f - \min_i f_i \leq \frac{5A}{\gamma} q^\top f + 7\gamma \sum_i q_i \frac{(y_i - f_i)^2}{y_i + f_i}. \tag{15}$$

Its requirement $\gamma \geq 2A$ is satisfied here. For $y \in (0, 1)$ and $z \in [0, 1]$ with $z \neq y$, we have the strict scalar inequality

$$\frac{(z - y)^2}{z + y} < 2\,\mathrm{KL}(z\|y). \tag{16}$$

Indeed, $g(z) = \mathrm{KL}(z\|y)$ satisfies $g(y) = g'(y) = 0$ and $g''(w) = 1/[w(1 - w)]$. At every interior point between z and y, $w(1 - w) < w \leq \max\{z, y\} \leq z + y$. Taylor's integral remainder then gives $g(z) > (z - y)^2/[2(z + y)]$. At $z = 0$ or $z = 1$, the same argument follows by an improper integral, with the strict remainder already present on an interior subinterval.

All coordinates of q are positive. For $f \neq y$, combining (15) and (16) yields

$$\begin{aligned}(1 - 5A/\gamma)q^\top f - \min_i f_i - 14\gamma \sum_i q_i\,\mathrm{KL}(f_i\|y_i) \\ \leq 7\gamma \sum_i q_i \left[\frac{(y_i - f_i)^2}{y_i + f_i} - 2\,\mathrm{KL}(f_i\|y_i)\right] < 0.\end{aligned}$$

For the remaining vector $f = y$, let $h = y_b > 0$ and $\Delta_i = y_i - h$. The reference formula gives

$$q^\top y - h = \sum_{i \neq b} \frac{h\Delta_i}{Ah + \gamma\Delta_i} \leq \frac{(A - 1)h}{\gamma}.$$

Since $q^\top y \geq h$, it follows that

$$(1 - 5A/\gamma)q^\top y - h \leq \frac{(A - 1)h}{\gamma} - \frac{5Ah}{\gamma} = -\frac{(4A + 1)h}{\gamma} < 0.$$

Thus the continuous envelope objective is negative at every $f \in [0, 1]^A$. Compactness implies that its attained maximum is strictly negative, proving the claim and completing Lemma 1. □

B.2 From the online oracle to a predictable KL budget

The history $\mathcal{H}_t$ contains the current context, predictions, and allocation, but precedes the current action draw. Realizability and the pre-action protocol imply

$$\mathbb{E}[\ell_t(a_t) \mid \mathcal{H}_t, a_t = a] = f^*(x_t, a) = f^*_{t,a}.$$

No independence assumption is imposed on the context sequence. The oracle assumption (2) (Assumption 2), evaluated at f^* and then averaged, gives

$$\mathbb{E}\sum_{t=1}^{T}\left[\ell_{\log}(\widehat{y}_{t,a_t}, \ell_t(a_t)) - \ell_{\log}(f^*_{t,a_t}, \ell_t(a_t))\right] \le \mathrm{Reg}_{\mathrm{KL}}(T).$$

The expected version of the oracle assumption gives the same inequality. Boundary comparator means use the usual log-loss conventions; realizability excludes the corresponding impossible outcome almost surely.

Lemma 8 (Predictable KL budget)**.** *Under Assumptions 1 and 2,*

$$\mathbb{E}\sum_{t=1}^{T}\sum_{a=1}^{A} p_{t,a}\,\mathrm{KL}(f^*_{t,a}\|\widehat{y}_{t,a}) \le \mathrm{Reg}_{\mathrm{KL}}(T).$$

This holds for every $[0,1]$-valued loss, including nonbinary outcomes.

Proof. For $Y \in [0,1]$ with mean z, the affine dependence of cross-entropy on its second argument gives

$$\mathbb{E}[\ell_{\log}(y, Y) - \ell_{\log}(z, Y)] = z\log(z/y) + (1-z)\log((1-z)/(1-y)) = \mathrm{KL}(z\|y).$$

Apply this identity conditionally on $(\mathcal{H}_t, a_t)$, average over the action, and sum over rounds. The preceding oracle inequality proves the claim. □

B.3 Regret and calibration

Proof of Theorem 2. The finite allocation invariant in Theorem 5 gives $C_t(p_t) \le 0$ for the distribution actually sampled, at every round. This is a proved property of the algorithm, not an additional assumption. Substitute $f = f^*_t$ in (5) to obtain, conditionally on $\mathcal{H}_t$,

$$p_t^\top f^*_t - \min_a f^*_{t,a} \le \frac{5A}{\gamma} p_t^\top f^*_t + 14\gamma \sum_a p_{t,a}\,\mathrm{KL}(f^*_{t,a}\|\widehat{y}_{t,a}).$$

For brevity in this proof, write $\mathcal{R}_T = \mathbb{E}[\mathrm{Reg}_{\mathrm{CB}}(T)]$ and $\mathcal{L}^*_T = \mathbb{E}L^*_T$. The learner's expected cumulative mean loss is $\mathcal{L}^*_T + \mathcal{R}_T$. Taking expectations, summing, and using Lemma 8 therefore gives

$$\mathcal{R}_T \le \frac{5A}{\gamma}(\mathcal{L}^*_T + \mathcal{R}_T) + 14\gamma\,\mathrm{Reg}_{\mathrm{KL}}(T).$$

Since $1 - 5A/\gamma > 0$, rearranging proves (10). Each one-step statement conditions on the current context, so this argument permits contexts chosen adaptively from the preceding history.

For the calibrated bound, put $B = \mathrm{Reg}_{\mathrm{KL}}(T) > 0$ and $x = \sqrt{\bar{L}/(AB)}$. Let F denote the right side of (10), with $\mathcal{L}^*_T$ replaced by $\bar{L}$ and γ chosen in Theorem 2. Direct substitution gives

$$\frac{F}{AB} = \begin{cases} x^2 + 280, & 0 \le x \le 10\sqrt{3}, \\ \dfrac{29x^2}{\sqrt{3}(x - 5\sqrt{3})}, & x \ge 10\sqrt{3}. \end{cases}$$

In the first range, $x^2 \le 10\sqrt{3}\,x$. In the second,

$$10\sqrt{3}\,x + 280 - \frac{29x^2}{\sqrt{3}(x - 5\sqrt{3})} = \frac{(x - 10\sqrt{3})(x + 140\sqrt{3})}{\sqrt{3}(x - 5\sqrt{3})} \ge 0.$$

Consequently, $F \le 10\sqrt{3}\sqrt{A\bar{L}B} + 280AB$, which is (11). □

The two expressions agree at $x = 10\sqrt{3}$, so $10\sqrt{3}$ is the smallest square-root coefficient compatible with additive coefficient 280 for this rational upper bound and calibration. This sharpens a shared accounting step: (15) also implies the same regret bound for FastCB. These constants alone do not establish a regret separation between the two algorithms.

Proof of Corollary 3. Since $0 \le L_T^* \le T$, choose $\bar{L} = T$ in Theorem 2. Also, $\mathrm{Reg}_{\mathrm{CB}}(T) \le T$ on every sample path. Combining the bounds proves (12). For a finite class of $M \ge 2$ predictors, the mixture construction in Lemma 9 below gives $\mathrm{Reg}_{\mathrm{KL}}(T) = \log M$. If $T \ge A \log M$, then $A \log M \le \sqrt{AT \log M}$ and the additive term is absorbed into the square-root term. If $T < A \log M$, use the bound T instead. This proves (13). □

Scope of the minimax comparison. The lower bound of Agarwal et al. (2012, Theorem 5.1) constructs realizable finite regression classes with A actions and IID contexts. Replacing rewards by one minus the reward gives our loss convention without changing regret. The predictors can be interior-valued. For $T \ge \log M / \log A$, the lower bound is $\Omega(\sqrt{AT \log M / \log A})$ for a class of cardinality at most M. It is a worst-case statement over finite classes, not a lower bound for every dictionary. Thus (13) is minimax optimal for fixed A, with a $\sqrt{\log A}$ gap to this lower bound when A varies. An arbitrary supplied oracle bound $\mathrm{Reg}_{\mathrm{KL}}(T)$ does not by itself specify an information-theoretic lower bound.

B.4 A FINITE-DICTIONARY ORACLE

Let $\mathcal{F} = \{f_1, \dots, f_M\}$ consist of arbitrary maps from context–action pairs to $(0, 1)$. Initialize $w_{1,j} = 1/M$ and use

$$\widehat{y}_t(x, a) = \sum_{j=1}^{M} w_{t,j} f_j(x, a),$$

$$w_{t+1,j} = \frac{w_{t,j} \exp[-\ell_{\log}(f_j(x_t, a_t), \ell_t(a_t))]}{\sum_k w_{t,k} \exp[-\ell_{\log}(f_k(x_t, a_t), \ell_t(a_t))]}.$$

Only the selected action's observed loss enters the update. The proof also allows boundary-valued experts, provided an interior expert retains positive weight so that the mixture predictions remain interior.

Lemma 9 (Log-loss mixture oracle)**.** *For every adaptive sequence of selected examples, the mixture satisfies*

$$\sum_t \ell_{\log}(\widehat{y}_t(x_t, a_t), \ell_t(a_t)) - \min_{j \le M} \sum_t \ell_{\log}(f_j(x_t, a_t), \ell_t(a_t)) \le \log M.$$

Proof. For $\ell \in [0, 1]$, the function $y \mapsto e^{-\ell_{\log}(y,\ell)} = y^{\ell}(1-y)^{1-\ell}$ is concave. For $\ell = 0, 1$ it is affine; otherwise its second derivative is $-\ell(1-\ell)y^{\ell-2}(1-y)^{-\ell-1} \le 0$. Jensen's inequality gives

$$\ell_{\log}\left(\sum_j w_{t,j} f_j(x_t, a_t), \ell_t(a_t)\right) \le -\log \sum_j w_{t,j} e^{-\ell_{\log}(f_j(x_t,a_t),\ell_t(a_t))}.$$

Under the weight update, the right side telescopes. The cumulative loss is at most

$$-\log\left[\frac{1}{M}\sum_j \exp\left(-\sum_t \ell_{\log}(f_j(x_t, a_t), \ell_t(a_t))\right)\right]$$
$$\le \log M + \min_j \sum_t \ell_{\log}(f_j(x_t, a_t), \ell_t(a_t)).$$

The argument is pathwise and therefore permits adaptive examples. □

This construction supplies a fully specified oracle bound for arbitrary finite nonlinear dictionaries. It does not assert a logarithmic online regret bound for neural networks trained by nonconvex stochastic gradient descent.

C THE SIMULTANEOUS MINIMUM OF CERTIFIED EXPLORATION

Compact notation for the allocation proofs. The main text keeps the round index explicit. In the allocation proofs only, we fix one round t and use action indices $i, j \in [A]$, with the dictionary

$$y_i = \widehat{y}_{t,i}, \qquad b = b_t, \qquad q_i = q_{t,i}, \qquad U_i = U_{t,i}, \qquad D_i = D_{t,i},$$
$$H_i = H_{t,i}, \qquad G = G_t, \qquad C_y = C_t, \qquad h = y_b = \widehat{y}_{t,b_t}.$$

Thus every fixed-vector allocation result applies separately at every round, regardless of how the contexts or predictions change over time. We also use the proof abbreviation $\rho = 5A/\gamma$; it is not an algorithm input.

Fix $y \in (0,1)^A$, $A \geq 2$, $b = \min \operatorname{argmin}_i y_i$, and $\gamma \geq 10A$, and set $r = (1-\rho)/(14\gamma)$. Throughout this section, "optimal" refers to the fixed certificate $C_y(p) \leq 0$, not to an information-theoretic optimum over all contextual-bandit algorithms. Set

$$z_q = \sum_i q_i U_i, \qquad \Delta_i^U = U_i - U_b \geq 0, \qquad H_i(z) = D_i^{-1}(z) \quad (0 \leq z \leq z_q).$$

The inverse is well-defined by the next lemma. Define

$$G(z) = U_b + \sum_{i \neq b} \Delta_i^U H_i(z), \qquad F(z) = G(z) - z.$$

Lemma 10 (Monotonicity, concavity, and inverse domain)**.** *For each i, D_i is continuous on $[0,\infty)$, strictly increasing, and strictly concave on $(0,\infty)$. In particular, H_i is increasing, continuous, and convex. Moreover,*

$$0 < H_i(z_q) < q_i \qquad (i \in [A]).$$

Proof. Write $g_i(x) = \log(1 - y_i + y_i e^x)$ and $\zeta = 1/(14\gamma s)$. For $s > 0$, differentiation gives

$$D_i'(s) = g_i(r) - g_i(r-\zeta) - \zeta g_i'(r-\zeta) = \int_{r-\zeta}^{r} \{g_i'(x) - g_i'(r-\zeta)\}\, dx > 0,$$
$$D_i''(s) = -\frac{1}{(14\gamma)^2 s^3} g_i''(r-\zeta) < 0.$$

Continuity at zero follows from $D_i(s)/s \to U_i - \log(1-y_i) < \infty$. Strict feasibility $C_y(q) < 0$, proved in Appendix B.1, is equivalent to $z_q < \min_i D_i(q_i)$. Since $z_q > 0$ and $D_i(0) = 0$, inversion on $[0, z_q]$ is legitimate and gives the displayed strict inequalities. □

C.1 PROOF OF THE SIMULTANEOUS EXPLORATION MINIMUM

We prove Theorem 4, including strict improvement over q and tightness of every nongreedy certificate constraint.

Proof. The function F is convex because each H_i is convex and each Δ_i^U is nonnegative. Also

$$F(0) = U_b > 0, \qquad F(z_q) = \sum_{i \neq b} \Delta_i^U \{H_i(z_q) - q_i\} \leq 0.$$

If at least one gap is positive, the last inequality is strict. Convexity then implies there is exactly one root on $[0, z_q]$: two roots before z_q would give a nonnegative secant slope after the second root, contradicting $F(z_q) < 0$. If all gaps are zero, $F(z) = U_b - z$ and the unique root is $U_b = z_q$. Since $G(z) \geq U_b$, the root lies in $[U_b, z_q]$.

Define $p_i^{\mathrm{opt}} = H_i(z_\star)$ for $i \neq b$ and $p_b^{\mathrm{opt}} = 1 - \sum_{i \neq b} H_i(z_\star)$. By the inverse lemma, $H_i(z_\star) < q_i$ for every i, so $p_b^{\mathrm{opt}} > q_b > 0$ and $p^{\mathrm{opt}} \in \Delta_A$. The root identity gives $\sum_i p_i^{\mathrm{opt}} U_i = z_\star$. For nongreedy coordinates the corresponding D_i equals $z_\star$; for the greedy coordinate, $D_b(p_b^{\mathrm{opt}}) > D_b(q_b) > z_q \geq z_\star$. Thus $C_y(p^{\mathrm{opt}}) = 0$: every nongreedy certificate constraint is tight, whereas the greedy constraint is strict.

