Learning to Persuade a Biased Receiver

Yuqi Pan
Harvard University
yuqipan@g.harvard.edu

Sadie Zhao
Harvard University
sadie_zhao@g.harvard.edu

Milind Tambe
Harvard University
milind_tambe@harvard.edu

Yiling Chen
Harvard University
yiling@seas.harvard.edu

Abstract

We study a repeated information design setting in which the receiver, who is also the decision-maker, updates beliefs in a systematically biased way. More specifically, a distorted posterior in our model can be written as a convex combination of the prior and the Bayesian posterior, governed by a fixed but unknown parameter. Over repeated interactions, the sender chooses persuasive signaling schemes, observes only the receiver's realized actions, and seeks to minimize regret relative to a full-information oracle that knows the receiver's biased updating rule. We propose a safe exploration algorithm for learning the receiver's bias while maintaining high persuasion value. The algorithm exploits the asymmetric cost of probing: conservative probes incur only local loss, whereas overly aggressive probes may lose the persuasive opportunity entirely. For general finite state and action spaces and arbitrary bounded utilities, our method achieves $O(\log\log T)$ regret. A matching $\Omega(\log\log T)$ lower bound shows that this rate is optimal. We further discuss the influence on receiver welfare, as well as extensions to jointly unknown prior and bias, and contextual settings with time-varying priors and utilities.

1 Introduction

Information designers, such as AI platforms, LLM-based assistants, and automated warning systems, often observe utility-relevant state information and decide how to present it to a human decision-maker (receiver). Effective communication depends not only on how informative the signal is, but also on how the receiver updates her belief in response to algorithmic evidence. Classical Bayesian persuasion models the receiver as updating by Bayes' rule [27]. Yet when algorithmic systems advise human receivers, they may underweight algorithmic advice, consistent with algorithm aversion phenomenon [16, 31]: humans may anchor on prior judgments or treat algorithmic explanations as only partially credible. Thus the same message can induce different actions across receivers, not because the evidence differs, but because receivers differ in how strongly they incorporate it [2, 4, 14].

As a running example, consider an online marketplace that communicates fraud-risk information to merchants through risk scores or warning labels. After seeing the message, a merchant decides whether to ship the order or request additional verification. The platform and the merchant both care about fraud, but their objectives are not identical: the platform also values marketplace-wide integrity and buyer trust, while each merchant trades off verification costs against fraud risk. Moreover, merchants may react differently to the same warning: one may treat it as strong evidence of fraud, while another may discount the warning and stay close to her prior assessment. If the platform knew a merchant's updating strength, it could calibrate messages accordingly.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Motivated by such settings, we study a repeated information-design problem with an unknown biased receiver. In each of T rounds, the sender picks a signaling scheme that maps the state i.i.d. sampled from the prior to a randomized signal. After observing the realized signal, the receiver forms a biased posterior belief and takes a best-response action accordingly. Here the receiver's biased updating rule is fixed across rounds but unknown to the sender, capturing repeated interactions with the same receiver or with a stable receiver type. The sender observes this action and uses the history of past signals and actions to update future signaling schemes, whose goal is to achieve high long-run utility.

To quantify biased belief updating, we adopt a standard linear model [11, 15, 20, 25, 34]. If a signal induces a Bayesian posterior ν from prior μ_0, a receiver with bias level $\alpha^\star \in [0, 1]$ instead acts on a distorted belief $\alpha^\star \nu + (1-\alpha^\star)\mu_0$. The case $\alpha^\star = 1$ corresponds to Bayesian updating, while $\alpha^\star = 0$ means that the receiver completely ignores the signal and keeps the prior. Our setting is related to, but conceptually distinct from, existing work on persuasion with biased receivers [11, 15, 22, 29]. Most closely, Chen et al. [11] study the same linear belief distortion, but their objective is diagnostic, i.e., deciding whether the receiver's bias exceeds a threshold. In our setting, a new exploration–exploitation tradeoff arises: a diagnostic signal may reveal $\alpha^\star$ but perform poorly for current utility, while a utility-maximizing signal may reveal little about how the receiver processes evidence.

Moreover, feedback is indirect and discretized: an observed action is informative only when different candidate bias levels would induce different best responses under the realized posterior. The central challenge is therefore to learn the unknown updating strength from coarse action feedback without sacrificing too much persuasion utility along the way.

