ADAPTIVE SELF-CONSISTENCY: FROM BLACK-BOX SAMPLING TO DISTRIBUTION-VALUED FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-consistency samples many reasoning trajectories and aggregates their final answers, treating the LLM as a black box that returns one answer per trajectory. Yet the final answer of each trajectory is sampled from a softmax vector that is available from the model's log-probabilities. We refer to this as the grey-box setting in which each trajectory reveals this answer distribution rather than a single draw from it. We formulate efficient inference in this setting as sequential mode identification with distribution-valued observations: sample trajectories one at a time and stop as soon as the LLM's modal answer is identified at a prescribed confidence level. We characterize the asymptotic stopping rate of mode identification with distribution-valued observations exactly and show that it is never worse than the black-box rate. We then propose the ASC-D algorithm, a betting stopping rule that attains this asymptotic stopping rate. On MMLU-Redux, ASC-D uses 46.4–95.6% fewer trajectories than answer-only adaptive self-consistency baselines and achieves the highest fixed-budget correct-certification rate across three open-source models.

1 INTRODUCTION

Large language models (LLMs) have demonstrated strong performance on challenging mathematical and logical reasoning tasks. When solving a complex problem, an LLM often produces a sequence of intermediate reasoning steps before giving its final answer, a generation pattern commonly known as chain-of-thought reasoning (Wei et al., 2022; Kojima et al., 2022). Because decoding is stochastic, repeatedly querying the same LLM can produce different reasoning trajectories and different final answers. Self-consistency (SC) exploits this diversity by sampling multiple reasoning trajectories and returning their most frequent final answer (Wang et al., 2022). Because each trajectory requires a separate LLM generation, the inference cost of SC grows with its sampling budget. Standard SC fixes this budget in advance, thus it may cause a computation waste on easy problems that require only a few trajectories. Adaptive self-consistency (ASC) addresses this inefficiency by sampling trajectories sequentially and stopping once the observed answers satisfy sufficient agreement (Aggarwal et al., 2023).

Both SC and ASC retain only the final-answer label from each reasoning trajectory, which we refer to as the *black-box* observation model. Nevertheless, at the end of a trajectory, the probability vector from which the LLM decodes its final answer (token) is readily available and informative. Specifically, when Llama-3.2-3B-Instruct answers a four-choice ARC-Challenge question, we can read the next-token softmax over the four option tokens at the final-answer position, e.g., one recorded trajectory gives $(0.26, 0.43, 0.21, 0.10)$ for options A–D, see Appendix C for details on how we obtain the probabilities. This raises a natural question: can such distribution-valued feedback further reduce the number of trajectories required for reliable self-consistency?

We address this question by introducing adaptive self-consistency with *distribution-valued feedback*. Figure 1 contrasts it with the black-box model. Both start from the same reasoning process: a prompt x yields a sampled trajectory r, which induces a probability vector $\boldsymbol{\theta}$ over the K candidate answers. The black-box model observes only one answer label drawn from $\boldsymbol{\theta}$, whereas our setting observes the vector itself (as well as the answer label). The two settings share the same answer frequencies and the same modal answer, and differ only in how much of $\boldsymbol{\theta}$ they reveal. Thus, we call our observation model *grey-box*. In a *white-box* setting, where the token-level probabilities for **all**

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.

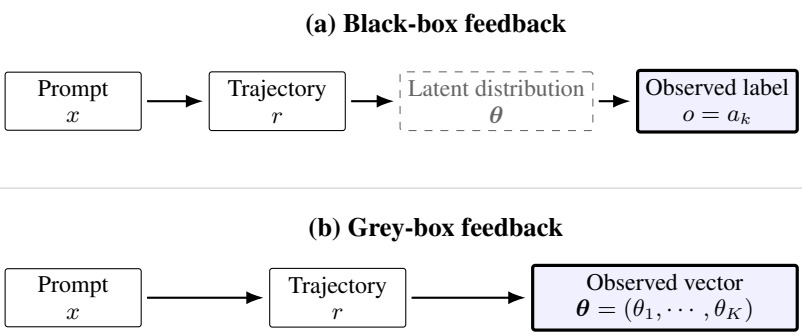

Figure 1: Black-box and grey-box feedback from the same reasoning process.

possible reasoning trajectories are available, we could in principle identify the modal answer without repeated sampling by directly calculating the final answer probabilities. However, enumerating and evaluating all possible trajectories is generally infeasible. Grey-box feedback therefore offers a practical middle ground between black-box and white-box: it retains trajectory sampling while using each trajectory's probabilities for all candidate answers.

Building on this richer observation model, we study how much grey-box feedback can reduce the number of trajectories required for reliable mode identification and whether this improvement can be attained by a practical adaptive stopping rule. We summarize our main contributions below.

1. **The benefit of grey-box feedback.** For general distributions of answer-probability vectors, we derive a lower bound on the asymptotic sample complexity of any algorithm and show that grey-box feedback never requires more trajectories than black-box feedback. Specifically, for small frequency gap Δ between the modal answer and the runner-up, the sample complexity ranges from the black-box order $\Theta(\Delta^{-2})$, when every trajectory commits to a single answer, down to $\Theta(\Delta^{-1})$, when trajectories report the same beliefs.

2. **An algorithm that fully realizes the benefit.** We develop an algorithm for adaptive self-consistency with distribution-valued feedback (ASC-D), the grey-box counterpart of ASC. Instead of counting answer labels, ASC-D bets on pairwise differences of the probability vectors and stops once the leading answer has accumulated enough evidence against every alternative. It controls the probability of returning a non-modal answer at level δ at every stopping time, and its expected number of trajectories attains the lower bound asymptotically, without any knowledge of the distribution of the probability vectors. The gain identified above is therefore attained in full by a practical procedure.

3. **Empirical validation.** We validate our theoretical predictions through synthetic experiments and evaluations on MMLU-Redux. On MMLU-Redux, ASC-D uses 46.4–95.6% fewer trajectories than answer-only baselines as the modal gap narrows. Under a fixed maximum budget of $N = 128$, it achieves the highest correct-certification rate at every tested confidence level across Llama-3.2-1B, Llama-3.1-8B, and Qwen3-1.7B.

Organization The rest of the paper is organized as follows. Section 2 presents the problem setup, derives and compares asymptotic lower bounds on the expected stopping time under different observation models. Section 3 introduces the ASC-D algorithm and its theoretical guarantees. Section 4 reports the experiment results on both synthetic and real-world datasets, and Section 5 concludes with a discussion of limitations. Appendix A reviews related literature. Appendix B provides the proof of main results in the paper. Appendix C gives a real-world example of grey-box feedback. Appendices D and E provide omitted details and additional results for the synthetic and real-world experiments, respectively.

2 PROBLEM SETUP

2.1 DISTRIBUTION-VALUED FEEDBACK

For a fixed prompt x with a known set of candidate answers $\mathcal{A} = \{a_1, \cdots, a_K\}$, each LLM call first generates a reasoning trajectory r_n and then produces a final answer. Formally, conditional on r_n and x, the model assigns a probability to each candidate answer, yielding the trajectory-specific distribution

$$\boldsymbol{\theta}_n := \mathbb{P}(\cdot \mid r_n, x) \in \Delta^{K-1}, \qquad \theta_{n,k} := \mathbb{P}(a_k \mid r_n, x).$$

Standard black-box self-consistency returns only the realized answer $o_n \mid \boldsymbol{\theta}_n \sim \mathrm{Cat}(\boldsymbol{\theta}_n)$ sampled from the categorical distribution with parameter $\boldsymbol{\theta}_n$, whereas the grey-box setting observes the entire vector $\boldsymbol{\theta}_n$. Because different reasoning trajectories can induce different answer distributions, these vectors may vary across calls. We write $\boldsymbol{\theta}$ for a generic trajectory-specific distribution.

General Model Since trajectories are sampled independently, the vectors $(\boldsymbol{\theta}_n)_{n\geq 1}$ are i.i.d. drawn from an unknown distribution F on Δ^{K-1}. Let $\mathcal{M}$ denote the set of all distributions on Δ^{K-1} and

$$\boldsymbol{\pi} := \mathbb{E}_F[\boldsymbol{\theta}] = (p_1, \cdots, p_K)$$

for the mean vector of F. We assume throughout that F has a unique mode, and index the candidate answers so that $p_1 > p_2 \geq \cdots \geq p_K$, and a_1 is the modal answer. In this model, the black-box answers are i.i.d. with $\mathbb{P}(o_n = a_k) = \mathbb{E}_F[\theta_k] = p_k$. Thus, $\boldsymbol{\pi}$ is the answer-frequency vector of black-box self-consistency, and both observation models target the same modal answer a_1.

2.2 SEQUENTIAL MODE IDENTIFICATION

The goal of self-consistency is to identify the modal answer a_1 of the LLM, which is the answer that a majority vote over infinitely many trajectories would return. A sequential procedure samples trajectories one at a time, observes $(\boldsymbol{\theta}_n)_{n\geq 1}$ in the grey-box setting or $(o_n)_{n\geq 1}$ in the black-box setting, and consists of a stopping time τ with respect to the observations together with a recommendation $\hat{a}_\tau$. Stopping is the only decision: every trajectory costs one LLM call, and the procedure should stop as soon as the mode is identified at the prescribed error probability δ.

δ-PAC Procedures A procedure is *δ-PAC over $\mathcal{M}$* if

$$\mathbb{P}_F(\hat{a}_\tau \neq a_1) \leq \delta \qquad \text{for every } F \in \mathcal{M} \text{ with a unique mode,}$$

that is, its error guarantee holds whatever the trajectories do.

Characteristic Time Following Garivier & Kaufmann (2016); Jain et al. (2022), we measure the efficiency of a δ-PAC procedure by the asymptotic constant of its expected stopping time as the confidence grows, $\liminf_{\delta\to 0} \mathbb{E}_F[\tau]/\log(1/\delta)$, and define the *characteristic time* of F as the best constant over all δ-PAC procedures,

$$T^\star(F) := \liminf_{\delta\to 0} \frac{\mathbb{E}_F[\tau]}{\log(1/\delta)}.$$

By definition, every δ-PAC procedure needs at least $T^\star(F)\cdot\log(1/\delta)$ trajectories on F asymptotically before stopping. Throughout we write

$$\Delta := p_1 - p_2, \qquad \bar{p} := \frac{p_1+p_2}{2}, \qquad X_j := \theta_1 - \theta_j \in [-1,1] \quad (j \neq 1), \tag{1}$$

where Δ is the gap between the top two answers and X_j is the margin of the mode over its j-th challenger in a single trajectory.

The Characteristic Time under Black-Box In the black-box setting, the observations are i.i.d. $\mathrm{Cat}(\boldsymbol{\pi})$ whatever the law F, so the procedure sees only $\boldsymbol{\pi}$. Shah et al. (2020); Jain et al. (2022) show that the characteristic time of mode identification from categorical samples is

$$T_{\mathrm{bb}}(F) := \frac{1}{p_1 \cdot \log\frac{2p_1}{p_1+p_2} + p_2 \cdot \log\frac{2p_2}{p_1+p_2}} = \frac{4\bar{p}}{\Delta^2}\cdot\big(1+O(\Delta^2)\big), \tag{2}$$

where the expansion holds as $\Delta \to 0$, see Appendix B for details. We write $T_{\mathrm{bb}}(F)$ for this black-box characteristic time, which depends on F only through $\boldsymbol{\pi} = \mathbb{E}_F[\boldsymbol{\theta}]$. The denominator is the Kullback–Leibler divergence from $\mathrm{Cat}(\boldsymbol{\pi})$ to the closest categorical law in which a_2 ties with a_1; the quadratic dependence on Δ is the price of resolving a small gap from the hard votes. Since grey-box feedback can always be reduced to black-box feedback by drawing $o \sim \mathrm{Cat}(\boldsymbol{\theta})$, (2) is an upper bound on the grey-box characteristic time. Whether, and by how much, grey-box feedback improves on this bound is the question we turn to next.