Now let p be any certified distribution and let $z = p^\top U$. Certification means $D_i(p_i) \geq z$ for every i. Suppose $z < z_\star$. Then $z < z_q$, so $p_i \geq H_i(z)$ and

$$z = U_b + \sum_{i \neq b} \Delta_i^U p_i \geq U_b + \sum_{i \neq b} \Delta_i^U H_i(z) = G(z) > z,$$

where $F(z) > 0$ holds below its unique root. This is impossible. Thus $z \geq z_\star$, and monotonicity of D_i now directly implies $p_i \geq H_i(z_\star)$, without requiring $H_i(z)$ to be defined above z_q. The coordinatewise statement proves all the exploration objectives in Theorem 4. Equality in the unweighted sum forces equality in every nongreedy coordinate and hence in the greedy coordinate, proving uniqueness. □

D FINITE COMPUTATION OF THE MINIMUM EXPLORATION ALLOCATION

The exact inverses in the preceding theorem are not primitive operations. We now give a finite algorithm with internal accuracy fixed by the horizon, and prove its feasibility, relative optimality, and termination. No numerical tolerance or iteration count is supplied by the user. We use the fixed-round notation of Appendix C. First we bound the contraction rate of the shared scalar map. We then control inverse sensitivity and finite stopping. All computational statements in this section concern exact real arithmetic.

D.1 CONTRACTION AND FEASIBILITY

Theorem 11 (Uniformly contractive, feasible allocation updates)**.** *On $[U_b, z_q]$, G is nondecreasing and*

$$0 \leq G'(z) \leq \frac{A-1}{112A} < \frac{1}{112}.$$

Starting from $p^{(0)} = q$, form $z_n = (p^{(n)})^\top U$ and update

$$p_i^{(n+1)} = H_i(z_n) \quad (i \neq b), \qquad p_b^{(n+1)} = 1 - \sum_{i \neq b} p_i^{(n+1)}.$$

Every iterate is certified, every nongreedy probability decreases, and $p^{(n)} \to p^{\mathrm{opt}}$. In particular,

$$0 \leq z_n - z_\star \leq 112^{-n}(z_q - z_\star).$$

The same feasibility and coordinatewise monotonicity hold if each inverse is replaced by any upper approximation in $[H_i(z_n), p_i^{(n)}]$.

Proof. We first prove the uniform rate. Introduce the tilted Bernoulli means

$$w_i = \frac{y_i e^r}{1 - y_i + y_i e^r}, \qquad \kappa = \frac{A-1}{\gamma}, \qquad \delta = 1 - (1-\rho)(1 + e^r \kappa).$$

The maps $y \mapsto w(y)$ and $y \mapsto U(y)$ are increasing. Since $g_i(r) = \int_0^r g_i'(x)\,dx \leq r w_i$ and $w_i - w_b \leq e^r(y_i - y_b)$, FastCB's gap bound yields

$$14\gamma z_q \leq (1-\rho)\sum_i q_i w_i \leq (1-\rho)(1+e^r\kappa) w_b = (1-\delta) w_b.$$

Here $w_b \geq y_b$ was used. We have $r \leq \rho$, so $(1-\rho)e^r \leq 1$. Therefore

$$\frac{4A}{\gamma} \leq \rho - \kappa \leq \delta \leq \rho \leq \frac{1}{2}.$$

For $s = H_i(z) > 0$, let $\zeta = 1/(14\gamma s)$ and

$$\phi_i(\zeta) = -\log(1 - w_i + w_i e^{-\zeta}), \qquad m_i = \phi_i'(\zeta) = \frac{w_i e^{-\zeta}}{1 - w_i + w_i e^{-\zeta}}.$$

Then $14\gamma z = \phi_i(\zeta)/\zeta \geq m_i$, while direct calculation gives

$$D_i'(s) = \phi_i(\zeta) - \zeta\phi_i'(\zeta) = \mathrm{KL}(m_i \| w_i) \geq \frac{(w_i - m_i)^2}{2w_i} \geq \frac{(w_i - 14\gamma z)^2}{2w_i}.$$

The KL lower bound follows by integrating its second derivative $1/[x(1-x)] \geq 1/w_i$ on the interval $[m_i, w_i]$. Viewing U as a function of w gives $dU/dw \leq e^r - 1$, hence $\Delta_i^U \leq (e^r - 1)(w_i - w_b)$. Consequently

$$\begin{aligned} G'(z) &\leq 2(e^r - 1) \sum_{i \neq b} \frac{(w_i - w_b)w_i}{\{w_i - (1-\delta)w_b\}^2} \\ &\leq \frac{(A-1)(e^r - 1)}{2\delta(1-\delta)}. \end{aligned}$$

The second line uses, for $x \geq 0$ and $0 < \delta \leq 1/2$,

$$\frac{x(1+x)}{(x+\delta)^2} \leq \frac{1}{4\delta(1-\delta)};$$

after clearing denominators the difference is $\{(1-2\delta)x - \delta\}^2 \geq 0$. Finally, $e^r - 1 \leq re^r$ and $1 - \delta = (1-\rho)(1 + e^r\kappa)$ imply

$$\frac{(A-1)(e^r-1)}{2\delta(1-\delta)} \leq \frac{(A-1)e^r}{28\gamma\delta(1+e^r\kappa)} \leq \frac{A-1}{28\gamma\delta} \leq \frac{A-1}{112A}.$$

For the middle inequality, $\kappa \geq r \geq 1 - e^{-r}$, so $e^r/(1+e^r\kappa) \leq 1$.

For feasibility, suppose $p^{(n)}$ is certified. Then $D_i(p_i^{(n)}) \geq z_n$, so the stated exact or upper-approximate update decreases every nongreedy coordinate. Because $U_b \leq U_i$, this gives $z_{n+1} \leq z_n$; the greedy coordinate increases. Each nongreedy coordinate after the update has $D_i \geq z_n \geq z_{n+1}$, and the greedy one has $D_b(p_b^{(n+1)}) \geq D_b(p_b^{(n)}) \geq z_n$. Thus the update is certified. For exact inverses $z_{n+1} = G(z_n)$, and the uniform derivative bound proves the geometric convergence. □

D.2 Inverse sensitivity and reusable brackets

The contraction proof defines δ. Put $W = z_q - U_b$.

Lemma 12 (Inverse sensitivity on the feasible interval)**.** *For $i \in [A]$ and $z \in [U_b, z_q]$,*

$$\frac{zH_i'(z)}{H_i(z)} \leq \frac{1}{\delta}, \qquad \frac{W}{\delta U_b} < \frac{1}{3}.$$

Consequently, whenever $U_b \leq l \leq u \leq z_q$, $H_i(u)/H_i(l) \leq \exp\{(u-l)/(\delta U_b)\}$.

Proof. Let $s = H_i(z)$, $\zeta = 1/(14\gamma s)$, and $w_i(x) = \mathrm{sigmoid}(\mathrm{logit}(y_i) + r - x)$; thus $w_i(0) = w_i$. Writing $d = 14\gamma z$ gives

$$d = \frac{1}{\zeta}\int_0^\zeta w_i(x)\,dx, \qquad D_i'(s) = \zeta\{d - w_i(\zeta)\}.$$

The function $\log w_i(x)$ is concave, since its second derivative is $-w_i(x)(1-w_i(x))$. Its graph lies above the chord between zero and ζ. Integrating the resulting geometric interpolation and using the logarithmic-mean/geometric-mean inequality yields $d \geq \sqrt{w_i w_i(\zeta)}$. Therefore

$$\frac{sD_i'(s)}{D_i(s)} = 1 - \frac{w_i(\zeta)}{d} \geq 1 - \frac{d}{w_i} \geq \delta,$$

where the last inequality follows from $d \leq 14\gamma z_q \leq (1-\delta)w_b \leq (1-\delta)w_i$. Inverting proves the first assertion.

For the second, let $h = y_b$ and $\kappa = (A-1)/\gamma$. FastCB's gap bound and $U_b \geq rh$ imply

$$W \leq (e^r - 1)h\kappa \leq e^r\kappa U_b.$$

Since $\delta \geq 4A/\gamma$ and $r \leq 1/(280A) \leq 1/560$, $W/(\delta U_b) \leq e^r/4 < 1/3$. Finally, integrating $H_i'/H_i \leq 1/(\delta U_b)$ proves the ratio bound. □

Lemma 13 (Rational sensitivity and reusable inverse brackets)**.** *Let $w_i = y_i e^{r-U_i}$ be the tilted Bernoulli mean. For $z \in [U_b, z_q]$ and $U_b \le l \le u \le z_q$,*

$$\frac{H_i'(z)}{H_i(z)} \le \frac{w_i}{z(w_i - 14\gamma z)}, \qquad H_i(l) \ge H_i(u)\frac{l}{u}\frac{w_i - 14\gamma u}{w_i - 14\gamma l}. \tag{17}$$

Consequently, if $a_i \le H_i(u) \le c_i$, then a bracket at the smaller target l is

$$a_i \frac{l}{u}\frac{w_i - 14\gamma u}{w_i - 14\gamma l} \;\le\; H_i(l) \;\le\; c_i \frac{l}{u}. \tag{18}$$

Its endpoints require only arithmetic on cached quantities.

Proof. The proof of Lemma 12 gives the sharper inequality $sD_i'(s)/D_i(s) \ge 1 - 14\gamma z/w_i$ at $s = H_i(z)$, before replacing its right side by δ. Inverting yields the derivative bound in (17). All denominators are positive, since $14\gamma z_q \le (1-\delta)w_b < w_b \le w_i$. Integration gives

$$\log \frac{H_i(u)}{H_i(l)} \le \int_l^u \frac{w_i\, dz}{z(w_i - 14\gamma z)} = \log \frac{u(w_i - 14\gamma l)}{l(w_i - 14\gamma u)},$$

which proves the lower bracket. For the upper bracket, convexity of H_i and $H_i(0) = 0$ imply $H_i(l) \le (l/u)H_i(u) \le (l/u)c_i$. □

D.3 The finite allocation routine

Set $\tau = 1/(16T)$ and $\lambda = (A-1)/(112A)$ internally, and initialize $p = q$, $l = U_b$. At each sweep let $u = p^\top U$. For each $i \ne b$, compute just one bracket

$$a_i \le H_i(u) \le c_i \le p_i, \qquad c_i - a_i \le \tau a_i.$$

Form the lower estimate of the fixed-point map and update the lower target:

$$g^- = U_b + \sum_{i \ne b} \Delta_i^U a_i, \qquad l \leftarrow \max\left\{ l, \frac{g^- - \lambda u}{1-\lambda} \right\}.$$

The contraction residual supplies a lower bound on $z_\star$, eliminating all inverse solves at a second target. A lower bound on each optimal coordinate is then obtained without a transcendental evaluation:

$$L_i = a_i \frac{l}{u}\frac{w_i - 14\gamma u}{w_i - 14\gamma l}.$$

Set $p_i' = c_i$ for $i \ne b$ and $p_b' = 1 - \sum_{i \ne b} c_i$. Return p' if

$$c_i \le (1 + 1/T)L_i \qquad \text{for every } i \ne b;$$

otherwise replace p by p' and repeat. An exploration lower-bound diagnostic is $\sum_{i \ne b} L_i$; the sum of the inverse lower endpoints a_i at the upper target need not lower-bound the optimum.

Within this round, (18) transports the previous bracket to each new, smaller target before refinement. Its upper endpoint is no larger than the current allocation cap. The lower endpoint may also be tightened by z/Λ_i, where $\Lambda_i = U_i - \log(1-y_i) = -\log(1-w_i)$. Algorithm 3 gives the complete subroutine, including a saturation check that can avoid root refinements. Brackets are reset when the round's predictions change. Safeguarded Newton steps with rational upper bounds accelerate the subroutine, while ensuring each refinement round either meets the stopping condition or halves its preceding width. These computations introduce no public parameter.

Theorem 14 (Finite approximation of the certified exploration minimum)**.** *The preceding routine uses no parameter beyond the existing γ, T and terminates after at most*

$$N_T = 1 + \lceil \log_{100} T \rceil$$

sweeps. Every intermediate allocation is certified and every nongreedy probability is nonincreasing. The returned allocation $\widehat{p}$ obeys

$$\begin{gathered} C_y(\widehat{p}) \le 0, \qquad p_i^{\mathrm{opt}} \le \widehat{p}_i \le \min\{q_i, (1+1/T)p_i^{\mathrm{opt}}\} \quad (i \ne b), \\ 1 - p_b^{\mathrm{opt}} \le 1 - \widehat{p}_b \le (1+1/T)(1-p_b^{\mathrm{opt}}). \end{gathered} \tag{19}$$

Each sweep solves only one inverse problem for each $i \neq b$. Each such problem requires at most

$$M_i = \left\lceil \log_2\left(\frac{(64T+1)q_i\Lambda_i}{U_b}\right)\right\rceil, \qquad \Lambda_i = U_i - \log(1-y_i), \tag{20}$$

bracket refinement rounds. Each round needs a constant number of evaluations of D_i and, if accelerated, its derivative. Thus at most $N_T \sum_{i\neq b} M_i$ scalar refinement rounds are needed, in addition to $\mathcal{O}(AN_T)$ time operations. For $T = 5000$, $N_T = 3$.

Proof. We first verify feasibility and monotonicity. If p is certified, then $D_i(p_i) \geq u$, so the upper caps are valid. Their refined upper endpoints satisfy $H_i(u) \leq c_i \leq p_i$. Every nongreedy coordinate therefore decreases, the greedy coordinate increases, and $u' = (p')^\top U \leq u$. For $i \neq b$, $D_i(p_i') \geq u \geq u'$; for b, $D_b(p_b') \geq D_b(p_b) \geq u \geq u'$. This proves certification without any positive error allowance. It also implies $u \geq z_\star$ by Theorem 4.

The updated lower target remains valid because

$$g^- \leq G(u) \leq G(z_\star) + \lambda(u - z_\star) = (1-\lambda)z_\star + \lambda u.$$

Hence $l \leq z_\star$. Lemma 13 now gives $L_i \leq H_i(l) \leq H_i(z_\star) = p_i^{\mathrm{opt}}$. The stopping condition, together with feasibility and $c_i \leq q_i$, proves (19).