Our results. We first isolate the core structure in a binary-state, binary-action setting. There, the full-information optimal scheme is pinned down by the posterior cutoff at which the biased receiver switches actions. A standard binary search over this cutoff yields $O(\log T)$ regret. Our main insight to obtain lower regret is that probing losses are asymmetric: a conservative probe that stays on the safe side still induces the desired action and loses only a local utility gap, whereas an aggressive probe may lose the persuasion opportunity entirely. This mirrors online posted-price auctions [28]. Exploiting this asymmetry, our Safe Exploration algorithm approaches the unknown cutoff using increasingly fine probes from the safe side and achieves $O(\log \log T)$ regret. We further prove a matching $\Omega(\log \log T)$ lower bound, which also implies an $\Omega(\log \log T)$ lower bound for the general setting. We then extend the safe-exploration idea to general finite state and action spaces with arbitrary bounded utilities. In this setting, there is no single posterior cutoff: each action corresponds to a polyhedral region in the Bayesian-posterior simplex, and these regions move with $\alpha^\star$. We construct interval-safe schemes whose support remains incentive-compatible throughout the current bias uncertainty interval, and perturb only binding movable constraints to create local one-dimensional probes while preserving Bayes plausibility. The resulting General Safe Exploration algorithm runs in polynomial time and achieves $O(\log \log T)$ regret. Finally, we discuss several extensions concerning receiver welfare, jointly unknown prior and bias, and contextual environments with time-varying priors and utilities.

Related work. Our work is closely related to the literature on non-Bayesian belief updating [2, 5, 6, 12, 13, 19, 24, 33, 35], Bayesian persuasion and information design [7, 17, 18, 23, 27, 32], persuasion with biased receivers [11, 15, 22, 29, 34], and learning in information design [3, 9, 10, 21, 26, 30, 36, 38]. Due to space limits, we defer a detailed discussion to Appendix A.

2 Problem setup

Primitives. Let Ω and A be finite state and action spaces, and S a signal space. The common prior is $\mu_0 \in \Delta(\Omega)$. Sender and receiver utility functions $u_S, u_R : A \times \Omega \to \mathbb{R}$ are publicly known. We write $\mathcal{I} = (\Omega, A, \mu_0, u_S, u_R)$ for the known problem instance. The receiver's true bias $\alpha^\star \in (0, 1]$ is fixed but unknown to the sender.

A signaling scheme is a map $\pi : \Omega \to \Delta(S)$. The realized state $\omega \in \Omega$ is observed by the sender but not the receiver. The sender then draws a signal $s \in S$ following $\pi(\cdot \mid \omega)$ and sends it to the receiver. A standard Bayesian receiver forms the posterior belief $\nu_s \in \Delta(\Omega)$ from the signal s, where $\nu_s(\omega) = \mathbb{P}[\omega|s] = \mu_0(\omega)\pi(s|\omega) / \sum_{\omega' \in \Omega} \mu_0(\omega')\pi(s|\omega')$. Equivalently, a signaling scheme can be viewed as a distribution over posteriors $\rho \in \Delta(\Delta(\Omega))$. We now introduce the well-known splitting lemma.

Lemma 1 (Splitting Lemma [27])**.** *A distribution of posteriors ρ is Bayes-plausible (i.e., can be induced by some signaling scheme π) iff the expected posterior equals the prior: $\mathbb{E}_{\nu\sim\rho}[\nu] = \mu_0$.*

Therefore, the sender equivalently chooses a Bayes-plausible distribution $\rho \in \Delta(\Delta(\Omega))$ satisfying $\mathbb{E}_{\nu\sim\rho}[\nu] = \mu_0$. We adopt this posterior-based view throughout the paper.

Biased belief updating. We consider an $\alpha^\star$-biased receiver, where bias level $\alpha^\star \in (0, 1]$ is fixed across rounds but unknown to the sender. We adopt the widely-used linear distortion model [11, 15, 20, 25, 34], which captures partial incorporation of signal evidence. Mathematically, given Bayesian posterior ν and bias level $\alpha \in [0, 1]$, the receiver produces a distorted posterior $\widehat{\nu} = (1-\alpha)\mu_0 + \alpha\nu$, and chooses the corresponding best action:

$$a^\star(\nu; \alpha) \in \arg\max_{a\in A} \sum_{\omega\in\Omega} \left((1-\alpha)\mu_0(\omega) + \alpha\nu(\omega)\right) u_R(a, \omega),$$

with a fixed tie-breaking rule satisfying mild conditions, specified in Section 4. Throughout the paper, we reserve ν for Bayesian posteriors and $\widehat{\nu}$ for distorted posteriors. Note that although the receiver optimizes with respect to the distorted posterior $\widehat{\nu}$, the sender can only design the Bayesian posterior ν, so we always discuss the action region within the Bayesian posterior space.