2.3 Asymptotic Lower Bounds

How much the grey-box feedback gain depends on how the probability vectors spread around their mean $\boldsymbol{\pi}$, and two extreme laws with this mean bracket the possibilities. At one end, the *degenerate law* $\delta_{\boldsymbol{\pi}}$ puts all its mass at $\boldsymbol{\pi}$: every trajectory reports the same beliefs. At the other end, the *vertex law* $F_0 := \sum_{k=1}^{K} p_k \cdot \delta_{\boldsymbol{e}_k}$ puts $\boldsymbol{\theta}$ at a vertex of the simplex with probability p_k: every trajectory commits to one answer, and $\boldsymbol{\theta}$ carries no more information than the sampled label o. The following theorem gives the grey-box characteristic time for a general F and shows that these two laws are exactly the two extremes of the characteristic time.

Theorem 2.1 (Asymptotic Lower Bound under Grey-box Feedback)**.** *For every δ-PAC procedure and every $F \in \mathcal{M}$ with a unique mode, under the grey-box feedback model,*

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_F[\tau]}{\log(1/\delta)} \geq T_{\mathrm{gb}}(F) := \frac{1}{\min_{j=2,\cdots,K} \mathcal{G}_j(F)}, \quad \mathcal{G}_j(F) := \max_{\lambda \in [0,1]} \mathbb{E}_F\left[\log\left(1 + \lambda \cdot (\theta_1 - \theta_j)\right)\right]. \tag{3}$$

where we use $T_{\mathrm{gb}}(F)$ to denote the characteristic time for distribution F under the grey-box feedback model. Moreover,

$$T_{\mathrm{gb}}(\delta_{\boldsymbol{\pi}}) = \frac{1}{\log(1+\Delta)} \leq T_{\mathrm{gb}}(F) \leq T_{\mathrm{gb}}(F_0) = T_{\mathrm{bb}}(F), \tag{4}$$

i.e., the two endpoints are the grey-box characteristic times of the degenerate law and the vertex law, the most and the least consistent laws with the same answer frequencies.

For the first part of the theorem: a standard change-of-measure argument (Kaufmann et al., 2016) gives a lower bound through KL minimization over alternative distributions (in which a_1 is not the unique mode). Making a competitor a_j at least as likely as a_1 is equivalent to making the mean of $X_j = \theta_1 - \theta_j$ nonpositive. We show that it suffices to optimize over distributions of this margin. By utilizing the dual representation of Honda & Takemura (2015), we then turn this infinite-dimensional optimization over probability distributions into the one-dimensional maximization over $\lambda \in [0, 1]$ in (3). Thus, $\mathcal{G}_j(F)$ is the minimum KL divergence needed for a_j to be at least as likely as a_1, and the hardest competitor determines the characteristic time. The sandwich bound in (4) can then be derived simply through the Jensen's inequality.

The Value of the Grey-Box Observation Model Under the vertex law F_0, every trajectory commits to a single answer and the grey-box observation carries no more information than the sampled answer; accordingly $T_{\mathrm{gb}}(F_0) = T_{\mathrm{bb}}(F)$, and the upper bound is attained exactly at this black-box endpoint. Any law that keeps mass away from the vertices stops strictly faster. The gain is bounded, however. Even when every trajectory reports $\boldsymbol{\pi}$ exactly, a δ-PAC procedure cannot exclude the law that agrees with $\delta_{\boldsymbol{\pi}}$ on a fraction $1/(1+\Delta)$ of the trajectories and commits to a_2 on the rest, whose divergence from $\delta_{\boldsymbol{\pi}}$ is $\log(1+\Delta)$, and it says that without further assumptions the dependence on the gap improves from $\Theta(\Delta^{-2})$ to at best $\Theta(\Delta^{-1})$.

Connection with Best-of-N Best-of-N is another widely applied method for test-time compute, which samples N trajectories and selects the answer with the highest verifier score (Cobbe et al., 2021). With an ideal verifier, it succeeds whenever a correct answer appears among the samples. Assuming that the modal answer is correct, Huang (2026) show that the worst-case sample complexity scales as $\Theta(\Delta^{-1})$ for best-of-N, compared with $\Theta(\Delta^{-2})$ for self-consistency.

In contrast, our Theorem 2.1 shows that, at the degenerate law $\delta_{\boldsymbol{\pi}}$, grey-box self-consistency has characteristic time $1/\log(1+\Delta) = \Theta(\Delta^{-1})$. Thus, maximally consistent grey-box feedback yields

the same linear inverse-gap dependence as the worst-case benchmark for ideal best-of-N, using the LLM's own answer probabilities available under our distribution-valued observation model, without training or evaluating a separate verifier.

Connection with Confidence-Weighted ASC Between the black-box and grey-box observation models lies confidence-weighted self-consistency (Taubenfeld et al., 2025; Fu et al., 2026; Ota et al., 2026), in which each trajectory reveals its sampled answer together with the confidence it assigns to that answer. We formalize this feedback as observing the sampled answer together with its conditional probability given the reasoning trajectory:

$$o \mid \boldsymbol{\theta} \sim \mathrm{Cat}(\boldsymbol{\theta}), \qquad w := \theta_o \in (0, 1].$$

The pair (o, w) can be generated from $\boldsymbol{\theta}$ and reduces to o once w is dropped. Thus, when all three feedback models are used to identify the same modal answer a_1, the reduction argument of Section 2.2 places the confidence-feedback characteristic time between the black-box and grey-box times. The following proposition characterizes this intermediate rate, where we write Q_F for the law of (o, w) for $\boldsymbol{\theta} \sim F$.

Proposition 2.1 (Asymptotic Lower Bound under Confidence Feedback)**.** *For every $F \in \mathcal{M}$ with unique mode a_1, any procedure that observes (o, w) and is δ-PAC over $\mathcal{M}$ satisfies:*

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_F[\tau]}{\log(1/\delta)} \geq T_{\mathrm{cw}}(F) := \left[\min_{j \neq 1} \inf_{F' \in \mathcal{M} : \mathbb{E}_{F'}[\theta_j] \geq \mathbb{E}_{F'}[\theta_1]} \mathrm{D}_{\mathrm{KL}}(Q_F \| Q_{F'}) \right]^{-1}. \tag{5}$$

Moreover, $T_{\mathrm{gb}}(F) \leq T_{\mathrm{cw}}(F) \leq T_{\mathrm{bb}}(F)$.

Proposition 2.1 shows that confidence feedback can reduce the characteristic time relative to observing only the sampled answer, while revealing the full answer distribution can yield a further reduction.

The Gain under Dirichlet Heterogeneity To further quantify these gains, we now consider a Dirichlet model. We hold the answer frequencies $\boldsymbol{\pi}$ fixed and vary a single concentration parameter c, which controls the agreement across trajectories. This lets us compare all three characteristic times on the same family of instances.

Assumption 1 (Dirichlet Heterogeneity)**.** *Assume $F = \mathrm{Dir}(c\boldsymbol{\pi})$ for a concentration $c > 0$, i.e., $\boldsymbol{\theta} \sim \mathrm{Dir}(cp_1, \cdots, cp_K)$.*

Under Assumption 1, we still have $\mathbb{E}[\theta_k] = p_k$ and

$$\mathrm{Var}(\theta_k) = \frac{p_k \cdot (1 - p_k)}{c + 1}, \qquad \mathrm{Cov}(\theta_k, \theta_j) = -\frac{p_k \cdot p_j}{c + 1} \quad (k \neq j). \tag{6}$$

The parameter c measures how much the trajectories agree with one another. As $c \to \infty$, F converges to the degenerate law $\delta_{\boldsymbol{\pi}}$, and as $c \to 0$ it converges to the vertex law F_0. Between the two endpoints, (6) shows that observing θ_k instead of the one-hot coordinate of o, whose variance is $p_k(1 - p_k)$, reduces the variance by exactly the factor $c + 1$.

Assumption 1 makes the gain explicit through the single parameter c, interpolates between the black-box endpoint F_0 and the fully consistent endpoint $\delta_{\boldsymbol{\pi}}$. Besides, although $T_{\mathrm{gb}}(\mathrm{Dir}(c\boldsymbol{\pi}))$ still has no closed form for general $\boldsymbol{\pi}$, in the hard regime $p_1 \approx p_2$ it reduces to the black-box constant divided by a single factor, as shown in the following proposition.

Proposition 2.2 (Characteristic Time under the Dirichlet Assumption)**.** *Let $F = \mathrm{Dir}(c\boldsymbol{\pi})$. Fix $c > 0$, $\bar{p}$ and $(p_3, \cdots, p_K)$, and let $\Delta \to 0$. Then*

$$T_{\mathrm{gb}}(F) = \frac{T_{\mathrm{bb}}(F)}{c + 1} \cdot \big(1 + O(\Delta)\big), \tag{7}$$

and the maximizer in (3) is $\lambda^\star = \frac{(c+1)\Delta}{2\bar{p}} \cdot (1 + O(\Delta))$ for $j = 2$. For fixed Δ and $c \to \infty$, $T_{\mathrm{gb}}(F) \to 1/\log(1 + \Delta)$.

The following table compares the three feedback models on the same Dirichlet instances. Confidence feedback gives an intermediate gain, with a substantial gap from grey-box feedback at moderate c. Besides, the approximation $T_{\mathrm{bb}}(F)/(c + 1)$ closely matches $T_{\mathrm{gb}}(F)$ for small c. As c increases, however, the approximation tends to zero, while $T_{\mathrm{gb}}(F)$ approaches the limit $1/\log(1+\Delta)$.

Table 1: Characteristic times for $F = \mathrm{Dir}(c\boldsymbol{\pi})$ with $\boldsymbol{\pi} = (0.4, 0.3, 0.2, 0.1)$. For all c, $T_{\mathrm{bb}}(F) = 139.52$ and $1/\log(1+\Delta) = 10.49$. The values of $T_{\mathrm{cw}}(F)$ and $T_{\mathrm{gb}}(F)$ are computed numerically through simulation. The last row is the fixed-c, small-gap approximation in (7).

c	0.01	0.1	1	2	3	5	10	20	50	100	1000
$T_{\mathrm{cw}}(F)$	139.52	139.50	138.22	125.66	109.15	84.99	55.76	33.70	15.62	11.85	10.76
$T_{\mathrm{gb}}(F)$	138.15	126.95	70.33	47.24	35.70	24.17	15.01	12.34	11.16	10.82	10.52
$T_{\mathrm{bb}}(F)/(c+1)$	138.14	126.84	69.76	46.51	34.88	23.25	12.68	6.64	2.74	1.38	0.14

3 The ASC-D Algorithm

3.1 From the Rate to the Algorithm

The rate in Theorem 2.1 tells us which statistic to compute: Consider a gambler who starts with wealth 1 and, on every trajectory i, bets a fraction λ of his wealth on "a_k beats a_j", receiving $1 + \lambda \cdot (\theta_{i,k} - \theta_{i,j})$ per unit bet. After n trajectories his wealth is

$$W_n^{(k,j)}(\lambda) := \prod_{i=1}^{n} \big(1 + \lambda \cdot (\theta_{i,k} - \theta_{i,j})\big), \qquad \lambda \in [0, 1), \tag{8}$$

which is nonnegative because $\theta_{i,k} - \theta_{i,j} \geq -1$, and $\mathcal{G}_j(F)$ is the best expected log-wealth per trajectory of the bet on "a_1 beats a_j". The process has the two properties a stopping statistic needs.