For termination, first consider $W = 0$. Then $u = z_\star = l = U_b$ and $L_i = a_i$, so $c_i/L_i \leq 1 + \tau \leq 1 + 1/T$ and the first sweep stops. Assume henceforth $W > 0$. Let u_j denote the upper target after j updates, $e_j = u_j - z_\star$, and $s_\star = z_\star - U_b \leq W$. Since $c_i \leq (1+\tau)H_i(u_j)$, contraction gives

$$e_{j+1} \leq (1+\tau)\lambda e_j + \tau s_\star.$$

Put $a = (1+\tau)\lambda < 1/100$ and $m = \lceil \log_{100} T \rceil$. Since $e_0 \leq W$ and $a^m \leq 1/T = 16\tau$, including $m = 0$ when $T = 1$,

$$e_m \leq 16\tau W + \frac{\tau W}{1-a} < 18\tau W.$$

The inverse lower endpoints satisfy $a_i \geq H_i(u)/(1+\tau)$, so

$$0 \leq G(u) - g^- \leq \tau\{G(u) - U_b\} \leq \tau(s_\star + \lambda e).$$

Thus the lower target computed in sweep $m+1$ obeys

$$u_m - l \leq \frac{u_m - g^-}{1-\lambda} \leq \frac{(1+\tau\lambda)e_m + \tau s_\star}{1-\lambda} < 20\tau W.$$

The final inequality uses $\tau \leq 1/16$ and $\lambda < 1/112$. The integrand in the rational sensitivity bound is at most $1/(\delta U_b)$ on the feasible interval. Consequently,

$$\frac{c_i}{L_i} \leq (1+\tau)\frac{u_m(w_i - 14\gamma l)}{l(w_i - 14\gamma u_m)} \leq (1+\tau)\exp\left(\frac{u_m - l}{\delta U_b}\right) < \exp(23\tau/3) \leq \exp(8\tau) \leq 1 + \frac{1}{T},$$

where $W/(\delta U_b) < 1/3$ was proved in Lemma 12. The last inequality follows from $\log(1 + 1/T) \geq 1/(T+1) \geq 1/(2T) = 8\tau$. This establishes termination by sweep $m+1$.

Finally, $D_i(s) \leq \Lambda_i s$ gives $H_i(z) \geq U_b/\Lambda_i$ for every target used by the routine. The bracket cap is at most q_i, including after warm transport. After n refinement rounds, either the routine has stopped or its width is at most $q_i 2^{-n}$. If this is at most $(\tau/4)H_i(z)/(1+\tau/4)$, the lower endpoint is at least the root minus the width, and the subroutine's relative stopping condition holds. The value M_i in (20) suffices. Inflating the final upper endpoint by at most $1 + \tau/4$, subject to its valid cap, retains the required bracket contract because $(1+\tau/4)^2 \leq 1 + \tau$. □

Scope of the computational guarantee. The outer sweep bound is uniform over interior predictions. The inner counts M_i depend on the current predictions, and do not imply a uniform $\mathcal{O}(A \log T)$ runtime without additional bounds on prediction values. All displayed guarantees concern finite operations in exact real arithmetic; machine-certified floating-point signs require a separate rounding analysis. Finite inverse approximation itself contributes *zero* certificate violation and hence no additive statistical regret term.

Cumulative allocation accuracy. On the learner's realized prediction sequence, the finite guarantee yields

$$0 \le \sum_{t=1}^{T} \big[(1-\widehat{p}_{t,b_t}) - (1-p^{\mathrm{opt}}_{t,b_t})\big] \le T^{-1} \sum_{t=1}^{T} (1-p^{\mathrm{opt}}_{t,b_t}) \le 1.$$

This is a comparison at identical prediction states, not a comparison of separately trained learners. The first-order regret bound has no residual term because every returned allocation is certified.

D.4 Scalar inverse subroutine

Algorithm 3 completes the MinCE pseudocode. Its accuracy is determined internally by the existing horizon T; no inverse tolerance or iteration count is supplied by the user. We use the compact notation for one fixed round from Appendix C. The action-specific quantities D_i, w_i, Λ_i are cached when that round's predictions are received; γ is the shared exploration parameter.

The subroutine returns $s^- \le H_i(z) \le s^+ \le c$ with $s^+ \le (1+\tau)s^-$. It reuses brackets only between sweeps for this same action and prediction vector. The cache is cleared at the start of every contextual-bandit round. Since the allocation targets decrease within a round, every reuse satisfies $z \le z^{\mathrm{old}}$.

For the pseudocode, define the arithmetic upper enclosure

$$\mathsf{Upper}_i(s,d;z) = s + \frac{s w_i (z-d)}{d(w_i - 14\gamma z)}, \qquad 0 < d = D_i(s) \le z < w_i/(14\gamma).$$

Lemma 15 below proves its validity. The helper $\mathrm{Refine}(s)$ evaluates $d = D_i(s)$ and $g = D_i'(s)$. If $d < z$, it sets $(s^-, d^-, g^-) \leftarrow (s, d, g)$ and $s^+ \leftarrow \min\{s^+, \mathsf{Upper}_i(s,d;z)\}$; otherwise it sets $s^+ \leftarrow s$. Thus it caches the current lower value and derivative and requires one joint function/derivative evaluation.

Algorithm 3 InverseBracket: warm brackets and safeguarded Newton refinement

Input: Cached D_i, w_i, Λ_i; exploration parameter γ; target $z > 0$; certified cap c with $D_i(c) \ge z$; internal $\tau = 1/(16T)$.

$s^- \leftarrow z/\Lambda_i$.
if a bracket $(z^{\mathrm{old}}, a_i^{\mathrm{old}}, c_i^{\mathrm{old}})$ is cached **then**
  $s^- \leftarrow \max\{s^-, a_i^{\mathrm{old}}(z/z^{\mathrm{old}})(w_i - 14\gamma z^{\mathrm{old}})/(w_i - 14\gamma z)\}$.
  $c \leftarrow \min\{c, c_i^{\mathrm{old}}(z/z^{\mathrm{old}})\}$.
end if
$s^+ \leftarrow \min\{c, (1+\tau/4)s^-\}$.
if $D_i(s^+) < z$ **then**
  $(d^-, g^-) \leftarrow (D_i(s^-), D_i'(s^-))$.
  $s^+ \leftarrow \min\{c, \mathsf{Upper}_i(s^-, d^-; z)\}$.
  while $s^+ - s^- > (\tau/4)s^-$ **do**
    $\Delta_{\mathrm{old}} \leftarrow s^+ - s^-$; $s \leftarrow s^- + (z-d^-)/g^-$.
    If $s \notin (s^-, s^+)$, set $s \leftarrow (s^- + s^+)/2$.
    $\mathrm{Refine}(s)$.
    if $s^+ - s^- > \Delta_{\mathrm{old}}/2$ and $s^+ - s^- > (\tau/4)s^-$ **then**
      $\mathrm{Refine}((s^- + s^+)/2)$.
    end if
  end while
end if
$s^+ \leftarrow \min\{c, (1+\tau/4)s^+\}$; cache (z, s^-, s^+) for the next sweep.
return (s^-, s^+).

Why the shortcuts are valid. The inequality $D_i(s) \le \Lambda_i s$ makes z/Λ_i a valid lower endpoint. The reused endpoints follow from Lemma 13 and require no evaluation of D_i. The single check before the loop can certify that a probability within $1 + \tau/4$ of the lower endpoint is already sufficient; this often occurs when the logarithmic correction in D_i is nearly saturated. If it succeeds, no refinement is needed. If it fails, the evaluated lower endpoint supplies the initial rational upper bound. Finally,

the upper-endpoint inflation remains within the proved inverse-width budget: $(1+\tau/4)^2 \le 1+\tau$ for $T \ge 1$.

Lemma 15 (An arithmetic upper enclosure for the scalar inverse)**.** *Write $J_i(s) = 14\gamma D_i(s)$ and let d denote a scaled inverse target. For every $a > 0$ with $0 < j = J_i(a) \le d < w_i$,*

$$J_i^{-1}(d) \le a\frac{d(w_i - j)}{j(w_i - d)} = a + \frac{aw_i(d-j)}{j(w_i - d)}. \tag{21}$$

This enclosure is valid on the entire inverse domain $(0, w_i)$, including when the evaluated value j lies below $14\gamma U_b$.

Proof. For $x > 0$, put

$$\phi_i(x) = -\log(1 - w_i + w_i e^{-x}), \qquad m_i(x) = \phi_i'(x) = \frac{w_i e^{-x}}{1 - w_i + w_i e^{-x}}.$$

With $x = 1/(14\gamma s)$, we have $J_i(s) = \phi_i(x)/x$. Thus J_i is strictly increasing from zero to w_i as s ranges over $(0, \infty)$. The function $\log m_i(x)$ is concave on $[0, \infty)$. Its chord bound, followed by the logarithmic-mean/geometric-mean inequality, gives

$$J_i(s) = \frac{1}{x}\int_0^x m_i(v)\,dv \ge \frac{1}{x}\int_0^x w_i^{1-v/x} m_i(x)^{v/x}\,dv \ge \sqrt{w_i m_i(x)}.$$

Consequently, for every $s > 0$,

$$\frac{sJ_i'(s)}{J_i(s)} = 1 - \frac{m_i(x)}{J_i(s)} \ge 1 - \frac{J_i(s)}{w_i}.$$

After inversion,

$$\frac{(J_i^{-1})'(v)}{J_i^{-1}(v)} \le \frac{w_i}{v(w_i - v)} \qquad (0 < v < w_i).$$

Integrating from j to d and using $J_i^{-1}(j) = a$ proves (21). The argument uses no lower restriction on j beyond positivity. □

Safeguarded Newton acceleration. Algorithm 3 and the compiled implementation use Newton lower updates and the arithmetic upper enclosure (21). The pseudocode uses D_i and z; the compiled implementation uses $J_i = 14\gamma D_i$ and $d = 14\gamma z$. At an evaluated lower endpoint a with $j = J_i(a) \le d$, concavity makes $a + (d-j)/J_i'(a)$ a lower inverse bound. The rational formula supplies an upper inverse bound using the same cached a, j. Intersecting this bound with the current upper endpoint requires no additional evaluation of J_i. In particular, the general refinement loop does not evaluate J_i at the cap or use chord interpolation.

Each proposed point is restricted to the bracket and checked by evaluating J_i and J_i'. A new lower endpoint immediately yields a tighter arithmetic upper bound. If the round has neither satisfied the relative-width test nor halved its preceding width, a midpoint refinement enforces halving. Each round therefore uses at most two joint function/derivative evaluations and retains the same bracket contract and finite refinement bound proved in Theorem 14. The acceleration adds no tolerance, iteration limit, or other user parameter. These guarantees concern exact arithmetic; numerical certificate checks and the recorded fallback remain necessary in the floating-point code.

D.5 A UNIFORM COORDINATEWISE CHARACTERIZATION

The inverse sensitivity estimate also yields a coordinatewise description that holds throughout the interior prediction domain. This is an immediate consequence of the preceding analysis, rather than a new allocation routine.

Proposition 16 (A scalar-inverse benchmark for each nongreedy action)**.** *Fix $A \ge 2$, $y \in (0,1)^A$, $b = \min \operatorname{argmin}_i y_i$, and $\gamma \ge 10A$, with U_i, D_i, H_i defined as above. Then the exact minimum-exploration allocation and the finite MinCE allocation satisfy, for every $i \ne b$,*

$$H_i(U_b) \le p_i^{\mathrm{opt}} \le 1.285\, H_i(U_b), \qquad H_i(U_b) \le \widehat{p}_i \le 1.285(1 + 1/T)H_i(U_b).$$

For fixed A, γ, each benchmark $H_i(U_b)$ depends only on the greedy prediction $h = y_b$ and the prediction y_i of that action.

Proof. Use $z_q = q^\top U$, $W = z_q - U_b$, and δ from Theorem 11 and Lemma 12. Since $U_b \le z_\star \le z_q$ and H_i is increasing, the lower bound follows from $p_i^{\mathrm{opt}} = H_i(z_\star)$. The inverse sensitivity estimate on this same interval gives

$$\frac{p_i^{\mathrm{opt}}}{H_i(U_b)} \le \exp\left(\frac{z_\star - U_b}{\delta U_b}\right) \le \exp\left(\frac{W}{\delta U_b}\right).$$

For completeness, the proof of Lemma 12 establishes $W \le e^r \kappa U_b$, where $\kappa = (A-1)/\gamma$, while the contraction proof gives $\delta \ge 4A/\gamma$. Also, writing $x = \gamma/A \ge 10$,

$$r = \frac{1 - 5A/\gamma}{14\gamma} = \frac{x-5}{14Ax^2} \le \frac{1}{280A} \le \frac{1}{560}.$$

The first inequality is equivalent to $(x-10)^2 \ge 0$. Consequently,

$$\frac{W}{\delta U_b} \le e^r \frac{A-1}{4A} < \frac{e^{1/560}}{4}.$$

Since $\exp(e^{1/560}/4) < 1.285$, this proves the first bound. The finite-allocation bound follows from Theorem 14. Finally, U_b depends only on h, A, γ, while the function D_i depends only on y_i, A, γ. Thus $H_i(U_b)$ depends only on h, y_i, A, γ. □

The proposition describes each optimal exploration probability through one scalar inverse, with a uniform multiplicative factor below 1.285; no boundary limit is required. The vector formed by assigning the nongreedy coordinates $H_i(U_b)$ and the remaining mass to b is *not necessarily certified*. MinCE therefore retains its coupled allocation routine. These benchmarks are used only to interpret the allocation and introduce no algorithm input or additional tuning.

E When can certified exploration improve in order?

This section gives a necessary exploration bound for every distribution satisfying the same one-step certificate as MinCE. It does not impose a parametric model, require a fixed context, or optimize over a function class. Its optimality statements concern this certificate class, rather than all contextual-bandit algorithms.

Fix $A \ge 2$, $y \in (0,1)^A$, $b \in \operatorname{argmin}_i y_i$, $h = y_b$, $\gamma \ge 10A$, and $\rho = 5A/\gamma$. Write $\Delta_i = y_i - h$ and

$$q_i = \frac{h}{Ah + \gamma\Delta_i} \quad (i \ne b), \qquad q_b = 1 - \sum_{i \ne b} q_i.$$

The feasible set is

$$\mathcal{P}_y = \{p \in \Delta_A : C_y(p) \le 0\},$$

where the exact certificate has the variational identity

$$C_y(p) = \sup_{f \in [0,1]^A} \left\{ (1-\rho)p^\top f - \min_j f_j - 14\gamma \sum_j p_j \,\mathrm{KL}(f_j \| y_j) \right\}.$$

The variational expression is used only in the proof; the algorithm evaluates its explicit coordinatewise formula.