Timing. The interaction proceeds for T rounds. Let $h_0 = \emptyset$ and $h_t = \{(\rho_\tau, \omega_\tau, \nu_\tau, a_\tau)\}_{\tau=1}^t$ denote the sender's history up to the end of round t, consisting of previous signaling schemes, realized states, induced posteriors, and receiver's adopted actions. Let Π denote the sender's adaptive policy, which is a mapping from histories to Bayes-plausible distributions over Bayesian posteriors. Π may depend on $\mathcal{I}$ and T, but not the unknown $\alpha^\star$. In $t = 1, 2, \ldots, T$, the interaction proceeds as follows:

1. Based on h_{t-1}, the sender publicly commits to a Bayes-plausible distribution (equivalently, a signaling scheme) over posteriors $\rho_t = \Pi(h_{t-1})$.
2. A state is drawn according to the prior: $\omega_t \sim \mu_0$. The sender observes ω_t.
3. Given ω_t, the sender draws a signal from the chosen signaling scheme and sends it to the receiver. The realized signal induces a posterior belief $\nu_t \in \Delta(\Omega)$.
4. After observing the signal, the receiver forms the distorted posterior $(1-\alpha^\star)\mu_0 + \alpha^\star\nu_t$ and chooses an action $a_t = a^\star(\nu_t; \alpha^\star)$.
5. The sender and receiver obtain utility $u_S(a_t, \omega_t)$ and $u_R(a_t, \omega_t)$. The sender observes the receiver's action a_t, and the history is updated to h_t.

Objective and regret. For a fixed instance $\mathcal{I}$ and bias α, define the sender's reduced-form expected utility at Bayesian posterior ν as

$$V_{\mathcal{I}}^{\alpha}(\nu) = \sum_{\omega\in\Omega} \nu(\omega) u_S(a^\star(\nu; \alpha), \omega) .$$

The full-information one-round benchmark knows $\alpha^\star$ and plays an optimal Bayes-plausible posterior distribution in every round, which is defined as:

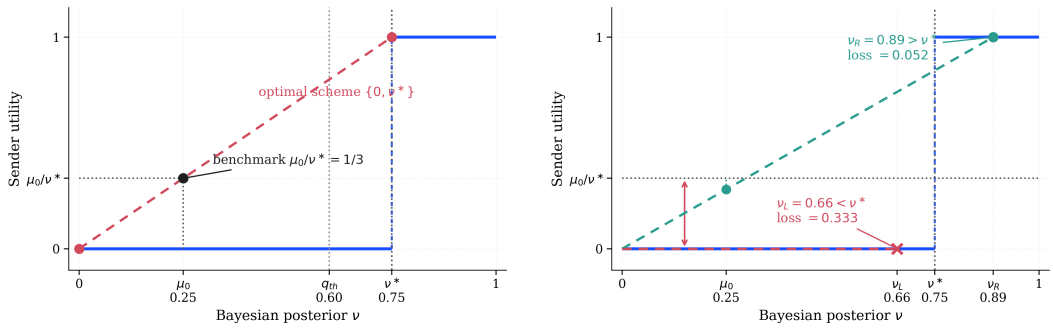
$$\mathrm{OPT}_{\mathcal{I}}(\alpha^\star) = \sup_{\rho\in\Delta(\Delta(\Omega)):\mathbb{E}_{\nu\sim\rho}[\nu]=\mu_0} \mathbb{E}_{\nu\sim\rho}\left[V_{\mathcal{I}}^{\alpha^\star}(\nu)\right]$$

Let $\mathbb{P}^{\Pi}_{t,\mathcal{I},\alpha^\star}$ denote the distribution of history h_t under $\Pi, \mathcal{I}, \alpha^\star$. The expected regret is

$$\mathrm{Reg}_T^{\mathcal{I}}(\Pi; \alpha^\star) = T\,\mathrm{OPT}_{\mathcal{I}}(\alpha^\star) - \sum_{t=1}^{T} \mathbb{E}_{h_{t-1}\sim\mathbb{P}^{\Pi}_{t-1,\mathcal{I},\alpha^\star}}\left[\mathbb{E}_{\nu_t\sim\Pi(h_{t-1})}\left[V_{\mathcal{I}}^{\alpha^\star}(\nu_t)\right]\right].$$

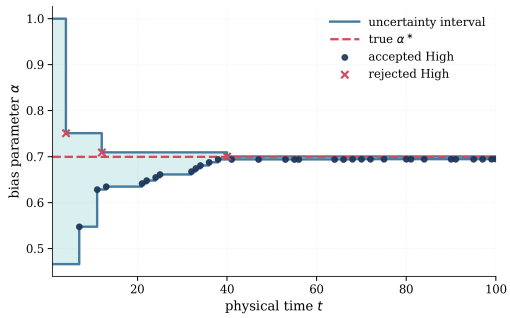
3 Warm-up: safe exploration via a one-dimensional threshold