Validity at Stopping Time If a_k does not beat a_j, i.e., $\mathbb{E}_F[\theta_k - \theta_j] \leq 0$, then $\mathbb{E}_F[1 + \lambda(\theta_{i,k} - \theta_{i,j})] \leq 1$, so (8) is a nonnegative supermartingale with initial value 1, and Ville's inequality gives

$$\mathbb{P}_F\Big(\sup_{n \geq 1} W_n^{(k,j)}(\lambda) \geq \frac{1}{\alpha}\Big) \leq \alpha.$$

Wealth above $1/\alpha$ is therefore evidence at level α that a_k beats a_j, and the guarantee holds uniformly over time, which is what a data-dependent stopping rule requires.

Growth at the Optimal Rate If a_1 beats a_j, the law of large numbers gives $\frac{1}{n} \log W_n^{(1,j)}(\lambda) \to \mathbb{E}_F[\log(1 + \lambda(\theta_1 - \theta_j))]$, so at the optimal fraction the wealth grows at rate $\mathcal{G}_j(F)$ and crosses $1/\delta$ after $\log(1/\delta)/\mathcal{G}_j(F)$ trajectories, as illustrated by Theorem 2.1.

However, we note that two quantities in this argument are unknown to the procedure: the optimal fraction λ and the modal answer itself.

Unknown Optimal Fraction The optimal fraction depends on F. As F is itself unknown, we consider to average the wealth over a finite grid $\Lambda \subset [0, 1)$,

$$\overline{W}_n^{(k,j)} := \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} W_n^{(k,j)}(\lambda). \tag{9}$$

A convex combination of supermartingales is a supermartingale, so validity is preserved, and $\log \overline{W}_n^{(k,j)} \geq \max_{\lambda \in \Lambda} \log W_n^{(k,j)}(\lambda) - \log |\Lambda|$, so the averaged wealth grows at the best rate on the grid, $\max_{\lambda \in \Lambda} \mathbb{E}_F[\log(1 + \lambda(\theta_1 - \theta_j))]$, only at the price of a constant that does not affect the rate. The choice of the grid will be discuss in the following subsection.

Unknown Mode Answer As the algorithm cannot know in advance which answer to bet on (i.e., which answer is the modal answer), so it calculates (9) for every ordered pair (k, j). The current leader is the answer with the largest cumulative mass $S_{n,k} := \sum_{i \leq n} \theta_{i,k}$, the grey-box counterpart of the plurality vote, and the procedure stops once the leader's averaged wealth against every challenger exceeds $(K-1)/\delta$. An error means that some $a_k \neq a_1$ is returned, which requires $\overline{W}_n^{(k,1)} \geq (K-1)/\delta$ for a pair in which a_k does not beat a_1. As each of these $K-1$ events has

probability at most $\delta/(K-1)$ by Ville's inequality, and a union bound to the $K-1$ components then gives the error probability δ.

Algorithm 1 provides the detailed procedure. If we use only the sampled answer from each trajectory instead of its full answer distribution, ASC-D reduces to the known-K adaptation of CITE (Ota et al., 2026). Using full answer distributions allows ASC-D to also use the probabilities assigned to answers that were not sampled. Our later analysis will quantify this information gain and shows that ASC-D attains the optimal characteristic time $T_{\mathrm{gb}}(F)$ with a properly designed betting grid.

Algorithm 1 ASC-D: adaptive self-consistency with distribution-valued feedback

Require: prompt x, candidate answers $\{a_1, \cdots, a_K\}$, confidence δ, grid Λ.
$\ell_{k,j}(\lambda) \leftarrow 0$ for all $k \neq j$ and $\lambda \in \Lambda$; $S_k \leftarrow 0$ for all k
for $n = 1, 2, \cdots$ **do**
  sample a trajectory $r_n \sim \mathbb{P}(\cdot \mid x)$ and read off $\boldsymbol{\theta}_n = \mathbb{P}(\cdot \mid r_n, x)$ over the candidate set
  $S_k \leftarrow S_k + \theta_{n,k}$ for all k ▷ cumulative mass of each answer
  $\ell_{k,j}(\lambda) \leftarrow \ell_{k,j}(\lambda) + \log\big(1 + \lambda \cdot (\theta_{n,k} - \theta_{n,j})\big)$ for all $k \neq j$ and $\lambda \in \Lambda$
  $\hat{k} \leftarrow \arg\max_k S_k$ ▷ current leader
  $\log \overline{W}_j \leftarrow \log \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} \exp\big(\ell_{\hat{k},j}(\lambda)\big)$ for all $j \neq \hat{k}$ ▷ averaged against each challenger
  if $\min_{j \neq \hat{k}} \log \overline{W}_j \geq \log \frac{K-1}{\delta}$ **then**
    return $a_{\hat{k}}$
  end if
end for

3.2 Performance Guarantees

Design of the Grid The grid Λ should approximate each challenger's optimal betting fraction $\lambda_j^\star$ in (3). Two insights guide its design (see Appendix B for details). *(i) The middle can be coarse*; *(ii) Both ends should be dense*. We therefore utilize the geometric grid

$$\Lambda_{r,m} := \big\{r^{-i} : i = 1, \cdots, m\big\} \cup \big\{1 - r^{-i} : i = 1, \cdots, m\big\}, \qquad r > 1,\ m \in \mathbb{N}. \tag{10}$$

Asymptotic Upper Bound of ASC-D Define

$$\mathcal{G}_j^\Lambda(F) := \max_{\lambda \in \Lambda} \mathbb{E}_F\big[\log\big(1 + \lambda \cdot (\theta_1 - \theta_j)\big)\big], \qquad T_{\mathrm{gb}}^\Lambda(F) := \frac{1}{\min_{j \neq 1} \mathcal{G}_j^\Lambda(F)}, \tag{11}$$

The following theorem provides the performance guarantees of Algorithm 1.

Theorem 3.1 (Performance Guarantees of ASC-D)**.** *Let $(\tau^{\mathrm{D}}, \hat{a}_{\tau^{\mathrm{D}}})$ be the stopping time and the output of Algorithm 1 with grid $\Lambda \subset [0,1)$, then for every $\delta \in (0,1)$, $\mathbb{P}_F(\tau^{\mathrm{D}} < \infty,\ \hat{a}_{\tau^{\mathrm{D}}} \neq a_1) \leq \delta$.*

(i) Let $\Lambda = \Lambda_{r,m}$ with $r^{-m} \leq \min_{j \neq 1} \lambda_j^\star$. We have $\limsup_{\delta \to 0} \frac{\mathbb{E}_F[\tau^{\mathrm{D}}]}{\log(1/\delta)} \leq T_{\mathrm{gb}}^\Lambda(F) \leq r \cdot T_{\mathrm{gb}}(F)$.

(ii) Let $\Lambda = \Lambda_{r_\delta, m_\delta}$ with $r_\delta := 1 + \big(\log(1/\delta)\big)^{-1/4}$, and $m_\delta := \big\lceil (\log(1/\delta))^{1/2} \big\rceil$. We have

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_F[\tau^{\mathrm{D}}]}{\log(1/\delta)} \leq T_{\mathrm{gb}}(F). \tag{12}$$

Together with Theorem 2.1, (12) gives, for every F with a unique mode, $\lim_{\delta \to 0} \frac{\mathbb{E}_F[\tau^{\mathrm{D}}]}{\log(1/\delta)} = T_{\mathrm{gb}}(F)$. Thus, ASC-D can attain the optimal characteristic time of distribution-valued feedback exactly.

4 Experiments

We evaluate whether distribution-valued feedback improves the sample efficiency of mode identification in both synthetic and real-world settings. In both settings, all methods target the same population mode and are evaluated on the same trajectories. We test sample efficiency in two ways: the number of generations needed to certify the mode at a fixed confidence level (i.e., the stopping time), and the probability of correct certification of the mode within a fixed maximum trajectory budget N (i.e., $\mathbb{P}(\hat{a}_\tau = a_1, \tau \leq N)$). Section 4.1 studies the effect of the feedback type in synthetic datasets, and Section 4.2 tests whether the same gains persist on the probability vectors generated from live LLM calls.

4.1 Controlled Comparison of Feedback Types

We set $K = 4$ and consider $\boldsymbol{\pi}_\Delta = (0.35 + \Delta/2, 0.35 - \Delta/2, 0.15, 0.15)$, where $0 < \Delta \leq 0.4$ is the probability gap between the top-two answers. At each time t, we draw $Z_t \sim \mathrm{Cat}(\boldsymbol{\pi}_\Delta)$, form $\boldsymbol{\theta}_t = \frac{1}{2}\boldsymbol{e}_{Z_t} + \frac{1}{2}\boldsymbol{\pi}_\Delta$, and sample $Y_t \mid \boldsymbol{\theta}_t \sim \mathrm{Cat}(\boldsymbol{\theta}_t)$. This construction gives $\mathbb{E}[\boldsymbol{\theta}_t] = \boldsymbol{\pi}_\Delta$ and $Y_t \sim \mathrm{Cat}(\boldsymbol{\pi}_\Delta)$. We compare the ASC-D with observation of $\boldsymbol{\theta}_t$, a confidence-weighted variant using $\theta_{t,Y_t} \cdot \boldsymbol{e}_{Y_t}$, and CITE-4 (Ota et al., 2026) using $\boldsymbol{e}_{Y_t}$ and its native certification rule. All procedures use the same 15-point grid, and we run 10,000 paired replays per condition.

Table 2: Controlled comparison of feedback types.

(a) Mean stopping time under $1 - \delta = 0.95$

Δ	ASC-D	Conf.-wt.	CITE-4
0.10	**188.42**	444.57	907.72
0.20	**46.20**	99.21	231.54
0.30	**26.79**	46.58	109.82
0.40	**20.62**	32.12	71.72

(b) Correct-certification rate under $\Delta = 0.1$

$1 - \delta$	ASC-D	Conf.-wt.	CITE-4
0.90	**98.44%**	71.37%	30.86%
0.95	**97.49%**	62.49%	24.22%
0.975	**96.40%**	54.83%	18.38%
0.99	**94.23%**	45.11%	12.73%

Table 2(a) shows that under the 95% confidence level, ASC-D uses 35.8–57.6% fewer than the confidence-weighted method, and 71.3–80.1% fewer than CITE-4, with larger savings at smaller gaps. For the harder setting $\Delta = 0.1$ with a maximum trajectory budget of $N = 512$, Table 2(b) shows that ASC-D correctly certifies the mode in 94.23–98.44% of trials across the tested confidence levels, substantially more often than the reduced-feedback methods. Note that the correct-certification rate decreases as the confidence requirement $1 - \delta$ becomes stricter, since the fixed budget constraint leaves less time to accumulate the required evidence. Together, these results establish ASC-D as the most sample-efficient among the evaluated methods. Appendix D further validates ASC-D's characteristic-time scaling and the concentration-dependent gain in Proposition 2.2.