Lemma 17 (A necessary exploration probability from one-arm alternatives)**.** *For every $p \in \mathcal{P}_y$ and $i \ne b$,*

$$p_i \ge \max_{0 \le z \le (1-\rho)h} \frac{(1-\rho)h - z}{(1-\rho)(y_i - z) + 14\gamma\,\mathrm{KL}(z \| y_i)}. \tag{22}$$

In particular,

$$p_i \ge \frac{(1-\rho)h}{(1-\rho)y_i + 14\gamma \log(1/(1-y_i))} > 0. \tag{23}$$

Proof. Fix $z \in [0, (1-\rho)h]$ and consider the alternative vector $f_i = z$ and $f_j = y_j$ for $j \neq i$. Its minimum is z, including when some predictions are tied. Also $p^\top y \geq h$, so feasibility implies

$$\begin{aligned} 0 &\geq (1-\rho)\{p^\top y - p_i(y_i - z)\} - z - 14\gamma p_i \,\mathrm{KL}(z\|y_i) \\ &\geq (1-\rho)h - z - p_i\{(1-\rho)(y_i - z) + 14\gamma \,\mathrm{KL}(z\|y_i)\}. \end{aligned}$$

Since $y_i \geq h > (1-\rho)h \geq z$, the denominator is strictly positive. Rearranging and maximizing gives (22). The maximum is attained because the ratio is continuous on a compact interval. Taking $z = 0$ gives (23). □

Theorem 18 (An upper-boundary condition for order improvements)**.** *For every $p \in \mathcal{P}_y$ and every $i \neq b$,*

$$\boxed{p_i \geq \frac{1-y_i}{281}\, q_i.}$$

Consequently, if $y_i \leq 1-d$ for all $i \neq b$, with $d \in (0,1)$, the exact MinCE allocation p^{opt} and the implemented MinCE allocation $\widehat{p}$ satisfy, for every $p \in \mathcal{P}_y$,

$$p_i^{\mathrm{opt}} \leq \widehat{p}_i \leq q_i \leq \frac{281}{d} p_i \quad (i \neq b), \qquad \sum_{i \neq b} \widehat{p}_i \leq \frac{281}{d} \min_{p \in \mathcal{P}_y} \sum_{i \neq b} p_i. \tag{24}$$

This conclusion requires no positive lower clipping level for h. The factor is independent of A, γ, and the horizon.

Proof. Fix $i \neq b$ and abbreviate $y = y_i$, $\Delta = y - h$, $\omega = \rho h$, $B = \Delta + \omega$, and $V = y(1-y)$. Both ω and V are positive, and $0 < \rho \leq 1/2$. For any $v \in [\omega, h]$, the choice $z = h - v$ in Lemma 17 gives

$$p_i \geq \frac{v - \omega}{(1-\rho)(\Delta + v) + 14\gamma \,\mathrm{KL}(h - v\|y)}. \tag{25}$$

We choose $v = \min\{h, \Delta + 2\omega\}$, which belongs to $[\omega, h]$. The elementary Bernoulli KL bound

$$\mathrm{KL}(x\|y) \leq \frac{(x-y)^2}{y(1-y)} \quad (0 \leq x \leq 1,\ 0 < y < 1) \tag{26}$$

follows by applying $\log u \leq u - 1$ to the two terms of the divergence, with endpoint values obtained by continuity.

First suppose $\Delta + 2\omega \leq h$. Then $v - \omega = B$ and $\Delta + v = 2B$, so (25) and (26) imply

$$p_i \geq \frac{B}{2(1-\rho)B + 56\gamma B^2/V}.$$

Because $Ah = \gamma\omega/5$, the FastCB probability is $q_i = h/[\gamma(\Delta + \omega/5)]$. Hence

$$\begin{aligned} \frac{p_i}{q_i} &\geq \frac{\gamma(\Delta + \omega/5)}{h\{2(1-\rho) + 56\gamma B/V\}} \\ &\geq \frac{V}{h\{280 + 10(1-\rho)V/(\gamma B)\}}. \end{aligned}$$

Here $\Delta + \omega/5 \geq B/5$. Moreover

$$\frac{V}{\gamma B} \leq \frac{h + \Delta}{\gamma(\Delta + \rho h)} \leq \frac{1}{\gamma\rho} = \frac{1}{5A} \leq \frac{1}{10}.$$

The second inequality uses $\rho(h + \Delta) \leq \rho h + \Delta$. Thus $p_i/q_i \geq V/(281h) \geq (1-y)/281$, since $y \geq h$.

Now suppose $\Delta + 2\omega \geq h$. Then $v = h$ and the zero alternative bound gives

$$p_i \geq \frac{(1-\rho)h}{(1-\rho)y + 14\gamma \log(1/(1-y))} \geq \frac{(1-\rho)h}{y\{(1-\rho) + 14\gamma/(1-y)\}},$$

where the last inequality also follows from (26) at $x = 0$. It follows that

$$\frac{p_i}{q_i} \geq (1-y)\frac{(1-\rho)(\Delta+\omega/5)/y}{14+(1-\rho)(1-y)/\gamma}.$$

Writing $x = \Delta/h$, the case condition is $x \geq 1-2\rho \geq 0$. The function $(x+\rho/5)/(1+x)$ is increasing in x, and therefore

$$\frac{(1-\rho)(\Delta+\omega/5)}{y} = \frac{(1-\rho)(x+\rho/5)}{1+x} \geq \frac{1-9\rho/5}{2} \geq \frac{1}{20}.$$

Since $\gamma \geq 10A \geq 20$, the denominator is at most $14 + 1/20 = 281/20$. This again proves $p_i/q_i \geq (1-y)/281$.

The two cases cover equality as well. Finally, Theorems 4 and 14 give $p_i^{\mathrm{opt}} \leq \widehat{p}_i \leq q_i$ for $i \neq b$. Combining this fact with the just-proved lower bound yields the coordinatewise claim in (24). Summing over $i \neq b$ and minimizing over $p \in \mathcal{P}_y$ gives its total-probability claim. The feasible set is nonempty because it contains q, and compact because the explicit certificate is continuous on the simplex, so the minimum exists. □

The bound identifies a necessary boundary condition for an unbounded multiplicative reduction relative to q: if all nongreedy predictions stay below $1-d$ for a fixed $d > 0$, the total exploration remains at least $d/281$ times that of q. The following appendix shows that a reduction of increasing order is attainable as nongreedy predictions approach one. These statements compare allocations at the same prediction vector.

E.1 Matching exploration laws with heterogeneous predictions

Use the fixed-round notation of Appendix C, with $\gamma \geq 10A$, $\rho = 5A/\gamma$, and $r = (1-\rho)/(14\gamma)$. For the proof write $\phi_y(z) = \log(1-y+ye^z)$, so

$$U_i = \phi_{y_i}(r), \qquad D_i(s) = s\{U_i - \phi_{y_i}(r - 1/(14\gamma s))\} \quad (s > 0), \qquad D_i(0) = 0.$$

The minimum total certified exploration is

$$E(y) = \min_{p \in \Delta_A:\, C_y(p) \leq 0} (1-p_b) = 1 - p_b^{\mathrm{opt}}.$$

Theorem 4 shows that this minimum is attained and that its nongreedy probabilities are each smaller than those of q. The next two results establish the finite and asymptotic assertions of Theorem 6.

Theorem 19 (Necessary and sufficient inverse-information exploration)**.** *Suppose $y_b = h \in [1/10, 1/2]$ and, for $i \neq b$, $y_i = 1 - e^{-L_i}$ with arbitrary $L_i \geq 4$. Then*

$$\frac{1}{560\gamma}\sum_{i\neq b}\frac{1}{L_i} \leq E(y) \leq \frac{3}{56\gamma}\sum_{i\neq b}\frac{1}{L_i}.$$

In particular, the required total nongreedy mass under this certificate is $\Theta(\gamma^{-1}\sum_{i\neq b} L_i^{-1})$, uniformly over A, γ, h and the heterogeneous collection (L_i) in the specified ranges. The upper bound is witnessed by the explicit allocation

$$\bar{p}_i = \frac{3}{56\gamma L_i} \quad (i \neq b), \qquad \bar{p}_b = 1 - \sum_{i\neq b}\bar{p}_i, \tag{27}$$

which satisfies $C_y(\bar{p}) < 0$ and $\bar{p}_i < q_i$ for every $i \neq b$.

Proof. We have $14\gamma \geq 140A \geq 280$, $14\gamma\rho = 70A$, and $0 < \rho \leq 1/2$. Put $H = \phi_h(r)$ and $Z = \sum_{i\neq b} L_i^{-1}$. Jensen's inequality and $\phi_y(r) \leq r$ give

$$\frac{1}{280\gamma} \leq hr \leq H \leq \frac{1}{14\gamma}, \qquad 0 < U_i \leq \frac{1}{14\gamma}.$$

Universal lower bound. Let p be any certified allocation and $E = 1 - p_b$. Because $\phi_{y_i}(z) \geq \log(1 - y_i) = -L_i$,

$$(1 - E)H \leq \sum_j p_j U_j \leq D_i(p_i) \leq p_i(L_i + U_i), \qquad i \neq b.$$

Consequently, with $R = \sum_{i \neq b}(L_i + U_i)^{-1}$,

$$E \geq (1 - E)HR, \qquad E \geq \frac{HR}{1 + HR}.$$

The elementary bounds $U_i \leq L_i/3$ and $L_i \geq 4$ give $R \geq 3Z/4$ and $HR \leq (A - 1)/(56\gamma) < 1/2$. Therefore

$$E \geq \frac{2}{3}HR \geq \frac{2}{3}\frac{1}{280\gamma}\frac{3}{4}Z = \frac{Z}{560\gamma}.$$

Feasible upper witness. For the allocation in (27), denote $E_0 = 1 - \bar{p}_b = 3Z/(56\gamma)$. The parameter restrictions imply

$$E_0 \leq \frac{3(A - 1)}{224\gamma} < \frac{1}{700}, \qquad \frac{E_0}{\rho} \leq \frac{3}{1120} < \frac{1}{350}.$$

For a comparator j, write $C_{y,j}(p) = 14\gamma[p^\top U - D_j(p_j)]$, so $C_y(p) = \max_j C_{y,j}(p)$. For $j \neq b$, the exact envelope is

$$C_{y,j}(\bar{p}) = 14\gamma\bar{p}_b\phi_h(r) + 14\gamma \sum_{i \neq b,j} \bar{p}_i\phi_{y_i}(r) + 14\gamma\bar{p}_j\phi_{y_j}(r - 4L_j/3).$$

Its first term is at most $he^{1/(14\gamma)} < 0.502$; its second is at most E_0. Furthermore,

$$\phi_{y_j}(r - 4L_j/3) = -L_j + \log\{1 + (1 - e^{-L_j})e^{r - L_j/3}\} \leq -L_j + \log 2 + r.$$

Thus

$$C_{y,j}(\bar{p}) \leq he^{1/(14\gamma)} + E_0 - \frac{3}{4} + \frac{3(\log 2 + 1/(14\gamma))}{4L_j} < 0.$$

For completeness, the final inequality follows from $he^{1/(14\gamma)} < 0.502$, $E_0 < 0.002$, and the last term being less than $3(0.7 + 0.004)/16 = 0.132$.

For the greedy comparator, put $z = \rho + E_0/(1 - E_0) < 1.003\rho$. The Bernoulli log-moment bound $\phi_h(s) \leq hs + s^2/8$ yields

$$\begin{aligned} C_{y,b}(\bar{p}) &\leq -h(1 - E_0)z + \frac{(1 - E_0)z^2}{112\gamma} + E_0(1 - \rho) \\ &\leq -\rho h + E_0(1 - \rho)(1 - h) + \frac{z^2}{112\gamma} \\ &\leq \rho\left\{-\frac{1}{10} + \frac{1}{350} + \frac{1.003^2\rho}{112\gamma}\right\} < 0. \end{aligned}$$

All A comparator envelopes are strictly negative, so $C_y(\bar{p}) < 0$.

Finally, $Ah + \gamma(y_i - h) \leq Ah + \gamma(1 - h) \leq \gamma$ because $\gamma \geq A$. Hence $q_i \geq h/\gamma \geq 1/(10\gamma)$, whereas $\bar{p}_i \leq 3/(224\gamma) < 1/(10\gamma)$. Thus the witness is certified and reduces every nongreedy probability of q. □

Theorem 20 (The sharp heterogeneous exploration law)**.** *Fix $A \geq 2$, $\gamma \geq 10A$, $h \in [1/10, 1/2]$, and positive numbers $(\lambda_i)_{i \neq b}$. For $L \to \infty$, let $y_b = h$ and $y_i = 1 - e^{-\lambda_i L}$, and put $H = \log(1 - h + he^r)$. Then*

$$E(y) \sim \frac{H}{L}\sum_{i \neq b}\frac{1}{\lambda_i}.$$

Proof. First note $0 < 14\gamma H \leq he^{1/(14\gamma)} < 0.502 < 1$ and

$$q_i \longrightarrow q_0 := \frac{h}{Ah + \gamma(1 - h)} > 0 \quad \text{for every } i \neq b.$$

We use the following elementary limit. If $s_L > 0$, $Ls_L \to c > 0$, and $y_L = 1 - e^{-\lambda L}$, then

$$D_{y_L}(s_L) \longrightarrow \min\{\lambda c, 1/(14\gamma)\}. \tag{28}$$

Indeed $s_L U_{y_L} \to 0$, and

$$\frac{1}{L}\log\{e^{-\lambda L} + (1 - e^{-\lambda L})e^{r-1/(14\gamma s_L)}\} \longrightarrow \max\{-\lambda, -1/(14\gamma c)\}.$$

This log-sum-exp limit follows by sandwiching the logarithm between the larger summand's logarithm and that logarithm plus $\log 2$.

Matching lower and upper limits. The lower-bound proof above, with $L_i = \lambda_i L$, shows

$$E(y) \geq \frac{HR_L}{1 + HR_L}, \qquad R_L = \sum_{i \neq b} \frac{1}{\lambda_i L + U_i},$$

so $\liminf LE(y) \geq H \sum_i \lambda_i^{-1}$. Choose any $\delta > 0$ small enough that $14\gamma(H + \delta) < 1$, and put $p_i = (H + \delta)/(\lambda_i L)$ for $i \neq b$, $p_b = 1 - \sum_{i \neq b} p_i$. For sufficiently large L, this is a distribution and $p_i < q_i$ for every $i \neq b$. We have $\sum_i p_i U_i \to H$, while (28) gives $D_i(p_i) \to H + \delta$ for every nongreedy comparator. For the greedy comparator,

$$C_{y,b}(p)/(14\gamma) = (1 - E)\phi_h(r - 1/(14\gamma(1 - E))) + \sum_{i \neq b} p_i U_i \longrightarrow \phi_h(-\rho/(14\gamma)) < 0.$$

All comparators are therefore strictly feasible for sufficiently large L. This gives $\limsup LE(y) \leq (H + \delta)\sum_i \lambda_i^{-1}$. Letting $\delta \downarrow 0$ proves the matching upper limit.

□

Interpretation and scope. The quantity $L_i = \mathrm{KL}(0\|y_i) = -\log(1 - y_i)$ is the log-loss cost of reversing action i's high predicted loss to zero. The finite theorem identifies both a necessary and sufficient exploration law under the prescribed certificate; the sharp theorem identifies its exact leading constant. These are not information-theoretic lower bounds over all contextual-bandit algorithms. They do not compare the regret of separately trained learners. The sharp limit fixes $A, \gamma, h, (\lambda_i)$ before $L \to \infty$; the explicit finite bound is uniform over its stated parameter range. The finite MinCE implementation has the same exploration orders by Theorem 14; its relative error is at most $1/T$.

F An extension to offline regression under IID contexts

The online guarantee permits adaptive contexts and uses a regression oracle whose regret is controlled on adaptive data. An offline oracle instead provides an estimation guarantee for independent samples. We obtain an offline version by freezing the predictor within doubling epochs, training on the preceding epoch, and applying the same allocation routine to a fixed affine rescaling of its predictions. This follows the epoch-based reduction of Simchi-Levi & Xu (2022); the argument below establishes the additional coverage property needed for MinCE.

F.1 Assumptions and algorithm

Assume that (x_t, ℓ_t) are independent draws from a fixed distribution on $\mathcal{X} \times [0,1]^A$, with $\mathbb{E}[\ell_t(a) \mid x_t] = f^*(x_t, a)$ and $f^* \in \mathcal{F}$. For any fixed randomized policy $p(a \mid x)$, an offline regression oracle receives n independent observations $(x, a, \ell(a))$, where $a \sim p(\cdot \mid x)$. It returns a predictor $\widehat{f} : \mathcal{X} \times \mathcal{A} \to [0,1]$ such that, with probability at least $1 - \vartheta$,

$$\mathbb{E}_x \sum_a p(a \mid x)\big(\widehat{f}(x,a) - f^*(x,a)\big)^2 \leq \xi(n, \vartheta). \tag{29}$$

The known bound $\xi(n, \vartheta) \in (0, 1]$ is nonincreasing in n and is valid for every sampling policy. Conditional on the history at an epoch's start, its samples have precisely this independent distribution. The oracle need not optimize over policies or compute confidence sets.

For $T \geq 2$, set $J = \lceil \log_2(T+1) \rceil$ and $\vartheta = 1/(TJ)$ internally. Epoch m has length $n_m = \min\{2^{m-1}, T - 2^{m-1} + 1\}$, for $m = 1, \ldots, J$. Every epoch before the last is complete. In epoch $m \geq 2$, fit $\widehat{f}_m$ using only the n_{m-1} examples from epoch $m-1$ and set

$$\xi_m = \xi(n_{m-1}, \vartheta), \qquad \gamma_m = \max\{10A, \sqrt{A/\xi_m}\}, \qquad \widehat{y}_{t,a} = \tfrac{1}{4} + \tfrac{1}{2}\widehat{f}_m(x_t, a). \tag{30}$$

The transformation preserves the greedy action and keeps allocation scores in $[1/4, 3/4]$. It introduces no adjustable parameter. Observed losses and regression targets remain on their original scale. The rescaling is relevant: the exploration lower bound in Theorem 6 contains $1 - \widehat{y}_{t,a}$, which can approach zero without it. Thus simply freezing the original online algorithm is not the offline guarantee proved here.

Algorithm 4 MinCE with an offline regression oracle

Input: Horizon T; offline oracle with error bound ξ in (29).

Receive x_1, sample a_1 uniformly from $\mathcal{A}$, observe $\ell_1(a_1)$, and store $(x_1, a_1, \ell_1(a_1))$. If $T = 1$, stop.
Set $J \leftarrow \lceil \log_2(T+1) \rceil$ and $\vartheta \leftarrow 1/(TJ)$.
for $m = 2, \ldots, J$ **do**
  Fit $\widehat{f}_m$ using only the examples from epoch $m-1$; compute ξ_m, γ_m from (30).
  for each round t in epoch m **do**
    Receive x_t and set $\widehat{y}_{t,a} \leftarrow 1/4 + \widehat{f}_m(x_t, a)/2$ for every a.
    Set $b_t \leftarrow \min \operatorname{argmin}_a \widehat{y}_{t,a}$ and form q_t using (3) with $\gamma = \gamma_m$.
    $p_t \leftarrow \mathrm{CertifiedAllocation}(\widehat{y}_t, q_t, b_t; \gamma_m, T)$.
    Sample $a_t \sim p_t$, observe $\ell_t(a_t)$, and store $(x_t, a_t, \ell_t(a_t))$ for the next fit.
  end for
end for

There are $J - 1 = \mathcal{O}(\log T)$ oracle calls. The predictor and sampling rule are fixed as functions of the context throughout an epoch; contexts themselves continue to change every round. The allocation routine is unchanged, including its horizon-determined numerical accuracy.

F.2 Regret guarantee

Theorem 21 (Offline-oracle regret)**.** *Under the assumptions above, for $T \geq 2$, Algorithm 4 satisfies*

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{CB}}(T)] \leq \min\left\{T,\ 2 + 44963 \sum_{m=2}^{J} n_m \sqrt{A\xi(n_{m-1}, 1/(TJ))}\right\}.$$

For $T = 1$, its expected regret is at most one. At every round after the first, its nongreedy probabilities are no larger than those of the reference allocation computed from the same rescaled predictions and the same γ_m.

The numerical constant is a conservative consequence of the uniform coverage bound, rather than a parameter used by the algorithm. This extension gives a worst-case regret guarantee; it does not assert the online first-order bound or preserve the unrescaled algorithm's upper-boundary exploration reductions.

Proof. Let $p_m(a \mid x)$ be the policy used in epoch m, and write $b_m(x) = \min \operatorname{argmin}_a \widehat{f}_m(x, a)$ and $\widehat{\Delta}_m(x, a) = \widehat{f}_m(x, a) - \widehat{f}_m(x, b_m(x))$. Theorem 6, the range $[1/4, 3/4]$, and $p_m(a \mid x) \leq q_m(a \mid x)$ for nongreedy actions imply

$$\frac{1}{p_m(a \mid x)} \leq \alpha\big(A + 2\gamma_m \widehat{\Delta}_m(x, a)\big) \quad \text{for every } a, \qquad \alpha = 1124. \tag{31}$$

Indeed, for a nongreedy action, $p_m(a \mid x) \geq q_m(a \mid x)/1124$, while

$$q_m(a \mid x) = \frac{1}{A + \gamma_m \widehat{\Delta}_m(x, a)/(1/2 + \widehat{f}_m(x, b_m(x)))} \geq \frac{1}{A + 2\gamma_m \widehat{\Delta}_m(x, a)}.$$

For the greedy action $p_m(b_m(x) \mid x) \geq q_m(b_m(x) \mid x) \geq 1/A$. The same reference formula also gives

$$\mathbb{E}_x \sum_a p_m(a \mid x)\widehat{\Delta}_m(x,a) \leq \frac{3A}{2\gamma_m}. \tag{32}$$

This follows termwise from $q_m(a \mid x)\widehat{\Delta}_m(x,a) \leq [1/2 + \widehat{f}_m(x, b_m(x))]/\gamma_m \leq 3/(2\gamma_m)$.

By the conditional guarantee (29) and a union bound over the $J-1$ fits, with probability at least $1 - 1/T$, all predictors satisfy

$$\mathbb{E}_x \sum_a p_{m-1}(a \mid x)(\widehat{f}_m(x,a) - f^*(x,a))^2 \leq \xi_m.$$

Call this event $\mathcal{G}$. The expectation in this display is over a fresh context, conditional on the fitted predictor and earlier history. In particular, the guarantee is not a claim about reuse of an independent test sample.

On $\mathcal{G}$, define, for any measurable deterministic policy π,

$$R(\pi) = \mathbb{E}_x[f^*(x,\pi(x)) - f^*(x,\pi^*(x))], \qquad \widehat{R}_m(\pi) = \mathbb{E}_x \widehat{\Delta}_m(x,\pi(x)).$$

We prove simultaneously for every such policy, by induction on m, that

$$R(\pi) \leq 2\widehat{R}_m(\pi) + \frac{\beta A}{\gamma_m}, \qquad \widehat{R}_m(\pi) \leq 2R(\pi) + \frac{\beta A}{\gamma_m}, \qquad \beta = 40\alpha. \tag{33}$$

If $\gamma_m = 10A$, both assertions follow from $R(\pi), \widehat{R}_m(\pi) \in [0,1]$ and $\beta/10 \geq 1$. Otherwise $\xi_m = A/\gamma_m^2$. Weighted Cauchy–Schwarz on the single event $\mathcal{G}$ yields, for every π,

$$e_m(\pi) := \mathbb{E}_x|\widehat{f}_m(x,\pi(x)) - f^*(x,\pi(x))| \leq \sqrt{\xi_m \mathbb{E}_x[1/p_{m-1}(\pi(x) \mid x)]}.$$

For $m = 2$, the inverse probability in this expression is A. For $m > 2$, (31), the induction hypothesis, and $\gamma_{m-1} \leq \gamma_m$ give

$$\mathbb{E}_x[1/p_{m-1}(\pi(x) \mid x)] \leq \alpha(1+2\beta)A + 4\alpha\gamma_m R(\pi).$$

This upper bound also covers the uniform initial epoch. Consequently,

$$e_m(\pi) \leq \frac{R(\pi)}{4} + \frac{\chi A}{\gamma_m}, \qquad \chi = 4\alpha + \sqrt{\alpha(1+2\beta)} \leq 13\alpha.$$

Here we used $2\sqrt{\alpha A R(\pi)/\gamma_m} \leq R(\pi)/4 + 4\alpha A/\gamma_m$. There is no union bound over policies: weighted Cauchy–Schwarz applies to every policy on the same population squared-error event.

Comparing π with π^* and using $R(\pi^*) = 0$ gives

$$R(\pi) \leq \widehat{R}_m(\pi) + e_m(\pi) + e_m(\pi^*) \leq \widehat{R}_m(\pi) + \tfrac{1}{4}R(\pi) + 2\chi A/\gamma_m.$$

Thus the first inequality in (33) follows from $8\chi/3 \leq \beta$. Conversely, comparison with b_m gives

$$\begin{aligned}\widehat{R}_m(\pi) &\leq R(\pi) - R(b_m) + e_m(\pi) + e_m(b_m) \\ &\leq \tfrac{5}{4}R(\pi) - \tfrac{3}{4}R(b_m) + 2\chi A/\gamma_m \leq 2R(\pi) + \beta A/\gamma_m,\end{aligned}$$

since $2\chi \leq \beta$. This completes the induction.

Integrating the first inequality in (33) over a representation of the randomized policy p_m as deterministic policies, then applying (32), gives

$$R(p_m) := \mathbb{E}_x \sum_a p_m(a \mid x)[f^*(x,a) - f^*(x,\pi^*(x))] \leq (\beta+3)A/\gamma_m.$$

Such a representation follows directly by inverse-CDF sampling with one uniform variable. If $\gamma_m > 10A$, this is $(\beta+3)\sqrt{A\xi_m}$. If $\gamma_m = 10A$, then $\sqrt{A\xi_m} \geq 1/10$ and the trivial bound $R(p_m) \leq 1$ proves the same conclusion.

The random variable $\sum_m n_m R(p_m)$ is a sum of population epoch regrets. On $\mathcal{G}$ it is at most $1 + (\beta + 3)\sum_{m \geq 2} n_m\sqrt{A\xi_m}$, and it is at most T otherwise. The tower property, applied unconditionally at the start of each epoch, identifies its expectation with $\mathbb{E}[\mathrm{Reg}_{\mathrm{CB}}(T)]$. The exceptional event contributes at most $T\Pr(\mathcal{G}^c) \leq 1$. Since $\beta + 3 = 44963$, this proves the bound. The greedy comparison is the unchanged allocation routine's coordinatewise guarantee. □

F.3 A FINITE-CLASS SPECIALIZATION

Corollary 22 (Finite regression classes)**.** *If $|\mathcal{F}| = N < \infty$ and the offline oracle minimizes empirical squared loss over $\mathcal{F}$, one may use*

$$\xi(n, \vartheta) = \min\{1, 16\log(N/\vartheta)/n\}.$$

For $T \geq 2$, Algorithm 4 then has $\mathbb{E}[\mathrm{Reg}_{\mathrm{CB}}(T)] = \mathcal{O}(\sqrt{AT\log(NT\log(T+1))})$, capped by T, with $\mathcal{O}(\log T)$ offline regression calls.

Proof. For a fixed sampling distribution and $f \in \mathcal{F}$, let $Z = (f(x,a) - \ell(a))^2 - (f^*(x,a) - \ell(a))^2$ and $\mu = \mathbb{E}Z = \mathbb{E}(f(x,a) - f^*(x,a))^2$. Then $\mathrm{Var}(Z) \leq \mathbb{E}Z^2 \leq 4\mu$ and $|Z - \mu| \leq 2$. Bernstein's inequality implies

$$\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} Z_i \leq 0\right\} \leq \exp(-3n\mu/28).$$

A union bound over the N functions, together with empirical risk minimization relative to f^*, proves (29) with the stated bound. For the regret rate, substitute $\vartheta = 1/(TJ)$ into Theorem 21 and use the geometric epoch schedule: $\sum_{m=2}^{J} n_m/\sqrt{n_{m-1}} = \mathcal{O}(\sqrt{T})$. □

The general theorem also applies to an infinite class, including neural regression, whenever its offline training procedure supplies (29). An arbitrary neural network trained by gradient descent does not automatically satisfy that assumption. No claim about its statistical error is needed to implement the wrapper, but a valid error bound is needed to instantiate its theorem.