4.2 Real-World Experiments

Exp. 1: Required number of trajectory generations across modal gaps. For each of 120 MMLU-Redux questions, we use 256 Qwen3-1.7B trajectories and rescale their stored answer logits with $T_a \in \{4, 8, 16, 32\}$, where larger T_a flattens the vectors and narrows the modal gap. At each temperature, the rescaled vectors define the empirical population $\widehat{F} = 256^{-1} \sum_{i=1}^{256} \delta_{\boldsymbol{\theta}_i}$, with mean $\widehat{\boldsymbol{\pi}} = 256^{-1} \sum_{i=1}^{256} \boldsymbol{\theta}_i$, mode index $a^\star = \arg\max_k \widehat{\pi}_k$, and modal gap $\Delta = \widehat{\pi}_{a^\star} - \max_{j \neq a^\star} \widehat{\pi}_j$. We then run 500 paired replays by sampling uniformly from $\widehat{F}$. ASC-D observes $\boldsymbol{\theta}_t$; its matched one-hot ablation observes $\boldsymbol{e}_{Y_t}$ with $Y_t \mid \boldsymbol{\theta}_t \sim \mathrm{Cat}(\boldsymbol{\theta}_t)$; CITE-4 and PPR-1v1 (Jain et al., 2022) observe the same Y_t. The number of trajectory generation (Num. Gen.) is the stopping time for each method under confidence level 95%, and answer accuracy (Ans. Acc.) is evaluated against the correct answer. We observe benefits of our ASC-D over other benchmarks among all the 120 questions. Specifically, we focus on the 90 questions that retain the same ordered top-two answers across temperatures and for which every method attains the 95% success target within the maximum budget of 32,768 trajectories. Table 3 reports the corresponding results on these 90 questions.

Table 3: Number of trajectory generations required for 95% confidence and the answer accuracy on the 90 common questions. Here, Δ_{geo} denotes the geometric mean of the modal gap.

		ASC-D		ASC-D (one-hot)		PPR-1v1		CITE-4	
T_a	Δ_{geo}	Num. Gen.	Ans. Acc.	Num. Gen.	Ans. Acc.	Num. Gen.	Ans. Acc.	Num. Gen.	Ans. Acc.
4	0.61	**24.67**	**72.51%**	41.82	71.83%	46.05	71.65%	58.19	71.70%
8	0.42	**27.82**	**72.40%**	90.73	70.97%	99.21	70.92%	113.92	70.94%
16	0.21	**42.17**	**72.36%**	346.55	70.78%	380.76	70.77%	403.87	70.78%
32	0.10	**76.87**	**72.16%**	1593.49	70.73%	1732.97	70.74%	1711.74	70.74%

The results yield two main findings. First, as T_a increases from 4 to 32, the geometric-mean gap decreases from 0.610 to 0.098, and ASC-D's advantage grows. It uses 41.0–95.2% fewer trajectories than its matched one-hot ablation, 46.4–95.6% fewer than PPR-1v1, and 57.6–95.5% fewer than CITE-4. Because full-vector and one-hot ASC-D use the same stopping rule and differ only in the observed feedback, their comparison isolates the benefit of distribution-valued feedback. These savings come without a loss in answer accuracy.

Second, regressing each question's log required budget on $\log(1/\Delta)$ gives a median slope of 0.766 for ASC-D, versus 1.969, 1.966, and 1.854 for its one-hot ablation, PPR-1v1, and CITE-4, respectively. This finite-range comparison supports the predicted near-linear rather than quadratic dependence on the inverse gap, and that the ASC-D slope below one is not an asymptotic claim. Appendix E reports bootstrap confidence intervals and the corresponding log–log plot.

Exp. 2: Correct certification within a fixed budget. At the natural answer temperature $T_a = 1$, we fix the maximum budget at $N = 128$ and evaluate the same 90 MMLU-Redux questions used in Exp. 1 across Qwen3-1.7B, Llama-3.2-1B, and Llama-3.1-8B. For each model and question, the first 128 probability vectors define $\widehat{F}$, $\widehat{\pi}$, and $a^\star$ as in Exp. 1. We run 200 paired replays at $1-\delta \in \{0.95, 0.975, 0.99\}$. At each step, we sample $\boldsymbol{\theta}_t$ uniformly from $\widehat{F}$ and draw $Y_t \mid \boldsymbol{\theta}_t \sim \mathrm{Cat}(\boldsymbol{\theta}_t)$. ASC-D observes $\boldsymbol{\theta}_t$, while CITE-4 and PPR-1v1 observe the matched draw Y_t. A replay succeeds only if the method stops within the fixed maximum budget and returns $a^\star$.

Table 4: Correct-certification rate within $N = 128$ trajectories on the 90 common questions.

Model	Method	$1-\delta = 0.95$	$1-\delta = 0.975$	$1-\delta = 0.99$
Llama-3.2-1B	ASC-D	**56.57%**	**53.38%**	**49.83%**
	CITE-4	37.51%	35.21%	32.93%
	PPR-1v1	36.11%	34.14%	31.94%
Llama-3.1-8B	ASC-D	**80.62%**	**78.84%**	**76.44%**
	CITE-4	71.69%	69.48%	66.94%
	PPR-1v1	70.34%	68.34%	66.01%
Qwen3-1.7B	ASC-D	**90.75%**	**89.83%**	**88.77%**
	CITE-4	89.29%	88.39%	87.44%
	PPR-1v1	88.69%	87.97%	87.03%

Table 4 shows that ASC-D achieves the highest correct-certification rate for every model and confidence level. Its advantage is largest on Llama-3.2-1B, where it exceeds CITE-4 and PPR-1v1 by up to 19.06 percentage points and 20.46 percentage points, respectively, highlighting the value of distribution-valued feedback.

5 Conclusion and Limitations

In this paper, we show that observing the LLM's probabilities for all candidate answers on each trajectory can reduce the number of trajectories needed to identify the LLM's most consistent answer. Our analysis characterizes the optimal asymptotic rate and proposes ASC-D, an algorithm that attains this rate with a suitable betting grid.

The limitations of this work are as follows: (i) ASC-D identifies the LLM's most consistent answer, which need not be the correct answer, and therefore inherits self-consistency's failure mode when the model is confidently wrong; (ii) the method requires log-probabilities over a finite candidate set, covering multiple-choice and short-answer tasks with specified candidates but not directly open-ended generation with an unbounded answer space; and (iii) our experiments are limited to small open-source models and multiple-choice benchmarks, with evaluation on larger models and open-ended reasoning tasks left for future work.

AI USE STATEMENT

In this work, we used generative AI tools for aiding or polishing writing and for retrieval and discovery (e.g., finding related work). We have not used generative AI tools for research ideation or execution, draft sections of the paper, generating synthetic datasets, and proving mathematical claims. Additionally, we used generative AI tools to check the correctness of the results, help implement some of the experiments, and evaluate the manuscript. We further manually revised the manuscript based on some of the suggestions to further improve its quality. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by the authors. 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 assumptions, problem formulation, and ASC-D procedure are specified in Sections 2 and 3, with complete proofs provided in Appendix B. Section 4 and Appendices C, D, and E provide the experimental protocols, model and dataset settings, temperatures, trajectory budgets, confidence levels, replay procedures, and additional analyses. The code, stored probability vectors and trajectories, experiment configurations, random seeds, and scripts for reproducing the reported tables and figures are available in our anonymous repository at https://anonymous.4open.science/r/ICLR2027-ASCD-7B09/. The repository README maps each reported experiment to the corresponding data, command, and expected output. The released stored trajectories allow the replay-based comparisons to be reproduced without regenerating model outputs.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Yiming Yang, et al. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12375–12396, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Paula Cordero-Encinar and Andrew B Duncan. Certified self-consistency: Statistical guarantees and test-time training for reliable reasoning in llms. *arXiv preprint arXiv:2510.17472*, 2025.

Yichao Fu, Xuewei Wang, Hao Zhang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. In *International Conference on Learning Representations*, volume 2026, pp. 94355–94377, 2026.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pp. 998–1027. PMLR, 2016.

Junya Honda and Akimichi Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *The Journal of Machine Learning Research*, 16(1):3721–3756, 2015.

Baihe Huang. Large language models as statistical decision-makers at inference time. 2026.

Baihe Huang, Shanda Li, Tianhao Wu, Yiming Yang, Ameet Talwalkar, Kannan Ramchandran, Michael Jordan, and Jiantao Jiao. Sample complexity and representation ability of test-time scaling paradigms. In *International Conference on Learning Representations*, volume 2026, pp. 54262–54298, 2026a.

Jingkai Huang, Will Ma, and Zhengyuan Zhou. Optimal bayesian stopping for efficient inference of consistent llm answers. *arXiv preprint arXiv:2602.05395*, 2026b.

Shubham Anand Jain, Rohan Shah, Sanit Gupta, Denil Mehta, Inderjeet J Nair, Jian Vora, Sushil Khyalia, Sourav Das, Vinay J Ribeiro, and Shivaram Kalyanakrishnan. Pac mode estimation using ppr martingale confidence sequences. In *International Conference on Artificial Intelligence and Statistics*, pp. 5815–5852. PMLR, 2022.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *International Conference on Learning Representations*, volume 2024, pp. 14751–14768, 2024.

Hirofumi Ota, Naoto Iwase, Yuki Ichihara, Junpei Komiyama, and Masaaki Imaizumi. CITE: Anytime-valid statistical inference in LLM self-consistency. *arXiv preprint arXiv:2605.05873*, 2026.

Dhruti Shah, Tuhinangshu Choudhury, Nikhil Karamchandani, and Aditya Gopalan. Sequential mode estimation with oracle queries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5644–5651, 2020.

David Siegmund. *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media, 2013.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111, 2025.

Abraham Wald. Sequential tests of statistical hypotheses. In *Breakthroughs in statistics: Foundations and basic theory*, pp. 256–298. Springer, 1992.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. Reasoning aware self-consistency: Leveraging reasoning paths for efficient llm sampling. 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)*, pp. 3613–3635, 2025.

Xinglin Wang, Shaoxiong Feng, Yiwei Li, Peiwen Yuan, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. Make every penny count: Difficulty-adaptive self-consistency for cost-efficient reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6919–6932, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

A Related Literature

Efficient and Adaptive Self-Consistency Adaptive self-consistency reduces the cost of fixed-budget self-consistency by stopping when the sampled answers show sufficient agreement (Aggarwal et al., 2023). Subsequent methods stop based on answer stability (Li et al., 2024), allocate sampling budgets using estimated question difficulty (Wang et al., 2025), or evaluate both answers and reasoning paths (Wan et al., 2025). Other approaches use scalar confidence signals to weight or filter reasoning trajectories (Taubenfeld et al., 2025; Fu et al., 2026). From a theoretical perspective, Huang et al. (2026a) and Huang (2026) compare the sample complexity of self-consistency and verifier-based best-of-N, showing a separation between their quadratic and linear dependence on the answer gap. Most closely related to our setting, Huang et al. (2026b) studies Bayesian stopping using prior information about answer frequencies, but observes only the final-answer labels. In contrast, our method does not require such prior information and instead gains efficiency by observing the full probability vector associated with each trajectory.

Sequential Mode Identification and Anytime-Valid Inference. Our problem is also related to classical sequential hypothesis testing, which studies how to make a reliable decision using as few observations as possible (Wald, 1992; Siegmund, 2013). Sequential mode identification specializes this problem to identifying the most likely outcome of an unknown discrete distribution. Shah et al. (2020) study mode identification under different oracle-query models, while Jain et al. (2022) develops an asymptotically optimal stopping rule using sampled answer labels and martingale confidence sequences. Recent work applies anytime-valid inference to LLM self-consistency: MMC certifies an absolute majority (Cordero-Encinar & Duncan, 2025), whereas CITE certifies a candidate as the unique mode and also considers scalar confidence (Ota et al., 2026). Our work extends this literature from categorical or scalar-weighted observations to probability-vector observations and characterizes the resulting optimal stopping rate.