G EXPERIMENTAL PROTOCOL AND ALLOCATION CALIBRATION

This appendix specifies the neural implementation shared by the synthetic and OpenML experiments.

G.1 NEURAL REGRESSION AND FEEDBACK

Every regressor has three fully connected hidden layers of width 32 with $\tanh$ activations and no skip connection. Synthetic experiments score each action's four-dimensional features with a shared scalar sigmoid output (4–32–32–32–1, 2,305 parameters). OpenML experiments use a shared context trunk and one sigmoid cost head per declared class. Hidden weights have independent zero-mean Gaussian initialization with variance reciprocal to fan-in; output weights have standard deviation 0.01, and all biases are zero. All weights are trained.

Each round makes one update using only the selected action's observed loss. There is no importance weighting, replay, full-label update, or supervised warm start. SquareCB and SpannerIGW minimize squared loss; SquareCB.L, FastCB, and MinCE minimize binary cross-entropy evaluated on raw logits. Thus SquareCB.L denotes the log-loss-oracle variant of SquareCB, as in Foster & Krishnamurthy (2021), rather than a separate neural architecture. Adam uses coefficients $(0.9, 0.999)$, numerical offset 10^{-8}, and global gradient-norm clipping at 5. Predictions returned to the allocation rule are clipped to $[\epsilon, 1-\epsilon]$ with $\epsilon = 10^{-6}$.

For each method and condition, seed 10000 chooses the learning rate from $\{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10\}$ by minimum cumulative incurred cost; ties favor the smaller rate. Methods share the context sequence, initialization, and action-uniform stream within each seed, but their selected actions and learned predictions can differ. Learning rates are chosen separately for each method; consequently, FastCB versus MinCE is not an allocation-only ablation with identical learned predictors. No allocation parameter or time-varying exploration schedule is tuned. The tuning criterion measures online bandit cost, not the predictor's full-information validation loss.

G.2 ALLOCATION PARAMETERS AND THEIR INTERPRETATION

Set $\delta = 1/T$ and $\bar{L} = T$. Bounded returned predictions imply the universal oracle-regret budgets

$$B_{\mathrm{sq}} = T(1-\epsilon)^2, \qquad B_{\log} = T\log(1/\epsilon),$$

because per-round prediction losses are bounded and comparator losses are nonnegative. These budgets apply to the returned predictions, regardless of the internal optimization procedure. They are linear in T and generally vacuous: they neither establish neural realizability nor a useful sublinear oracle-regret guarantee for Adam. Observed training loss is not substituted for oracle regret.

Following the allocation and Theorem 1 calibration of Foster & Rakhlin (2020), SquareCB and SquareCB.L use

$$\mu = A, \qquad \gamma_{\mathrm{sq}} = \sqrt{\frac{AT}{B_{\mathrm{sq}} + \log(2/\delta)}}, \qquad p_{t,a} = \frac{1}{A + \gamma_{\mathrm{sq}}(\widehat{y}_{t,a} - \widehat{y}_{t,b_t})} \quad (a \neq b_t),$$

with residual mass on b_t. For SquareCB.L, the squared-loss budget follows from bounded predictions; it is not obtained by equating log-loss regret with squared-loss regret. FastCB and MinCE use

$$\gamma = \max\left\{10A, \sqrt{\frac{A\bar{L}}{3B_{\log}}}\right\} = 10A.$$

The linear budgets keep SquareCB allocations close to uniform on many conditions. The comparison therefore reflects calibration as well as allocation design, and does not assess optimally tuned exploration parameters. FastCB and MinCE share the log-loss objective, oracle budget, and allocation parameter, making their comparison more directly informative about the proposed allocation.

For SpannerIGW, we specialize Algorithm 2 of Zhu et al. (2022) to $\phi(x, a) = e_a$ and action dimension $d = A$. Rewards and returned predictions are $-\ell_t(a)/\sqrt{A}$ and $-\widehat{y}_{t,a}/\sqrt{A}$, so the predictor vector has norm at most one and $B_{\mathrm{sp}} = B_{\mathrm{sq}}/A$. Internally, the oracle uses the original cost and the same squared-loss update. For reweighted one-hot vectors $z_a = e_a/\sqrt{w_a}$, any design u has $z_a^\top V(u)^{-1} z_a = 1/u_a$. The uniform design is therefore exactly G-optimal, with $C_{\mathrm{opt}} = 1$. The proof calibration gives

$$\gamma_{\mathrm{sp}} = \sqrt{\frac{AT}{B_{\mathrm{sq}}/A + 32\log(2/\delta)}}, \qquad q_{t,a}^{\mathrm{sp}} = \frac{1}{2A} + \frac{1}{2}\mathbf{1}\{a = b_t\},$$
$$p_{t,a} = \frac{q_{t,a}^{\mathrm{sp}}}{\nu_t + \gamma_{\mathrm{sp}}(\widehat{y}_{t,a} - \widehat{y}_{t,b_t})/(A\sqrt{A})}.$$

Scalar bisection finds the unique $\nu_t \in [1/2, 1]$ satisfying $\sum_a p_{t,a} = 1$. Reported losses retain their original units. This specialization requires no optimization over neural weights or function classes, but does not test SpannerIGW's large-action-space advantage.

G.3 Metrics and uncertainty

For synthetic experiments, let $\mu_t(a)$ be the known conditional mean reward. The plotted cumulative predictable pseudo-regret and the cumulative mean predicted-greedy probability are

$$R(t) = \sum_{\tau=1}^{t}\left[\max_a \mu_\tau(a) - \sum_a p_{\tau,a}\mu_\tau(a)\right], \qquad G(t) = \frac{1}{t}\sum_{\tau=1}^{t} p_{\tau,b_\tau}.$$

Greediness always refers to the learner's own predicted-best action, using first-index tie breaking; it does not measure selection of the true optimal action. OpenML performance is incurred classification loss $T^{-1}\sum_t \ell_t(a_t)$, not pseudo-regret, because the optimal neural comparator is unknown. All bands and per-condition error bars are sample SDs over ten seeds (denominator nine), not confidence intervals for the mean. Unique-best counts compare unrounded mean outcomes and do not constitute significance tests.

H Synthetic experiments

H.1 Reward models and changing contexts

Each condition has $T = 5{,}000$ rounds, $A = 5$ actions, and action features $x_{t,a} \in \mathbb{R}^4$ with $\|x_{t,a}\|_2 \leq 1$. For a unit vector $\theta^\star$, write $z_{t,a} = \langle \theta^\star, x_{t,a}\rangle$ and $\sigma(z) = (1 + e^{-z})^{-1}$. The three conditional reward

means are

$$\begin{aligned}
&\text{Linear reward:} && \mu_t(a) = 0.5 + 0.45 z_{t,a},\\
&\text{GLM (logistic) reward:} && \mu_t(a) = \sigma(3 z_{t,a}),\\
&\text{Nonlinear (exponential) reward:} && \mu_t(a) = 0.05 + 0.9\frac{e^{z_{t,a}+1} - 1}{e^2 - 1}.
\end{aligned}$$

Rewards are Bernoulli with these means, and the learner observes cost $\ell_t(a_t) = 1 - r_t(a_t)$. All three links are monotone in the same linear score: the exponential model varies the reward scale nonlinearly, but does not introduce a nonlinear optimal-action boundary. Neither $\theta^\star$ nor the reward model is supplied to the learner. For every context process, draw 32 normalized Gaussian reference directions and use direction $s \bmod 32$ as $\theta^\star$ for seed s.

Gaussian and uniform contexts. Gaussian action vectors are drawn independently at every round from $N(0, I/12)$ conditional on norm at most one, using rejection sampling. This is a truncated Gaussian, not an unconditioned Gaussian with covariance $I/12$. Uniform action vectors have independent coordinates in $[-1/2, 1/2]$ and are also resampled every round.

Adversarial (prior-mismatch) contexts. For each reward family, the reference prediction averages its mean reward over the 32 reference directions. Presample 256 uniform context matrices and rank them by the true reward gap incurred by the reference predictor's greedy action, breaking ties stably. Present the ten largest-gap matrices in that order for 500 rounds each, without resetting the learner. The entire sequence is fixed before interaction. Thus "adversarial" denotes this specific oblivious stress construction; it does not denote an adaptive or general worst-case adversary. The reference prior is used only to construct the environment and is neither the neural initialization nor information given to a learner.

All settings use the neural implementation in Appendix G. With $T = 5{,}000$ and $A = 5$, the allocation parameters are $\gamma_{\mathrm{sq}} = 2.23401355$, $\gamma = 50$ for FastCB and MinCE, and $\gamma_{\mathrm{sp}} = 4.39420774$. The random streams use seeds $130000 + s$ for reference functions, $120000 + s$ for fresh features, $150000 + s$ for the candidate context pool, $140001 + s$ for action uniforms, and $240002 + s$ for potential Bernoulli rewards. Potential rewards and action uniforms are matched across methods.

H.2 Regret and exploration results

Table 2 gives the uncropped final values underlying Figure 1. MinCE has the lowest mean regret in each displayed condition. Averaging conditions equally gives 56.07 for MinCE and 171.99 for FastCB, a 67.4% reduction. The strongest relative improvement is on Gaussian contexts with exponential rewards (80.8%), and the smallest is on prior-mismatch contexts with logistic rewards (41.8%). The latter also has substantial seed variability: MinCE's regret is 82.77 ± 87.82, compared with 142.20 ± 30.76 for FastCB. The lower mean therefore does not imply better performance on every seed.

Table 2: Final synthetic pseudo-regret at $T = 5{,}000$, mean $\pm$ sample SD over ten seeds. Bold denotes the lowest unrounded mean, not statistical significance.

Reward	Context	SquareCB	SquareCB.L	FastCB	SpannerIGW	MinCE
Linear	Gaussian	680.95 ± 4.91	681.08 ± 5.07	150.71 ± 7.79	371.42 ± 6.01	$\mathbf{40.01 \pm 9.81}$
Linear	Uniform	690.74 ± 5.88	699.75 ± 6.35	152.56 ± 12.91	373.03 ± 3.53	$\mathbf{36.79 \pm 9.73}$
Linear	Prior-mismatch	924.02 ± 150.77	924.34 ± 150.79	181.35 ± 29.03	491.16 ± 72.06	$\mathbf{51.22 \pm 40.70}$
Logistic	Gaussian	948.15 ± 4.80	947.78 ± 4.79	136.37 ± 10.29	512.01 ± 3.47	$\mathbf{34.06 \pm 9.55}$
Logistic	Uniform	968.74 ± 7.95	968.28 ± 8.34	136.16 ± 6.58	523.34 ± 4.12	$\mathbf{35.06 \pm 8.29}$
Logistic	Prior-mismatch	1239.21 ± 194.53	1239.02 ± 194.72	142.20 ± 30.76	664.56 ± 99.98	$\mathbf{82.77 \pm 87.82}$
Exponential	Gaussian	654.15 ± 6.86	653.71 ± 6.95	199.53 ± 8.49	372.37 ± 8.04	$\mathbf{38.31 \pm 6.45}$
Exponential	Uniform	666.93 ± 7.74	667.16 ± 7.87	205.08 ± 6.91	361.03 ± 4.79	$\mathbf{71.45 \pm 15.71}$
Exponential	Prior-mismatch	933.08 ± 191.34	932.70 ± 191.52	243.91 ± 23.54	505.89 ± 97.42	$\mathbf{114.95 \pm 68.39}$

Figure 2 shows that these gains coexist with near-greedy allocations. MinCE's time-averaged predicted-greedy probability ranges from 99.824% to 99.883% across conditions, averaging 99.858%; the corresponding FastCB average is 80.131%. The other averages are 26.700% for SquareCB, 26.665% for SquareCB.L, and 61.958% for SpannerIGW. On a 0–100% scale, the MinCE curves appear to

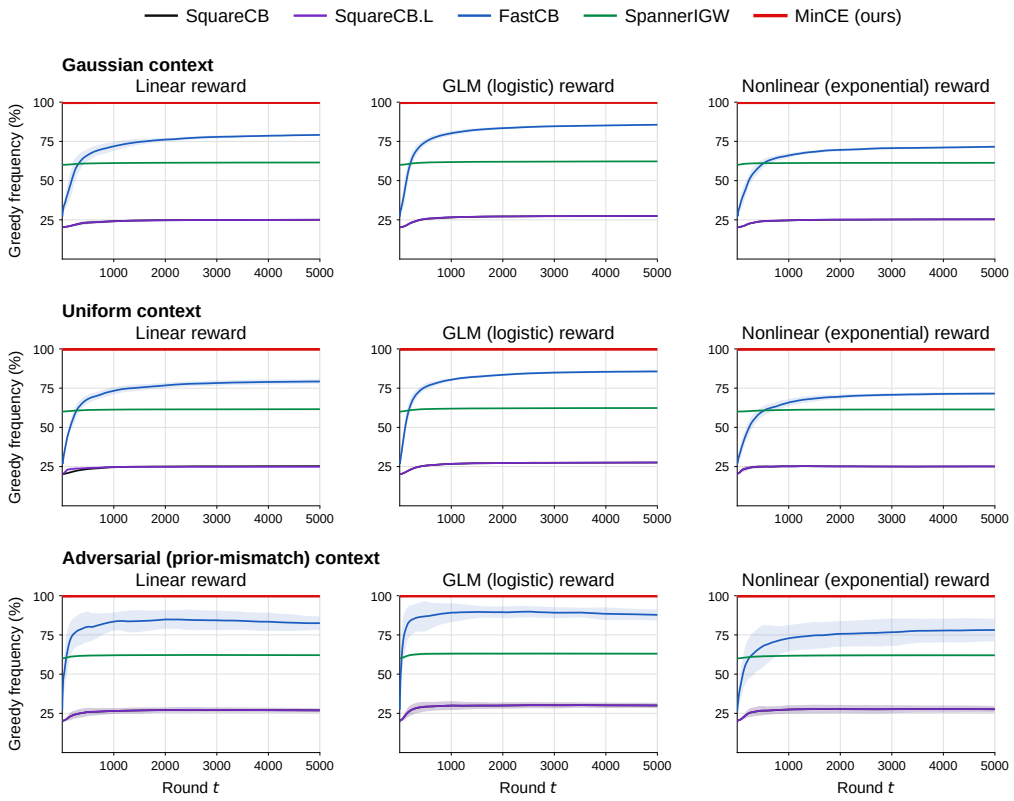

Figure 2: **Predicted-greedy probability in the synthetic experiments.** Curves show the cumulative average $G(t)$, with mean $\pm$ one sample SD over ten seeds. Greediness is relative to each learner's own predictions. Note that while MinCE (red) is near-greedy but strictly below, 100% on average; it continues to randomize.

reach 100%, but their residual exploration is nonzero. This statistic is a probability under each learner's current predictions, not a deterministic frequency guarantee or evidence that every greedy action is optimal.

The results are consistent with reduced exploration avoiding costly departures from informative predictions in these conditions. They do not show that the allocation change alone explains every gain: different feedback and separately chosen optimizer rates also change the learned predictors. In particular, the conservative calibration keeps the SquareCB variants highly exploratory. The large gaps to those baselines should be interpreted in that context.

H.3 Runtime and numerical implementation

Table 3 measures the complete online loop: prediction, allocation, action sampling, and the selected-action update. Each method and condition has three serial seed-0 replays, with randomized method order and recorded exact-trajectory checks. Initialization, data generation, compilation, input copies, and output are excluded.

Across the nine conditions, MinCE takes 0.187 s per 5,000-round episode on average, about 0.016 s more than FastCB (9.5% of its mean runtime). The median of conditionwise median-runtime ratios gives a 6.8% overhead. These are different aggregations of the same replays, not conflicting estimates. The additional computation is modest in this implementation, but is not free and may depend on architecture and action count. The table's SD is across three repeated timings within each condition, averaged over conditions; it does not measure variability across evaluation seeds.

Table 3: Synthetic runtime per 5,000-round episode. Means give equal weight to nine conditions; timing SDs are computed from three replays within each condition and then averaged. Ratios divide the overall mean by the FastCB mean.

Method	Mean time (s)	Mean timing SD (s)	Time ratio
SquareCB	0.1690	0.0044	0.990
SquareCB.L	0.1718	0.0056	1.006
FastCB	0.1708	0.0034	1.000
SpannerIGW	0.1722	0.0055	1.008
MinCE	0.1870	0.0060	1.095

The allocation routine uses internally determined accuracy $\tau = 1/(16T)$ and relative stopping factor $1 + 1/T$, with at most three sweeps. Across the 450,000 MinCE rounds in the displayed evaluation, the mean was two sweeps and no fallback or stagnation event was recorded. The largest computed certificate value was -4.52×10^{-7}. Finite-precision checks use ordinary double arithmetic, not interval arithmetic; the implementation has a safe FastCB fallback if its checks fail. These numerical diagnostics support the implementation checks and do not replace the exact-arithmetic analysis.

I OpenML classification experiments

I.1 Data, preprocessing, and bandit feedback

Each example (x_t, y_t) supplies a context and losses $\ell_t(a) = \mathbf{1}\{a \neq y_t\}$ over the declared classes. A run makes one random permutation of the examples, with T equal to the number of rows. Only $(x_t, a_t, \ell_t(a_t))$ reaches the update; even on a binary task, the unselected label is not inferred for training. Thus the contexts change over time and the sequence is oblivious to the learner's actions.

The datasets come from the 516-entry classification inventory evaluated by Foster & Krishnamurthy (2021, Appendix D.2); the collection includes synthetic tasks and related variants, not 130 independent application domains. Nominal inputs use ARFF declaration-order codes. Ignored and row-ID attributes are removed, missing inputs become zero, and inputs are converted to float32 through six-decimal serialization. Declared classes are retained even if absent from the observed rows. Each coordinate is divided by its running maximum absolute value, lower bounded by one and updated from the current context before prediction. Future examples and labels never determine this scale.

Architecture, optimizer, and allocation parameters are specified in Appendix G. Evaluation seed s uses PCG64 row-permutation seed s and action-uniform seed $9102000 + s$. The tuning seed 10000 supplies a different ordering of the same dataset, not an independent held-out dataset.

I.2 Aggregation, interpretation, and runtime

Every dataset–method result averages ten evaluation seeds. Aggregate loss and greediness give each dataset equal weight. To compute aggregate loss SD, first average dataset losses within each seed, then take the sample SD of the ten resulting averages. This is variability across randomized runs on the fixed collection, not uncertainty over a population of independent datasets. Best counts use integer mistake totals before rounding.

As shown in Table 1, MinCE lowers equal-dataset mean loss by 2.02 percentage points relative to FastCB (using unrounded values). Its predicted-greedy probability is higher by 15.58 percentage points. The pairwise comparison is 116 lower and 14 higher mean losses, with no ties; thus the improvement is widespread within this collection but not universal. The near-unit greedy probability measures the allocation, not prediction accuracy. Realized greedy-action frequencies are 51.110%, 50.812%, 84.008%, 72.774%, and 99.912% for SquareCB, SquareCB.L, FastCB, SpannerIGW, and MinCE, respectively, and differ from their probability averages through sampling.

Runtime includes prediction, allocation, sampling, and the online update, excluding data loading, initialization, and startup. Each condition has one serial seed-0 replay whose checkpoints and initial/final model states match evaluation; a single replay supplies no repeated-timing SD. The

recorded environment used one numerical thread and one serial worker on an AMD EPYC 9V74, Python 3.12.14, NumPy 2.3.5, and g++ 13.3.0 with `-O3 -march=native -fno-math-errno -ffp-contract=off -std=c++17`. The experiments report the median of per-dataset ratios to FastCB, giving a 5.60% MinCE overhead. These ratios should not be equated with the synthetic experiment's ratio of mean runtimes.

I.3 Results by dataset

Table 4 lists all 130 datasets alphabetically. The accompanying records retain dataset identifiers, input dimensions, horizons, per-seed outcomes, greedy frequencies, runtime, and optimizer rates; these are not repeated as separate paper tables.

Table 4: Incurred classification loss (%), mean ± sample SD over ten seeds. All methods use three hidden layers of width 32. Bold marks the uniquely smallest unrounded mean. Identically named dataset variants are distinguished by class count.

Dataset	SquareCB	SquareCB.L	FastCB	SpannerIGW	MinCE
abalone	95.88 ± 0.69	96.28 ± 0.29	90.59 ± 0.82	91.32 ± 0.55	$\mathbf{87.58 \pm 2.37}$
acute-inflammations	39.25 ± 4.26	38.92 ± 4.38	13.83 ± 4.48	29.00 ± 5.59	$\mathbf{13.67 \pm 5.50}$
adult	38.31 ± 0.20	37.60 ± 0.17	17.28 ± 0.10	28.37 ± 0.19	$\mathbf{16.83 \pm 0.08}$
airlines	47.46 ± 0.08	47.46 ± 0.08	38.99 ± 0.03	42.81 ± 0.07	$\mathbf{37.79 \pm 0.08}$
amazon-commerce-reviews	98.07 ± 0.39	$\mathbf{97.99 \pm 0.40}$	98.04 ± 0.25	98.08 ± 0.26	98.07 ± 0.33
analcatdata_cyyoung8092	41.75 ± 5.08	42.58 ± 4.79	29.79 ± 1.64	35.77 ± 4.81	$\mathbf{26.49 \pm 2.06}$
analcatdata_dmft	42.52 ± 7.66	40.59 ± 1.37	21.56 ± 0.58	40.01 ± 14.01	$\mathbf{19.62 \pm 0.28}$
analcatdata_olympic2000	46.67 ± 4.99	47.88 ± 6.11	31.67 ± 5.37	41.36 ± 3.78	$\mathbf{29.55 \pm 3.99}$
AP_Breast_Colon	49.46 ± 2.56	49.56 ± 2.61	$\mathbf{48.37 \pm 1.31}$	49.38 ± 1.90	48.79 ± 3.70
AP_Colon_Ovary	49.57 ± 2.32	50.56 ± 3.05	44.73 ± 0.81	45.74 ± 2.27	$\mathbf{41.40 \pm 0.59}$
AP_Lung_Kidney	46.79 ± 2.91	47.05 ± 2.95	35.16 ± 1.64	57.33 ± 6.43	$\mathbf{33.94 \pm 1.21}$
AP_Omentum_Lung	46.95 ± 2.57	48.13 ± 2.83	$\mathbf{41.43 \pm 2.29}$	47.64 ± 5.44	43.20 ± 8.70
AP_Prostate_Kidney	41.00 ± 5.67	41.49 ± 1.84	22.95 ± 0.93	45.59 ± 17.84	$\mathbf{21.28 \pm 0.45}$
AP_Prostate_Ovary	45.62 ± 4.77	44.68 ± 2.49	28.20 ± 0.71	47.75 ± 11.65	$\mathbf{26.25 \pm 0.67}$
artificial-characters	89.79 ± 0.76	89.30 ± 0.50	86.32 ± 1.25	88.48 ± 0.53	$\mathbf{85.66 \pm 2.08}$
Australian	38.32 ± 1.70	38.07 ± 1.58	17.58 ± 1.40	29.14 ± 1.43	$\mathbf{15.64 \pm 0.61}$
auto_price	49.75 ± 6.24	48.49 ± 4.21	38.30 ± 1.84	40.82 ± 3.48	$\mathbf{30.50 \pm 4.06}$
balance-scale	53.79 ± 2.13	53.30 ± 2.22	20.75 ± 2.90	39.25 ± 1.88	$\mathbf{16.27 \pm 0.90}$
balloon	44.31 ± 6.14	42.15 ± 1.13	23.43 ± 0.57	32.60 ± 0.65	$\mathbf{22.15 \pm 0.70}$
banana	42.04 ± 0.34	41.59 ± 0.54	23.62 ± 1.17	33.14 ± 0.82	$\mathbf{22.36 \pm 1.17}$
blood-transfusion-service-center	41.95 ± 1.59	44.39 ± 0.99	25.45 ± 0.97	39.25 ± 9.97	$\mathbf{23.82 \pm 0.58}$
car	67.76 ± 6.99	63.56 ± 1.40	25.60 ± 0.77	46.77 ± 1.08	$\mathbf{24.26 \pm 1.56}$
chscase_health	$\mathbf{44.40 \pm 6.38}$	47.60 ± 5.72	47.80 ± 5.77	49.80 ± 4.66	48.80 ± 2.15
Click_prediction_small	42.00 ± 9.28	32.08 ± 0.10	4.67 ± 0.01	23.17 ± 5.68	$\mathbf{4.47 \pm 0.00}$
climate-model-simulation-crashes	32.85 ± 2.28	34.72 ± 2.37	9.37 ± 0.46	23.70 ± 1.47	$\mathbf{8.69 \pm 0.16}$
cmc	65.93 ± 0.62	66.07 ± 0.97	60.10 ± 1.42	62.11 ± 1.39	$\mathbf{57.73 \pm 0.43}$
codrna	32.13 ± 0.05	31.64 ± 0.04	5.19 ± 0.03	19.64 ± 0.04	$\mathbf{5.13 \pm 0.03}$
covertype (2 classes)	41.06 ± 0.14	40.47 ± 0.12	23.23 ± 0.14	32.32 ± 0.12	$\mathbf{22.71 \pm 0.15}$
covertype (7 classes)	73.52 ± 0.09	73.25 ± 0.11	27.54 ± 0.23	51.75 ± 0.16	$\mathbf{25.85 \pm 0.27}$
credit-g	43.82 ± 1.34	44.20 ± 1.19	29.11 ± 0.92	37.04 ± 0.92	$\mathbf{28.06 \pm 0.91}$
diabetes	46.58 ± 3.41	48.10 ± 1.58	38.83 ± 1.90	43.15 ± 4.78	$\mathbf{35.90 \pm 0.83}$
ecoli (2 classes)	39.73 ± 4.15	39.14 ± 4.00	19.32 ± 2.95	30.95 ± 4.11	$\mathbf{17.50 \pm 2.48}$
ecoli (8 classes)	85.12 ± 1.69	85.30 ± 2.00	62.92 ± 5.05	80.48 ± 6.66	$\mathbf{54.23 \pm 7.14}$
eeg-eye-state	50.25 ± 1.11	49.84 ± 0.35	46.86 ± 0.26	47.83 ± 2.09	$\mathbf{44.88 \pm 0.01}$
electricity	42.65 ± 0.17	41.89 ± 0.20	26.60 ± 0.25	34.73 ± 0.14	$\mathbf{25.91 \pm 0.21}$
fertility	41.60 ± 6.67	38.80 ± 4.26	14.00 ± 2.45	31.10 ± 10.90	$\mathbf{13.40 \pm 1.96}$
first-order-theorem-proving	83.49 ± 0.65	80.11 ± 0.44	61.88 ± 0.40	69.61 ± 0.43	$\mathbf{59.20 \pm 0.45}$
fl2000	48.51 ± 7.68	49.10 ± 7.46	49.55 ± 4.97	46.87 ± 3.87	$\mathbf{41.49 \pm 4.27}$
flags	85.10 ± 3.70	86.96 ± 2.31	82.58 ± 3.67	86.65 ± 4.83	$\mathbf{78.66 \pm 5.17}$
fri_c0_100_5	43.50 ± 6.10	43.60 ± 6.19	28.60 ± 3.78	37.20 ± 5.27	$\mathbf{24.80 \pm 4.16}$
fri_c0_100_50	50.80 ± 6.12	52.30 ± 5.12	48.30 ± 4.67	$\mathbf{47.80 \pm 2.66}$	49.40 ± 4.84
fri_c0_250_5	45.44 ± 3.43	45.16 ± 3.20	26.88 ± 3.02	37.08 ± 3.57	$\mathbf{24.36 \pm 2.30}$
fri_c0_500_10	40.46 ± 1.62	42.20 ± 1.82	21.68 ± 1.97	31.94 ± 1.53	$\mathbf{21.00 \pm 1.82}$
fri_c0_500_5	41.86 ± 1.84	41.88 ± 1.92	21.34 ± 1.26	32.08 ± 1.76	$\mathbf{19.46 \pm 1.32}$
fri_c1_1000_10	44.79 ± 1.50	44.66 ± 1.63	30.32 ± 2.23	39.37 ± 2.27	$\mathbf{28.44 \pm 1.71}$
fri_c1_100_10	48.90 ± 7.20	51.10 ± 6.85	48.80 ± 5.87	45.00 ± 6.60	$\mathbf{37.00 \pm 4.88}$
fri_c1_500_50	49.72 ± 3.01	48.44 ± 1.79	44.28 ± 1.60	46.48 ± 1.53	$\mathbf{41.84 \pm 1.43}$
fri_c4_1000_10	45.06 ± 1.83	44.90 ± 1.88	34.05 ± 1.60	38.66 ± 1.57	$\mathbf{31.29 \pm 1.42}$
fri_c4_1000_25	47.90 ± 1.21	47.93 ± 1.19	41.17 ± 0.84	44.50 ± 1.15	$\mathbf{39.26 \pm 0.96}$
fri_c4_1000_50	47.85 ± 1.05	49.05 ± 0.57	40.23 ± 1.81	43.75 ± 1.31	$\mathbf{39.21 \pm 1.93}$
gina_prior2	90.22 ± 0.64	90.20 ± 0.55	88.49 ± 2.01	89.34 ± 1.13	$\mathbf{84.11 \pm 3.83}$
glass	84.25 ± 3.41	84.35 ± 2.90	75.47 ± 5.08	78.83 ± 5.74	$\mathbf{70.65 \pm 4.76}$
haberman	47.52 ± 5.95	44.18 ± 2.44	$\mathbf{29.15 \pm 0.72}$	38.95 ± 8.10	40.98 ± 22.17

Table 4 continued.