B Proof of Main Results

Characteristic Time for the Black-box Observation Model The reciprocal of the black-box constant is the KL divergence from $\mathrm{Cat}(\pi)$ to the merge alternative,

$$\frac{1}{T_{\mathrm{bb}}(F)} = p_1 \log \frac{p_1}{\bar{p}} + p_2 \log \frac{p_2}{\bar{p}} = g\Big(\bar{p} + \frac{\Delta}{2}\Big) + g\Big(\bar{p} - \frac{\Delta}{2}\Big), \qquad g(x) := x \log \frac{x}{\bar{p}}.$$

Expanding g around $\bar{p}$, we have $g(\bar{p}) = 0$, $g'(\bar{p}) = 1$ and $g''(\bar{p}) = 1/\bar{p}$. The first-order terms of the two summands are $+\frac{\Delta}{2}$ and $-\frac{\Delta}{2}$ and cancel, so only the second-order terms remain:

$$\frac{1}{T_{\mathrm{bb}}(F)} = 2 \cdot \frac{1}{2}\, g''(\bar{p}) \Big(\frac{\Delta}{2}\Big)^2 + O(\Delta^4) = \frac{\Delta^2}{4\bar{p}} + O(\Delta^4), \qquad \text{hence} \qquad T_{\mathrm{bb}}(F) = \frac{4\bar{p}}{\Delta^2} \cdot \big(1 + O(\Delta^2)\big).$$

The cancellation is the whole reason for the quadratic rate: the merge moves p_1 and p_2 by the same amount in opposite directions, so the divergence has no linear part in Δ.

Proof of Theorem 2.1. For distribution F with unique mode a_1, we define the alternative set $\mathrm{Alt}(F)$ as

$$\mathrm{Alt}(F) := \{F' \in \mathcal{M} : \arg\max_k \mathbb{E}_{F'}[\theta_k] \neq 1\} = \bigcup_{j=2}^{K} \{F' : \mathbb{E}_{F'}[X_j] \leq 0\},$$

i.e., $\mathrm{Alt}(F)$ stands for the set of distributions in $\mathcal{M}$ for which a_1 is not the unique mode. We let $\mathrm{Alt}_j(F) := \{F' : \mathbb{E}_{F'}[X_j] \leq 0\}$ for all $j \neq 1$. Thus $\mathrm{Alt}(F) = \cup_{j=2}^{K} \mathrm{Alt}_j(F)$.

First, take an alternative $F' \in \mathrm{Alt}(F)$, and let E be the event that the procedure recommends a_1 as the unique mode. We have the procedure is δ-PAC at both F and the alternative F', i.e.,

$$\mathbb{P}_F(E) \geq 1 - \delta, \qquad \mathbb{P}_{F'}(E) \leq \delta.$$

Then, by applying Lemma 1 of Kaufmann et al. (2016), we can conclude the lower bound of:

$$\mathbb{E}_F[\tau] \cdot \mathrm{D}_{\mathrm{KL}}(F \| F') \geq \mathrm{D}_{\mathrm{KL}}(\mathrm{Bern}(\mathbb{P}_F(E) \| \mathbb{P}_{F'}(E))) \geq (1 - 2\delta) \cdot \log \frac{1 - \delta}{\delta}.$$

Therefore, taking the supremum over the alternative set, we obtain

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_F[\tau]}{\log(1/\delta)} \geq \frac{1}{\inf_{F' \in \mathrm{Alt}(F)} \mathrm{D}_{\mathrm{KL}}(F \| F')}. \tag{13}$$

Next, it suffices to calculate the denominator in (13). We have

$$\inf_{F' \in \mathrm{Alt}(F)} \mathrm{D}_{\mathrm{KL}}(F \| F') = \min_{j \neq 1} \inf_{F' : \mathbb{E}_{F'}[X_j] \leq 0} \mathrm{D}_{\mathrm{KL}}(F \| F') \tag{14}$$

according to the definition of $\mathrm{Alt}(F)$. For a distribution G on $[-1, 1]$, we define

$$\mathcal{K}_{\inf}(G) := \inf \left\{ \mathrm{D}_{\mathrm{KL}}(G \| G') : \operatorname{supp} G' \subseteq [-1, 1],\ \mathbb{E}_{G'}[X] \leq 0 \right\}, \tag{15}$$

which represents the minimal KL divergence from the distribution G to the set of candidate distributions G' supported on $[-1, 1]$ with a mean less than or equal to 0. According to Proposition 1 of Honda & Takemura (2015), we have for any distribution G on $[-1, 1]$ with positive mean,

$$\mathcal{K}_{\inf}(G) = \max_{\lambda \in [0,1]} \mathbb{E}_G[\log(1 + \lambda X)].$$

Fix $j \neq 1$, and let G_j be the law of X_j under F. For every $F' \in \mathrm{Alt}_j(F)$, the law G'_j of X_j under F' has nonpositive mean. Data processing inequality therefore implies

$$\mathrm{D}_{\mathrm{KL}}(F \| F') \geq \mathrm{D}_{\mathrm{KL}}(G_j \| G'_j) \geq \mathcal{K}_{\inf}(G_j).$$

To obtain the reverse inequality, we construct an alternative F' that differs from F only in the distribution of X_j. Sampling from F can be viewed as two steps: first draw $X_j \sim G_j$, and then draw θ from its conditional distribution under F given X_j. Now we take any candidate law G' in

(15). Replace the first step by drawing $X_j \sim G'$, while leaving the second step unchanged. Let F' denote the resulting distribution of θ.

Under F', the margin X_j has exactly the law G', whose mean is nonpositive. Thus $F' \in \mathrm{Alt}_j(F)$. Moreover, the conditional distribution used in the second step is the same under F and F', so it contributes no additional KL divergence. The chain rule for relative entropy therefore gives

$$\mathrm{D}_{\mathrm{KL}}(F\|F') = \mathrm{D}_{\mathrm{KL}}(G_j\|G').$$

Hence every candidate G' in the one-dimensional problem gives a candidate F' in the original problem with the same KL divergence. Taking the infimum over G' proves the reverse inequality:

$$\inf_{F'\in\mathrm{Alt}_j(F)} \mathrm{D}_{\mathrm{KL}}(F\|F') \le \inf_{G':\mathbb{E}_{G'}[X]\le 0} \mathrm{D}_{\mathrm{KL}}(G_j\|G') = \mathcal{K}_{\inf}(G_j).$$

Combining these two directions of the inequalities yields the equation of

$$\inf_{F'\in\mathrm{Alt}_j(F)} \mathrm{D}_{\mathrm{KL}}(F\|F') = \mathcal{K}_{\inf}(G_j) = \max_{\lambda\in[0,1]} \mathbb{E}_F[\log(1+\lambda X_j)] = \mathcal{G}_j(F).$$

Taking the result back to (13) and (14), we obtain the main statement of the theorem.

We then focus on the inequalities of

$$\frac{1}{T_{\mathrm{bb}}(F)} \le \min_{j\neq 1} \mathcal{G}_j(F) \le \log(1+\Delta). \tag{16}$$

For the RHS of (16): for any F and $\lambda \in [0,1]$, since $\log x$ is concave in x, by applying the Jensen's inequality,

$$\mathbb{E}_F[\log(1+\lambda X_j)] \le \log(1+\lambda\mathbb{E}_F[X_j]) = \log(1+\lambda\cdot(p_1-p_j)) \le \log(1+p_1-p_j).$$

Let $j=2$ and we obtain $\min_{j\neq 1}\mathcal{G}_j(F) \le \log(1+\Delta)$. For $F=\delta_{\boldsymbol{\pi}}$, every margin is the positive constant p_1-p_j, therefore $\mathcal{G}_j(\delta_{\boldsymbol{\pi}}) = \log(1+p_1-p_j)$ and thus $F=\delta_{\boldsymbol{\pi}}$ can achieve the upper bound.

For the LHS of (16): concavity of the logarithm and $\sum_k \theta_k = 1$ give, for $0 \le \lambda < 1$,

$$\begin{aligned}\log(1+\lambda(\theta_1-\theta_j)) &= \log\left(\theta_1(1+\lambda)+\theta_j(1-\lambda)+\sum_{k\notin\{1,j\}}\theta_k\right)\\ &\ge \theta_1\log(1+\lambda)+\theta_j\log(1-\lambda).\end{aligned}$$

Taking expectations and optimizing over λ yields

$$\begin{aligned}\mathcal{G}_j(F) &\ge \sup_{0\le\lambda<1}\{p_1\log(1+\lambda)+p_j\log(1-\lambda)\}\\ &= D_j := p_1\log\frac{2p_1}{p_1+p_j}+p_j\log\frac{2p_j}{p_1+p_j},\end{aligned} \tag{17}$$

with the maximizer $\lambda = (p_1-p_j)/(p_1+p_j)$. Because $p_j \le p_2$ and $\log(1-\lambda)\le 0$, the supremum in (17) is at least its value with p_j replaced by p_2. Thus $D_j \ge D_2$ for every $j\neq 1$, and

$$\min_{j\neq 1}\mathcal{G}_j(F) \ge D_2 = \frac{1}{T_{\mathrm{bb}}(F)}.$$

Finally, under the vertex law $F_0 = \sum_k p_k\delta_{\boldsymbol{e}_k}$, the margin X_j equals 1, -1, and 0 with probabilities p_1, p_j, and $1-p_1-p_j$, respectively. Hence $\mathcal{G}_j(F_0) = D_j$, and thus the vertex law F_0 can achieve the lower bound.

□

Proof of Proposition 2.1. Let $F\in\mathcal{M}$ have unique mode a_1, and recall that Q_F is the distribution of the observed pair (o,w). We use the same alternative sets as in the proof of Theorem 2.1:

$$\mathrm{Alt}_j(F) := \{F'\in\mathcal{M} : \mathbb{E}_{F'}[X_j]\le 0\}, \qquad \mathrm{Alt}(F) = \bigcup_{j\neq 1}\mathrm{Alt}_j(F),$$

where $X_j = \theta_1 - \theta_j$. Thus $\mathrm{Alt}(F)$ contains the distributions for which a_1 is not the unique mode.

First take an alternative $F' \in \mathrm{Alt}(F)$ with a unique mode different from a_1, and let E be the event that the procedure recommends a_1. Since the procedure is δ-PAC under both F and F',

$$\mathbb{P}_F(E) \geq 1 - \delta, \qquad \mathbb{P}_{F'}(E) \leq \delta.$$

The same change-of-measure argument used in the proof of Lemma 1 Kaufmann et al. (2016), now applied to observations with laws Q_F and $Q_{F'}$, gives,

$$\mathbb{E}_F[\tau] \cdot \mathrm{D_{KL}}(Q_F \| Q_{F'}) \geq \mathrm{D_{KL}}(\mathrm{Bern}(\mathbb{P}_F(E) \| \mathbb{P}_{F'}(E))) \geq (1 - 2\delta) \cdot \log \frac{1-\delta}{\delta}.$$

Similarly, taking the supremum over the alternative set, and letting $\delta \to 0$, we obtain

$$\begin{aligned}\liminf_{\delta \to 0} \frac{\mathbb{E}_F[\tau]}{\log(1/\delta)} &\geq \frac{1}{\inf_{F' \in \mathrm{Alt}(F)} \mathrm{D_{KL}}(Q_F \| Q_{F'})} \\ &= \left[\min_{j \neq 1} \inf_{F' \in \mathrm{Alt}_j(F)} \mathrm{D_{KL}}(Q_F \| Q_{F'})\right]^{-1} = T_{\mathrm{cw}}(F).\end{aligned}$$

Finally, for the sandwich bound $T_{\mathrm{gb}}(F) \leq T_{\mathrm{cw}}(F) \leq T_{\mathrm{bb}}(F)$: the pair (o, w) can be generated from $\boldsymbol{\theta}$, and the sampled answer o is obtained by dropping w. The data processing inequality therefore gives, for every $F' \in \mathrm{Alt}(F)$,

$$\mathrm{D_{KL}}(F \| F') \geq \mathrm{D_{KL}}(Q_F \| Q_{F'}) \geq \mathrm{D_{KL}}\left(\mathrm{Cat}(\boldsymbol{\pi}) \| \mathrm{Cat}(\mathbb{E}_{F'}[\boldsymbol{\theta}])\right).$$

Taking infima over the same alternative set, the first term gives $1/T_{\mathrm{gb}}(F)$ by Theorem 2.1. The last term gives $1/T_{\mathrm{bb}}(F)$ in (2). Thus we obtain the desired result. Besides, we also note that this asymptotic lower bound is achievable through some properly designed algorithm (with similar idea to that of our ASC-D). We omit the details as this is not the main focus of the paper. □

Proof of Proposition 2.2. Throughout, $F = \mathrm{Dir}(c\pi)$ and $\phi_j(\lambda) := \mathbb{E}_F[\log(1 + \lambda X_j)]$, so that $\mathcal{G}_j(F) = \max_{\lambda \in [0,1]} \phi_j(\lambda)$.

Second moment of X_2. By (6),

$$\mathrm{Var}(X_2) = \mathrm{Var}(\theta_1) + \mathrm{Var}(\theta_2) - 2\mathrm{Cov}(\theta_1, \theta_2) = \frac{p_1(1-p_1) + p_2(1-p_2) + 2p_1p_2}{c+1} = \frac{p_1 + p_2 - \Delta^2}{c+1},$$

and since $\mathbb{E}[X_2] = \Delta$,

$$\mathbb{E}[X_2^2] = \frac{2\bar{p} - \Delta^2}{c+1} + \Delta^2 = \frac{2\bar{p}}{c+1} \cdot \big(1 + O(\Delta^2)\big). \tag{18}$$

Expansion of $\mathcal{G}_2$. The function ϕ_2 is concave on $[0,1]$, with $\phi_2'(\lambda) = \mathbb{E}[X_2/(1+\lambda X_2)]$ for $0 \leq \lambda < 1$. Fix $\lambda_0 \in (0,1)$. Since $1 + \lambda_0 X_2 \leq 2$,

$$\phi_2'(\lambda_0) = \mathbb{E}\Big[\frac{X_2}{1+\lambda_0 X_2}\Big] = \mathbb{E}\Big[X_2 - \frac{\lambda_0 X_2^2}{1+\lambda_0 X_2}\Big] \leq \Delta - \frac{\lambda_0}{2}\,\mathbb{E}[X_2^2],$$

which is negative for small Δ by (18). Since $\phi_2'(0) = \Delta > 0$, the maximizer $\lambda^\star$ lies in $(0, \lambda_0)$, and we have:

$$\Delta = \lambda^\star \mathbb{E}\left[\frac{X_2^2}{1+\lambda^\star X_2}\right] \geq \frac{\lambda^\star}{2}\mathbb{E}[X_2^2],$$

so $\lambda^\star \leq 2\Delta/\mathbb{E}[X_2^2] = O(\Delta)$.

On this interval, $\log(1+u) = u - \frac{u^2}{2} + R(u)$ with $|R(u)| \leq C_0|u|^3$ for some constant $C_0 > 0$ for any $|u| \leq \lambda_0$, and $|X_2|^3 \leq X_2^2$ since $|X_2| \leq 1$, so with $u = \lambda X_2$,

$$\phi_2(\lambda) = \lambda\Delta - \frac{\lambda^2}{2}\,\mathbb{E}[X_2^2] + r(\lambda), \qquad |r(\lambda)| \leq C_0\lambda^3\,\mathbb{E}[X_2^2]. \tag{19}$$

The quadratic part is maximized at $\hat{\lambda} := \Delta/\mathbb{E}[X_2^2] = \frac{(c+1)\Delta}{2\bar{p}} \cdot (1 + O(\Delta^2))$, which lies in $[0, \lambda_0]$ for small Δ, with value $\frac{\Delta^2}{2\mathbb{E}[X_2^2]} = \frac{(c+1)\Delta^2}{4\bar{p}} \cdot (1 + O(\Delta^2))$. Evaluating (19) at $\hat{\lambda}$ gives $\mathcal{G}_2(F) \geq$

$\frac{(c+1)\Delta^2}{4\bar{p}}(1+O(\Delta^2)) - C_0\hat{\lambda}^3\mathbb{E}[X_2^2]$, and the remainder is $O(\Delta^3)$ since $\hat{\lambda} = O(\Delta)$. Conversely, the quadratic part at $\lambda^\star$ is at most its value at $\hat{\lambda}$, and $r(\lambda^\star) = O(\Delta^3)$ because $\lambda^\star = O(\Delta)$. Thus $\mathcal{G}_2(F) \le \frac{\Delta^2}{2\mathbb{E}[X_2^2]} + O(\Delta^3)$. Therefore

$$\mathcal{G}_2(F) = \frac{(c+1)\Delta^2}{4\bar{p}} \cdot \big(1 + O(\Delta)\big). \tag{20}$$

For the maximizer, $\phi_2'(\lambda) = \mathbb{E}[X_2] - \lambda\mathbb{E}[X_2^2] + O(\lambda^2\mathbb{E}[X_2^2])$ for $\lambda \le \lambda_0$, so the stationary condition $\phi_2'(\lambda^\star) = 0$ gives $\lambda^\star = \frac{\Delta}{\mathbb{E}[X_2^2]} \cdot (1 + O(\lambda^\star)) = \frac{(c+1)\Delta}{2\bar{p}} \cdot (1 + O(\Delta))$.

Along this limit, $p_1 = \bar{p} + \Delta/2$ and $p_2 = \bar{p} - \Delta/2$. For $K > 2$, the fixed probabilities satisfy $p_j < \bar{p}$ for every $j \ge 3$, since $p_j \le p_2$. The pairwise lower bound $\mathcal{G}_j(F) \ge D_j$ in (17) then gives $\liminf_{\Delta\to 0} \mathcal{G}_j(F) > 0$ for each $j \ge 3$, whereas (20) gives $\mathcal{G}_2(F) \to 0$. Thus $\min_{j\neq 1} \mathcal{G}_j(F) = \mathcal{G}_2(F)$ for sufficiently small Δ. For $K = 2$ this result is immediate. Combining (20) with the black-box expansion in (2), we thus obtain now

$$\begin{aligned} T_{\mathrm{gb}}(F) &= \frac{1}{\mathcal{G}_2(F)} = \frac{4\bar{p}}{(c+1)\Delta^2}\big(1 + O(\Delta)\big) \\ &= \frac{T_{\mathrm{bb}}(F)}{c+1}\big(1 + O(\Delta)\big). \end{aligned}$$

□

Insights on the Grid Design The averaged wealth (9) grows at the best rate on the grid. For each challenger $j \neq 1$, let $\lambda_j^\star \in (0, 1]$ be a maximizer in (3), and write

$$\phi_j(\lambda) := \mathbb{E}_F[\log(1 + \lambda(\theta_1 - \theta_j))], \qquad \mathcal{G}_j(F) = \phi_j(\lambda_j^\star).$$

The optimal fraction depends on the unknown law F, so the grid must cover a range of possible values. The two insights are:

(i) **The middle can be coarse**. The function ϕ_j is concave and satisfies $\phi_j(0) = 0$. Hence

$$\phi_j(\lambda) \ge \frac{\lambda}{\lambda_j^\star}\mathcal{G}_j(F), \qquad 0 \le \lambda \le \lambda_j^\star.$$

Thus a point in $[\lambda_j^\star/r, \lambda_j^\star]$, for $r > 1$, retains at least a fraction $1/r$ of the optimal growth rate. The width of this interval is $\lambda_j^\star(1-1/r)$, so relatively coarse spacing suffices when the optimum lies in the middle of $[0, 1]$.

(ii) **Both ends should be dense**. When $\lambda_j^\star$ is small, the interval $[\lambda_j^\star/r, \lambda_j^\star]$ is narrow. Under Assumption 1, for example, $\lambda_2^\star \approx (c+1)\Delta/(2\bar{p})$ tends to zero with the gap for fixed c and $\bar{p}$. A uniform grid with spacing h cannot provide a positive point below the optimum once $\lambda_j^\star < h$. Geometrically spaced points r^{-i} instead cover progressively smaller fractions. At the other endpoint, highly consistent trajectories can favor fractions close to 1. We therefore also include points $1 - r^{-i}$, which approach 1 for large i.

Proof of Theorem 3.1. We write $\mathcal{F}_n := \sigma(\boldsymbol{\theta}_i : i \le n)$ and $M_i^{(k,j)} := \theta_{i,k} - \theta_{i,j} \in [-1, 1]$, so that $W_n^{(k,j)}(\lambda) = \prod_{i\le n}(1 + \lambda M_i^{(k,j)})$ and $\mathbb{E}_F[M_i^{(k,j)}] = p_k - p_j$. Throughout, $\phi_j(\lambda) := \mathbb{E}_F[\log(1 + \lambda M_1^{(1,j)})]$, so that $\mathcal{G}_j^\Lambda(F) = \max_{\lambda\in\Lambda} \phi_j(\lambda)$ in (11), and for a finite grid $\Lambda \subset [0, 1)$ we further define

$$g_\Lambda := \min_{j\neq 1} \mathcal{G}_j^\Lambda(F) = \frac{1}{T_{\mathrm{gb}}^\Lambda(F)}, \qquad B_\Lambda := \max_{\lambda\in\Lambda} \log\frac{1}{1-\lambda}.$$

Fix $k \neq 1$ and $\lambda \in \Lambda$. Since $p_k - p_1 < 0$,

$$\mathbb{E}_F\big[W_n^{(k,1)}(\lambda) \mid \mathcal{F}_{n-1}\big] = W_{n-1}^{(k,1)}(\lambda) \cdot \big(1 + \lambda \cdot (p_k - p_1)\big) \le W_{n-1}^{(k,1)}(\lambda),$$

so $(W_n^{(k,1)}(\lambda))_{n\ge 0}$ is a nonnegative supermartingale with $W_0^{(k,1)}(\lambda) = 1$, and so is its average $\overline{W}_n^{(k,1)}$ over $\lambda \in \Lambda$ defined in (9). Ville's inequality for the supermartingale gives

$$\mathbb{P}_F\Big(\sup_{n\ge 1} \overline{W}_n^{(k,1)} \ge \frac{K-1}{\delta}\Big) \le \frac{\delta}{K-1}.$$

On the event $\{\tau^{\mathrm{D}} < \infty, \hat{a}_{\tau^{\mathrm{D}}} = a_k\}$ the stopping condition of Algorithm 1 holds at time τ^{D} with leader $\hat{k}_{\tau^{\mathrm{D}}} = k$, and in particular $\overline{W}^{(k,1)}_{\tau^{\mathrm{D}}} \geq (K-1)/\delta$. Hence

$$\mathbb{P}_F(\tau^{\mathrm{D}} < \infty,\ \hat{a}_{\tau^{\mathrm{D}}} \neq a_1) \leq \sum_{k=2}^{K} \mathbb{P}_F\Big(\sup_{n\geq 1} \overline{W}^{(k,1)}_n \geq \frac{K-1}{\delta}\Big) \leq \delta.$$