Dataset	SquareCB	SquareCB.L	FastCB	SpannerIGW	MinCE
hayes-roth (2 classes)	48.94 ± 4.67	48.71 ± 4.04	43.03 ± 2.31	48.56 ± 6.70	$\mathbf{39.70 \pm 1.48}$
hayes-roth (4 classes)	73.81 ± 3.19	75.19 ± 3.03	65.75 ± 2.81	74.25 ± 9.15	$\mathbf{59.88 \pm 0.87}$
heart-statlog	45.67 ± 3.56	45.52 ± 3.81	30.44 ± 3.20	37.63 ± 1.92	$\mathbf{27.48 \pm 1.74}$
IMDB.drama	49.94 ± 4.62	47.56 ± 0.12	37.85 ± 0.03	47.46 ± 8.11	$\mathbf{36.21 \pm 0.00}$
ionosphere	41.94 ± 1.98	42.02 ± 1.75	21.82 ± 1.96	33.39 ± 1.61	$\mathbf{19.72 \pm 2.09}$
iris	62.67 ± 6.46	62.60 ± 6.10	46.07 ± 9.81	62.00 ± 7.59	$\mathbf{41.40 \pm 5.97}$
isolet	96.12 ± 0.18	96.13 ± 0.20	96.12 ± 0.23	$\mathbf{96.11 \pm 0.16}$	96.12 ± 0.06
JapaneseVowels	85.57 ± 1.35	83.44 ± 1.73	$\mathbf{47.79 \pm 8.32}$	77.31 ± 3.01	57.16 ± 13.10
kc1	38.98 ± 4.85	38.64 ± 1.20	16.79 ± 0.30	29.24 ± 4.71	$\mathbf{15.75 \pm 0.37}$
kc2	49.79 ± 4.56	42.68 ± 1.53	21.23 ± 1.34	32.07 ± 1.25	$\mathbf{18.97 \pm 0.54}$
kc3	35.46 ± 1.46	37.47 ± 1.40	10.70 ± 0.67	27.66 ± 8.01	$\mathbf{9.80 \pm 0.51}$
KDDCup99	79.71 ± 0.06	79.66 ± 0.06	$\mathbf{0.88 \pm 0.12}$	43.86 ± 0.06	1.63 ± 1.01
kr-vs-kp	36.82 ± 0.63	35.91 ± 0.61	12.73 ± 0.69	25.77 ± 0.83	$\mathbf{12.18 \pm 0.76}$
KungChi3	36.26 ± 2.58	40.89 ± 3.91	14.80 ± 1.07	28.54 ± 2.67	$\mathbf{13.82 \pm 1.01}$
la2s.wc	78.64 ± 1.26	78.67 ± 1.62	44.12 ± 4.50	62.87 ± 1.78	$\mathbf{38.87 \pm 4.75}$
leaf	97.06 ± 1.15	96.97 ± 0.99	97.18 ± 1.11	$\mathbf{96.15 \pm 0.66}$	96.79 ± 0.73
letter	96.07 ± 0.07	96.07 ± 0.16	95.96 ± 0.28	96.13 ± 0.19	$\mathbf{95.30 \pm 1.39}$
leukemia	43.19 ± 4.60	45.97 ± 4.46	38.75 ± 3.79	40.56 ± 2.68	$\mathbf{36.53 \pm 1.74}$
lymph	72.70 ± 5.01	72.64 ± 2.71	70.88 ± 3.08	69.66 ± 7.32	$\mathbf{42.43 \pm 3.90}$
machine_cpu	40.24 ± 3.04	39.28 ± 3.14	18.52 ± 7.26	30.43 ± 3.91	$\mathbf{15.84 \pm 6.23}$
MagicTelescope	40.43 ± 0.32	39.93 ± 0.33	21.24 ± 0.15	31.06 ± 0.25	$\mathbf{20.52 \pm 0.14}$
mammography	33.07 ± 5.09	30.53 ± 0.39	$\mathbf{1.96 \pm 0.07}$	17.60 ± 0.37	5.71 ± 12.22
mc1	29.87 ± 0.44	30.02 ± 0.35	0.78 ± 0.02	16.53 ± 0.33	$\mathbf{0.74 \pm 0.04}$
mc2	45.40 ± 4.12	46.02 ± 2.91	37.08 ± 2.62	41.12 ± 3.26	$\mathbf{34.04 \pm 2.88}$
mfeat-factors	90.28 ± 0.70	90.30 ± 0.39	90.20 ± 0.51	$\mathbf{89.81 \pm 0.64}$	90.64 ± 0.57
mfeat-morphological	30.73 ± 2.17	29.51 ± 1.04	1.57 ± 1.46	17.41 ± 0.86	$\mathbf{0.68 \pm 0.18}$
mnist_784	83.74 ± 1.68	79.97 ± 0.93	39.14 ± 5.01	63.71 ± 4.57	$\mathbf{37.13 \pm 17.55}$
monks-problems-1	47.09 ± 1.90	47.12 ± 1.91	34.78 ± 1.43	40.49 ± 1.28	$\mathbf{33.47 \pm 1.51}$
monks-problems-2	48.00 ± 4.72	47.92 ± 2.27	37.32 ± 0.87	42.16 ± 3.96	$\mathbf{34.76 \pm 0.51}$
monks-problems-3	43.39 ± 2.25	43.25 ± 2.31	27.89 ± 0.91	35.79 ± 1.72	$\mathbf{26.71 \pm 1.13}$
mozilla4	36.06 ± 0.17	35.82 ± 0.20	12.66 ± 0.16	25.01 ± 0.25	$\mathbf{12.25 \pm 0.15}$
mw1	33.92 ± 1.74	36.33 ± 1.29	9.23 ± 0.75	23.37 ± 1.38	$\mathbf{8.01 \pm 0.26}$
nomao	33.71 ± 0.27	32.92 ± 0.26	7.59 ± 0.14	21.63 ± 0.23	$\mathbf{7.40 \pm 0.12}$
nursery	62.53 ± 0.53	62.09 ± 0.77	12.16 ± 0.33	40.61 ± 0.65	$\mathbf{11.60 \pm 0.34}$
oh15.wc	90.27 ± 1.16	90.57 ± 0.83	88.18 ± 2.09	86.69 ± 2.90	$\mathbf{81.54 \pm 6.13}$
oh5.wc	89.39 ± 0.73	89.61 ± 0.97	87.70 ± 2.18	89.88 ± 1.78	$\mathbf{86.59 \pm 4.12}$
optdigits	90.13 ± 0.39	89.77 ± 0.57	87.79 ± 2.96	88.55 ± 1.74	$\mathbf{73.10 \pm 7.93}$
OVA_Lung	34.41 ± 5.24	34.41 ± 1.09	8.80 ± 0.26	42.18 ± 25.63	$\mathbf{8.27 \pm 0.14}$
ozone-level-8hr	33.55 ± 4.03	33.59 ± 0.84	6.80 ± 0.18	21.52 ± 1.85	$\mathbf{6.41 \pm 0.07}$
page-blocks	59.63 ± 2.44	59.41 ± 0.59	8.40 ± 0.24	37.67 ± 0.50	$\mathbf{8.06 \pm 0.20}$
parkinsons	46.77 ± 8.09	47.85 ± 3.85	27.54 ± 0.69	51.95 ± 17.18	$\mathbf{25.59 \pm 3.10}$
pasture	71.67 ± 4.30	70.83 ± 5.75	70.28 ± 6.81	66.94 ± 7.69	$\mathbf{63.89 \pm 7.29}$
pc1	32.47 ± 0.60	33.95 ± 0.74	7.46 ± 0.18	25.69 ± 6.07	$\mathbf{7.04 \pm 0.14}$
pc2	34.04 ± 8.51	30.09 ± 0.64	0.46 ± 0.01	16.66 ± 0.41	$\mathbf{0.42 \pm 0.01}$
pc3	35.52 ± 2.89	36.28 ± 0.65	10.97 ± 0.21	25.14 ± 2.55	$\mathbf{10.64 \pm 0.39}$
pc4	40.67 ± 7.38	36.85 ± 0.92	12.77 ± 0.38	25.37 ± 0.99	$\mathbf{11.90 \pm 0.36}$
pendigits	89.88 ± 0.44	87.83 ± 0.86	79.55 ± 8.34	83.97 ± 2.42	$\mathbf{72.90 \pm 11.86}$
phoneme	41.34 ± 0.68	41.34 ± 0.67	23.78 ± 0.36	32.94 ± 0.52	$\mathbf{23.00 \pm 0.34}$
PopularKids	63.05 ± 3.35	64.96 ± 2.67	53.93 ± 1.08	59.25 ± 4.87	$\mathbf{49.96 \pm 1.30}$
RandomRBF_10_1E-4	63.66 ± 0.08	63.06 ± 0.09	20.08 ± 0.24	43.77 ± 0.23	$\mathbf{19.74 \pm 1.22}$
ringnorm	41.09 ± 0.70	40.65 ± 0.70	22.56 ± 0.63	32.04 ± 0.68	$\mathbf{22.19 \pm 0.71}$
rsctc2010_3	78.84 ± 3.66	78.95 ± 3.75	79.26 ± 2.05	79.16 ± 2.52	$\mathbf{76.42 \pm 4.47}$
sa-heart	46.77 ± 5.11	46.93 ± 1.48	35.97 ± 1.56	41.17 ± 1.91	$\mathbf{33.85 \pm 2.07}$
satimage	69.05 ± 0.58	68.62 ± 0.54	21.57 ± 0.63	47.66 ± 0.55	$\mathbf{19.03 \pm 0.66}$
segment	83.06 ± 0.99	83.85 ± 1.48	64.04 ± 8.75	76.09 ± 4.26	$\mathbf{57.77 \pm 10.09}$
sleuth_case1201	51.80 ± 5.45	51.80 ± 5.45	$\mathbf{41.40 \pm 10.67}$	46.20 ± 8.30	47.40 ± 10.71
sleuth_ex2016	48.39 ± 4.45	$\mathbf{47.59 \pm 5.00}$	48.74 ± 4.47	48.62 ± 3.60	48.39 ± 1.48
sonar	49.95 ± 3.50	50.53 ± 3.20	50.05 ± 3.36	50.53 ± 3.92	$\mathbf{49.04 \pm 1.98}$
spambase	34.54 ± 0.84	34.03 ± 0.85	9.51 ± 0.38	23.27 ± 0.67	$\mathbf{9.45 \pm 0.39}$
SPECT	41.57 ± 4.60	40.67 ± 2.87	24.68 ± 2.09	32.28 ± 1.97	$\mathbf{18.91 \pm 0.87}$
SPECTF	43.95 ± 3.30	45.56 ± 2.89	29.48 ± 0.60	42.58 ± 10.02	$\mathbf{27.68 \pm 0.53}$
spectrometer	37.29 ± 7.09	35.57 ± 1.17	11.45 ± 0.48	24.95 ± 1.88	$\mathbf{10.58 \pm 0.21}$
splice	56.43 ± 1.25	54.01 ± 1.52	23.89 ± 1.73	41.26 ± 1.57	$\mathbf{20.06 \pm 1.07}$
Stagger3	29.30 ± 0.05	29.30 ± 0.05	0.00 ± 0.00	15.66 ± 0.04	$\mathbf{0.00 \pm 0.00}$
sylva_prior	30.24 ± 0.40	29.95 ± 0.42	1.07 ± 0.06	16.80 ± 0.26	$\mathbf{0.99 \pm 0.02}$
synthetic_control	82.50 ± 0.52	82.78 ± 0.60	83.23 ± 1.06	$\mathbf{82.17 \pm 1.64}$	83.20 ± 2.05
tae	67.42 ± 2.86	$\mathbf{65.83 \pm 2.32}$	66.49 ± 3.29	66.29 ± 5.95	67.09 ± 1.21
tic-tac-toe	49.29 ± 3.30	48.79 ± 2.33	36.91 ± 0.84	51.14 ± 8.67	$\mathbf{34.72 \pm 0.55}$
transplant	40.61 ± 3.63	39.39 ± 3.85	$\mathbf{20.92 \pm 4.48}$	33.44 ± 3.09	22.29 ± 8.35
vehicle	75.39 ± 1.58	75.35 ± 1.42	75.71 ± 1.01	75.34 ± 1.10	$\mathbf{75.11 \pm 1.51}$
vehicle_sensIT	37.58 ± 0.17	37.39 ± 0.19	15.15 ± 0.05	26.90 ± 0.10	$\mathbf{14.58 \pm 0.05}$
visualizing_galaxy	34.74 ± 1.25	33.78 ± 1.13	10.00 ± 1.21	25.63 ± 0.92	$\mathbf{8.76 \pm 1.15}$

Table 4 continued.

Dataset	SquareCB	SquareCB.L	FastCB	SpannerIGW	MinCE
vowel	91.33 ± 0.98	91.42 ± 1.10	$\mathbf{90.62 \pm 0.74}$	90.89 ± 0.74	91.03 ± 0.95
waveform-5000	53.51 ± 0.83	52.62 ± 0.75	21.43 ± 1.09	38.93 ± 0.92	$\mathbf{19.33 \pm 0.59}$
wine	65.22 ± 4.21	64.72 ± 5.22	63.26 ± 5.14	64.33 ± 3.96	$\mathbf{59.27 \pm 3.86}$
yeast	88.86 ± 2.56	88.23 ± 1.00	77.04 ± 1.47	83.30 ± 5.37	$\mathbf{68.91 \pm 0.12}$
zoo	83.47 ± 5.54	82.48 ± 5.58	60.00 ± 8.62	71.98 ± 5.03	$\mathbf{49.70 \pm 8.37}$