Part (i). Let $\Lambda = \Lambda_{r,m}$ with $r^{-m} \leq \min_{j\neq 1} \lambda_j^\star$. Fix $j \neq 1$, ϕ_j is concave on $[0,1]$ with $\phi_j(0) = 0$ and $\phi_j(\lambda_j^\star) = \mathcal{G}_j(F) > 0$, so $\lambda_j^\star > 0$. The grid contains a point $\lambda \in [\lambda_j^\star/r, \lambda_j^\star]$: if $\lambda_j^\star < r^{-1}$, then $r^{-i} \leq \lambda_j^\star < r^{-i+1}$ for some $2 \leq i \leq m$ because $r^{-m} \leq \lambda_j^\star$, and $\lambda = r^{-i}$ falls within the bandwidth. If $\lambda_j^\star \geq r^{-1}$, then $\lambda = r^{-1}$ falls within the bandwidth because $\lambda_j^\star \leq 1$. For such λ, concavity and $\phi_j(0) = 0$ give

$$\phi_j(\lambda) = \phi_j\Big(\frac{\lambda}{\lambda_j^\star} \cdot \lambda_j^\star + \Big(1 - \frac{\lambda}{\lambda_j^\star}\Big) \cdot 0\Big) \geq \frac{\lambda}{\lambda_j^\star} \cdot \mathcal{G}_j(F) \geq \frac{\mathcal{G}_j(F)}{r}.$$

Hence $\mathcal{G}_j^\Lambda(F) \geq \mathcal{G}_j(F)/r$ for every $j \neq 1$, i.e.,

$$g_{\Lambda_{r,m}} \geq \frac{1}{r \cdot T_{\mathrm{gb}}(F)} > 0 \qquad \text{and} \qquad T_{\mathrm{gb}}(F) \leq T_{\Lambda_{r,m}}(F) \leq r \cdot T_{\mathrm{gb}}(F). \tag{21}$$

Next, we show that every finite grid Λ with $g_\Lambda > 0$ satisfies $\limsup_{\delta\to 0} \mathbb{E}_F[\tau^{\mathrm{D}}]/\log(1/\delta) \leq T_{\mathrm{gb}}^\Lambda(F)$. For each $j \neq 1$ fix $\lambda_j \in \arg\max_{\lambda\in\Lambda} \phi_j(\lambda)$, so that $\phi_j(\lambda_j) = \mathcal{G}_j^\Lambda(F) \geq g_\Lambda$. The variables $Y_{i,j} := \log(1 + \lambda_j M_i^{(1,j)})$ are i.i.d. in i with mean $\phi_j(\lambda_j)$ and, since $\lambda_j < 1$ and $|M_i^{(1,j)}| \leq 1$, bounded by $|Y_{i,j}| \leq B_\Lambda$. Let $L := \log\big((K-1)|\Lambda|/\delta\big)$ and consider the good event

$$E_n := \bigcap_{j\neq 1} \Big\{ \sum_{i=1}^{n} Y_{i,j} \geq L \Big\} \cap \big\{\hat{k}_n = 1\big\}$$

that at time n, the leader is the true modal answer and the cumulative log wealth for each challenger j is greater than L. Condition on the good event E_n, we have for every $j \neq 1$, $\overline{W}_n^{(1,j)} \geq |\Lambda|^{-1} \cdot W_n^{(1,j)}(\lambda_j) = |\Lambda|^{-1} \cdot \exp\big(\sum_{i\leq n} Y_{i,j}\big) \geq (K-1)/\delta$ (since the wealth is nonnegative), so the stopping condition holds at time n with leader a_1 and $\tau^{\mathrm{D}} \leq n$. Therefore $\mathbb{P}_F(\tau^{\mathrm{D}} > n) \leq \mathbb{P}_F(E_n^c)$, and we bound the two parts of E_n^c by Hoeffding's inequality: for i.i.d. $Z_1, \cdots, Z_n \in [a,b]$ with mean μ and $t > 0$, we have

$$\mathbb{P}\Big(\sum_{i=1}^{n} Z_i \leq n \cdot (\mu - t)\Big) \leq \exp\Big(-\frac{2nt^2}{(b-a)^2}\Big).$$

First, $\hat{k}_n \neq 1$ requires $S_{n,1} - S_{n,k} = \sum_{i\leq n} M_i^{(1,k)} \leq 0$ for some $k \neq 1$, and $M_i^{(1,k)} \in [-1,1]$ has mean $p_1 - p_k \geq \Delta$, so

$$\mathbb{P}_F(\hat{k}_n \neq 1) \leq \sum_{k\neq 1} \mathbb{P}_F\Big(\sum_{i\leq n} M_i^{(1,k)} \leq 0\Big) \leq (K-1) \cdot \exp(-n\Delta^2/2).$$

Second, fix $\varepsilon \in (0,1)$ and let $n_0 := \lceil L/((1-\varepsilon)g_\Lambda)\rceil$. For $n \geq n_0$ we have $L \leq (1-\varepsilon) \cdot n g_\Lambda \leq n\phi_j(\lambda_j) - \varepsilon n g_\Lambda$, so $\sum_{i\leq n} Y_{i,j} < L$ implies $\sum_{i\leq n} Y_{i,j} \leq n \cdot (\phi_j(\lambda_j) - \varepsilon g_\Lambda)$, and Hoeffding's inequality with $[a,b] = [-B_\Lambda, B_\Lambda]$ and $t = \varepsilon g_\Lambda$ gives

$$\mathbb{P}_F\Big(\sum_{i\leq n} Y_{i,j} < L\Big) \leq \exp\Big(-\frac{n\varepsilon^2 g_\Lambda^2}{2B_\Lambda^2}\Big), \qquad n \geq n_0.$$

With $\gamma := \min\{\Delta^2/2, \varepsilon^2 g_\Lambda^2/(2B_\Lambda^2)\} > 0$ we obtain $\mathbb{P}_F(\tau^{\mathrm{D}} > n) \leq 2(K-1)\cdot e^{-\gamma n}$ for all $n \geq n_0$. Hence $\tau^{\mathrm{D}} < \infty$ almost surely, and

$$\mathbb{E}_F[\tau^{\mathrm{D}}] = \sum_{n\geq 0} \mathbb{P}_F(\tau^{\mathrm{D}} > n) \leq n_0 + \sum_{n\geq n_0} 2(K-1) \cdot \exp(-\gamma n) \leq \frac{\log\big((K-1)|\Lambda|/\delta\big)}{(1-\varepsilon)\cdot g_\Lambda} + 1 + \frac{2(K-1)}{1-\exp(-\gamma)}. \tag{22}$$

For a fixed grid the last two terms do not depend on δ, so dividing by $\log(1/\delta)$, letting $\delta \to 0$ and then $\varepsilon \to 0$ gives $\limsup_{\delta\to 0} \mathbb{E}_F[\tau^{\mathrm{D}}]/\log(1/\delta) \le T^{\Lambda}_{\mathrm{gb}}(F)$.

Part (ii). We write τ^{D}_δ for the stopping time with the grid $\Lambda_\delta := \Lambda_{r_\delta, m_\delta}$ and $\ell := \log(1/\delta)$ and let δ be small enough so that $\ell \ge 1$. Now according to the definition, we have $r_\delta = 1 + \ell^{-1/4}$ and $m_\delta := \lceil \ell^{1/2} \rceil$. Then $1 < r_\delta \le 2$ and $m_\delta \le 2\ell^{1/2}$. Since $\log(1+x) \ge x/2$ for $x \in [0,1]$,

$$m_\delta \log r_\delta \ge \ell^{1/2} \cdot \frac{1}{2}\ell^{-1/4} = \frac{1}{2}\ell^{1/4} \to \infty,$$

so $r_\delta^{-m_\delta} \to 0$ and the condition $r_\delta^{-m_\delta} \le \min_{j\neq 1} \lambda_j^\star$ of Part (i) holds for all δ small enough. For such δ, (21) gives $g_{\Lambda_\delta} \ge 1/(r_\delta T_{\mathrm{gb}}(F)) \ge 1/(2T_{\mathrm{gb}}(F))$, and the argument of Part (i) gives $\tau^{\mathrm{D}}_\delta < \infty$ almost surely together with (22).

It remains to control how the grid enters (22). First, $|\Lambda_\delta| \le 2m_\delta \le 4\ell^{1/2}$, so $\log((K-1)|\Lambda_\delta|) = O(\log \ell)$. Second, a point $\lambda = 1 - r_\delta^{-i}$ contributes $\log(1/(1-\lambda)) = i \log r_\delta \le m_\delta \log r_\delta \le m_\delta(r_\delta - 1) \le 2\ell^{1/4}$ to B_{Λ_δ}, and a point $\lambda = r_\delta^{-i}$ contributes at most $\log(1/(1-r_\delta^{-1})) = \log(1 + 1/(r_\delta - 1)) = \log(1+\ell^{1/4}) \le \ell^{1/4}$. Hence $B_{\Lambda_\delta} \le 2\ell^{1/4}$ and now we need

$$\gamma \ge \min\left\{ \frac{\Delta^2}{2}, \frac{\varepsilon^2}{32 T_{\mathrm{gb}}(F)^2\, \ell^{1/2}} \right\}.$$

Using $1/(1-\exp(-\gamma)) \le 1 + 1/\gamma$, the last two terms of (22) are $O(\ell^{1/2})$ for fixed ε. Dividing (22) by ℓ and using $1/g_{\Lambda_\delta} \le r_\delta T_{\mathrm{gb}}(F)$ gives:

$$\frac{\mathbb{E}_F[\tau^{\mathrm{D}}_\delta]}{\log(1/\delta)} \le \frac{r_\delta \cdot T_{\mathrm{gb}}(F)}{1-\varepsilon} \cdot \left(1 + \frac{\log((K-1)|\Lambda_\delta|)}{\ell}\right) + O(\ell^{-1/2}) \xrightarrow[\delta\to 0]{} \frac{T_{\mathrm{gb}}(F)}{1-\varepsilon}.$$

since $r_\delta \to 1$. Letting $\varepsilon \to 0$ then proves (12). □

C A REAL-WORLD GREY-BOX OBSERVATION EXAMPLE

We give a complete example from the released demo data to make the grey-box observation model operationally concrete. The record is trajectory 0 (seed 20260922) for ARC-Challenge test item `MCAS_2001_5_3`, generated by `mlx-community/Llama-3.2-3B-Instruct-bf16` at reasoning temperature $T_r = 0.6$ (top-$p = 0.95$, top-$k = 20$). Answer probabilities are evaluated at the natural temperature $T_a = 1$.

Original problem

> **Question:** Which of the following is LEAST responsible for the weathering of rocks?
> **A.** freezing
> **B.** plant growth
> **C.** rain
> **D.** lightning

The benchmark answer is D.

Prompt The model receives the following system and user messages through its chat template.

> **System:** You are a careful reasoner. The final answer must be exactly one of A, B, C, or D.
>
> **User:** Solve this multiple-choice problem carefully.
>
> Question:
> Which of the following is LEAST responsible for the weathering of rocks?
>
> Options:
> A. freezing
> B. plant growth
> C. rain
> D. lightning
>
> Give a concise explanation in at most 120 words. Do not repeat the question or options. During this stage, do not output the final A/B/C/D letter and do not quote the exact chosen option. End exactly with `<END_REASONING>`.

LLM output The sampled reasoning trajectory, before answer scoring, is:

> Plant growth is not a primary cause of weathering, as it primarily contributes to physical weathering through roots breaking apart the rock. Freezing, on the other hand, is a significant factor in chemical weathering, causing ice to expand and exert pressure on the rock, leading to cracking. Rain is a major contributor to physical weathering, as it can physically wear away the rock surface through erosion and abrasion. Lightning, while not directly causing weathering, can contribute to it by igniting wildfires that can further break down the rock.
>
> `END_REASONING`

The observed probability vector After this trajectory, we append the decision prefix

```
Final answer (one letter only):
```

The prefix ends in the token "`):`" (token ID 1680). Each allowed continuation is one token: `A` (362), `B` (426), `C` (356), or `D` (423). Thus the observation comes from one precisely identified next-token position: the transitions from the terminal `):` token to each of the four option tokens. If ℓ_k is the full-vocabulary log-probability of option token k at that position, we record

$$q_k = \exp(\ell_k), \qquad \theta_k = \frac{\exp(\ell_k/T_a)}{\sum_{j\in\{\mathrm{A,B,C,D}\}} \exp(\ell_j/T_a)}.$$

For this trajectory, the exact values are:

Table 5: A single real-world grey-box observation. The raw probabilities q_k are the full-vocabulary next-token probabilities assigned to the four allowed label tokens. Their total mass is 0.9630 and renormalizing over A–D at $T_a = 1$ gives the observed vector $\boldsymbol{\theta} = (0.2635, 0.4344, 0.2052, 0.0969)$.

Option token	Token ID	ℓ_k	q_k	θ_k
`A`	362	-1.3715	0.2537	0.2635
`B`	426	-0.8715	0.4183	0.4344
`C`	356	-1.6215	0.1976	0.2052
`D`	423	-2.3715	0.0933	0.0969
Sum			0.9630	1.0000

This observation is not a product of probabilities along the reasoning trajectory. It is a single next-token softmax, restricted and renormalized over the known candidate labels, after the sampled reasoning has been placed in the context.

Final answer Since $\arg\max_k \theta_k = \mathrm{B}$, greedy decoding returns **B**. The released record also stores an independently seeded paired black-box draw $o \sim \mathrm{Cat}(\boldsymbol{\theta})$, which happened to be A in this trajectory; ASC-D instead observes the entire vector in Table 5. The greedy answer B and the paired draw A are both incorrect relative to the benchmark answer D, illustrating that the grey-box observation exposes uncertainty rather than assuming that a single trajectory is correct.

D Additional Synthetic Experiments

The following experiments test two distinct theoretical predictions. The first validates the characteristic-time scaling for a general, non-Dirichlet feedback law, while the second examines how the gain from distribution-valued feedback changes with Dirichlet concentration.

Characteristic Time Validation To test the characteristic-time prediction beyond the Dirichlet model, we fix $K = 10$ and $\boldsymbol{\pi} = (0.4, 0.3, 0.03751_8)$ and consider the discrete family

$$Z \sim \mathrm{Cat}(\boldsymbol{\pi}), \qquad \boldsymbol{\theta} = (1-\rho)\boldsymbol{e}_Z + \rho\boldsymbol{\pi}, \qquad \rho \in [0,1]. \tag{23}$$

Every F_ρ has mean $\boldsymbol{\pi}$ and therefore induces the same black-box law $\mathrm{Cat}(\boldsymbol{\pi})$. Increasing ρ changes only the information in the probability vector, moving from the vertex law F_0 to the degenerate law $F_1 = \delta_{\boldsymbol{\pi}}$. For $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, we compute $T_{\mathrm{gb}}(F_\rho)$ and run 2,000 repetitions of ASC-D for $\delta \in \{0.2, 0.1, 0.05, 0.02, 0.01, 0.002, 10^{-3}, 2\times10^{-4}, 10^{-5}, 10^{-6}, 10^{-8}\}$. We use $\Lambda_{1.5,16}$ and a maximum budget of 20,000. For each ρ, we estimate the empirical characteristic time by fitting $\mathbb{E}[\tau^{\mathrm{D}}] = \alpha_\rho + \beta_\rho \log(1/\delta)$ over $\delta \le 10^{-3}$. Confidence intervals use 2,000 paired bootstrap samples.

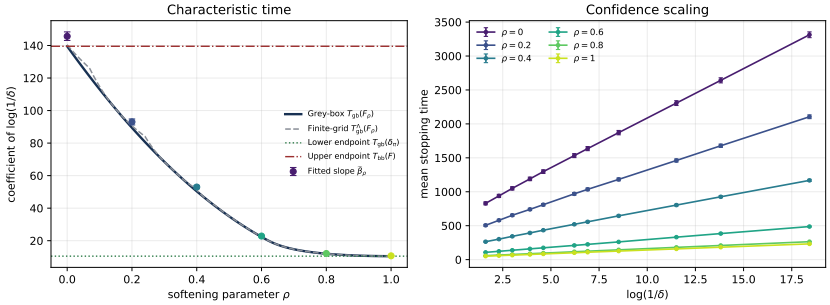

Figure 2: Characteristic-time scaling across the feedback family F_ρ.

Figure 2 compares theoretical and empirical characteristic times (left) and mean stopping time versus $\log(1/\delta)$ (right). Although all F_ρ induce the same black-box law, $T_{\mathrm{gb}}(F_\rho)$ decreases from the vertex upper bound toward the degenerate lower bound as ρ increases. The fitted slopes differ from $T_{\mathrm{gb}}(F_\rho)$ by at most 5.2%, while the finite grid increases the predicted leading constant by at most 1.53%. The approximately linear trends in the right panel support the predicted logarithmic dependence on $1/\delta$.

Effect of Dirichlet Concentration To examine the prediction in Proposition 2.2, we fix $K = 10$, $\delta = 0.05$, and $\boldsymbol{\pi} = (0.4, 0.3, 0.03751_8)$, and draw $\boldsymbol{\theta}_t \sim \mathrm{Dir}(c\boldsymbol{\pi})$ for $c \in \{0.01, 0.1, 1, 3, 10, 100, 1000\}$. For each c, we run 1,000 paired trials of ASC-D using either the full vector $\boldsymbol{\theta}_t$ or the one-hot sample $\boldsymbol{e}_{Y_t}$, where $Y_t \mid \boldsymbol{\theta}_t \sim \mathrm{Cat}(\boldsymbol{\theta}_t)$. Both versions use the same grid, stopping rule, and maximum budget of 65,536.

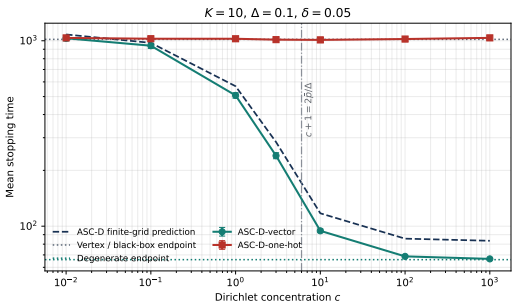

Figure 3: Mean stopping time versus Dirichlet concentration.

As shown in Figure 3, the one-hot stopping time is nearly constant because its marginal feedback law remains $\mathrm{Cat}(\boldsymbol{\pi})$, whereas full-vector ASC-D becomes faster as increasing concentration makes trajectories more consistent. The two methods perform similarly near the vertex endpoint; full-vector ASC-D is $4.22\times$ faster at $c = 3$ and $15.56\times$ faster at $c = 1000$, close to the limiting gain of $15.58\times$. The concentration-dependent gain and its eventual saturation are consistent with Proposition 2.2.

E ADDITIONAL REAL-WORLD ANALYSES

This section supplements Section 4.2 with uncertainty estimates for the modal-gap scaling analysis and budget-resolved results for the fixed-budget cross-dataset replication.

Uncertainty in Modal-Gap Scaling For each of the 90 common questions in Exp. 1, we regress log required budget on $\log(1/\Delta)$ across $T_a \in \{4, 8, 16, 32\}$. Table 6 reports the median within-question slope. Here the confidence intervals use 5,000 paired question-bootstrap resamples.

Table 6: Gap-scaling exponents on the 90 common questions.

	ASC-D	ASC-D (one-hot)	PPR-1v1	CITE-4
Median $\widehat{\beta}$	**0.766**	1.969	1.966	1.854
95% CI	$[0.692, 0.810]$	$[1.953, 2.001]$	$[1.949, 1.990]$	$[1.841, 1.861]$

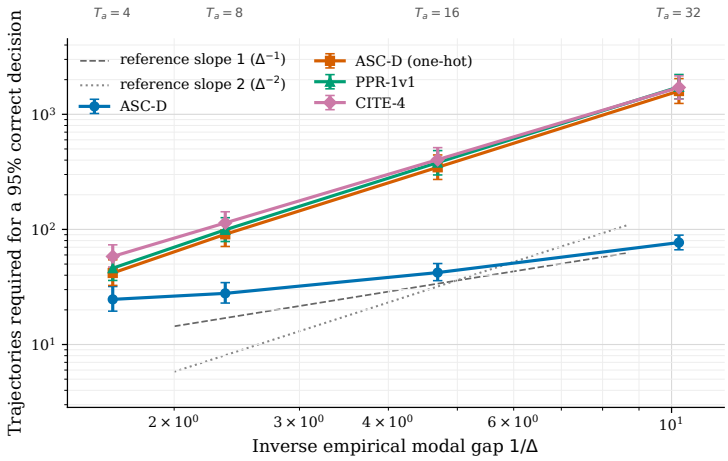

Figure 4: Required trajectory budget versus inverse empirical modal gap on the 90 common questions. Points show geometric means, error bars show 95% paired question-bootstrap confidence intervals, and reference lines have slopes one and two on the log–log scale.

The bootstrap intervals separate ASC-D from its matched one-hot ablation and the other hard-label methods. Figure 4 shows the same contrast directly: as the modal gap narrows, the required budget grows much more slowly for ASC-D than for ASC-D (one-hot), PPR-1v1, and CITE-4. These are finite-range empirical slopes and do not imply asymptotic scaling better than Δ^{-1}.

Cross-Dataset Replication on ARC-Challenge We replicate Exp. 2 on a fixed set of 100 four-choice ARC-Challenge questions using Qwen3-1.7B and Llama-3.2-1B. For each model and question, 128 probability vectors define the empirical population and its modal answer; we retain all 100 questions without screening by modal majority. We run 200 paired replays at $1 - \delta \in \{0.95, 0.975, 0.99\}$ with the same $N = 128$ trajectory budget as in Section 4.2. At each step, ASC-D observes a resampled probability vector, whereas CITE-4 and PPR-1v1 observe the same paired categorical draw. As in Exp. 2, a replay succeeds only if the method stops within the budget and returns the empirical mode.

Table 7: Correct-certification rate within $N = 128$ trajectories on 100 ARC-Challenge questions. Each entry aggregates 200 paired replays per question.

Model	Method	$1-\delta = 0.95$	$1-\delta = 0.975$	$1-\delta = 0.99$
Llama-3.2-1B	ASC-D	**74.48%**	**71.99%**	**69.06%**
	CITE-4	55.59%	53.40%	51.20%
	PPR-1v1	54.25%	52.43%	50.30%
Qwen3-1.7B	ASC-D	**97.48%**	**97.27%**	**96.94%**
	CITE-4	97.07%	96.72%	96.20%
	PPR-1v1	96.86%	96.55%	95.97%

Table 7 shows that ASC-D achieves the highest correct-certification rate for both models at every confidence level. The gains are larger for Llama-3.2-1B: ASC-D outperforms CITE-4 by 17.86–18.89 percentage points and PPR-1v1 by 18.76–20.23 percentage points. For Qwen3-1.7B, where all methods already achieve high certification rates, the gains are smaller but remain consistent, ranging from 0.41 to 0.97 percentage points. These results demonstrate that the benefit of distribution-valued feedback transfers from MMLU-Redux to ARC-Challenge and is especially substantial in more challenging certification settings.