We begin with a binary specialization that isolates the main learning structure. Let $\Omega = A = \{0, 1\}$. With a slight abuse of notation, write $\mu_0 \in (0, 1)$ for $\mu_0(1)$ and $\nu \in [0, 1]$ for $\nu(1)$. Moreover, we consider the canonical sender utility $u_S(a, \omega) = \mathbb{1}\{a = 1\}$, so the sender only cares about inducing

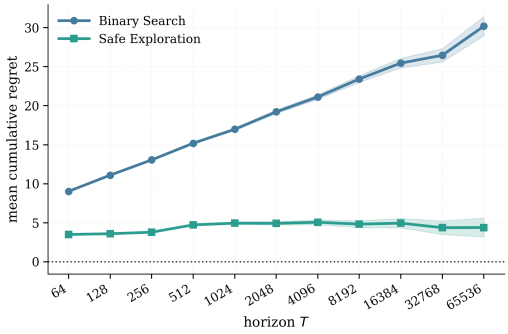


(a) Full-information benchmark

(b) Asymmetric probe loss



(c) Safe-side probing sample path



(d) $O(\log T)$ regret vs. $O(\log\log T)$ regret

Figure 1: Binary instance $B = (\mu_0, q_{\text{th}}) = (0.25, 0.60)$ and true bias $\alpha^\star = 0.70$.

action 1. On the receiver side, action 0 is set as optimal at the prior without loss of generality, otherwise the sender can induce action 1 without learning and regret is zero.

Since the receiver's utility difference between actions 1 and 0 is affine in the distorted belief, the action follows a cutoff rule. Let $q_{\text{th}} \in (\mu_0, 1)$ denote the corresponding cutoff: action 1 is chosen iff $(1-\alpha)\mu_0 + \alpha\nu \geq q_{\text{th}}$. Equivalently, this induces the Bayesian-posterior cutoff: $\nu_{\mathcal{B}}(\alpha) = \mu_0 + (q_{\text{th}} - \mu_0)/\alpha$. Define $R_1^\alpha = [\nu_{\mathcal{B}}(\alpha), 1]$ and $R_0^\alpha = [0, \nu_{\mathcal{B}}(\alpha)]$ as the Bayesian posterior regions that weakly induce action 1 and 0 respectively. For true bias level $\alpha^\star$, we write $\nu^\star = \nu_{\mathcal{B}}(\alpha^\star)$. Figure 1(a) illustrates this structure, where the blue step function is the sender's reduced-form utility function over Bayesian posterior. We denote the binary instance by $\mathcal{B} = (\mu_0, q_{\text{th}})$.

If $\alpha^\star < \alpha_{\min}(\mathcal{B}) := (q_{\text{th}} - \mu_0)/(1 - \mu_0)$, then $R_1^{\alpha^\star} = \emptyset$ and persuasion is infeasible, so the regret is trivially zero. We focus on $\alpha^\star \in [\alpha_{\min}(\mathcal{B}), 1]$, for any α, define corresponding two-point signaling scheme $\tau(\alpha)$ as the Bayes-plausible distribution over the two posteriors $\{0, \nu_{\mathcal{B}}(\alpha)\}$ that places mass $1 - \mu_0/\nu_{\mathcal{B}}(\alpha)$ on 0 and $\mu_0/\nu_{\mathcal{B}}(\alpha)$ on $\nu_{\mathcal{B}}(\alpha)$. The full-information optimum is attained by $\tau(\alpha^\star)$, with value $\mu_0/\nu_{\mathcal{B}}(\alpha^\star)$. Intuitively, any posterior strictly above $\nu^\star$ is wasteful because it induces the same action while using more positive posterior mass in the Bayes-plausibility constraint $\mathbb{E}[\nu] = \mu_0$. Any posterior below $\nu^\star$ is also wasteful because it produces no utility while still using posterior mass. Thus the optimal scheme puts persuasive mass at the lowest useful posterior $\nu^\star$ and offsets it with the lowest posterior 0, which is the red line in Figure 1(a). Learning the optimal scheme is therefore reduced to learning the unknown threshold $\nu^\star$.

Binary search (BS). Since $\nu_{\mathcal{B}}(\alpha)$ is monotone in α, we equivalently maintain an uncertainty interval for the unknown bias $\alpha^\star$. Let $J = [\underline{\alpha}, \overline{\alpha}] \ni \alpha^\star$ denote the current uncertain bias interval. Given J, the natural safe choice is obtained by restricting attention to posteriors that induce action 1 for every $\alpha \in J$. Define *interval-safe region* $R_1^J = \bigcap_{\alpha \in J} R_1^\alpha$, i.e., the region of Bayesian posteriors that can robustly induce action 1 within J. Since $\nu_{\mathcal{B}}(\alpha)$ is decreasing in α, $R_1^J = R_1^{\underline{\alpha}} = [\nu_{\mathcal{B}}(\underline{\alpha}), 1]$. Hence scheme $\tau(\underline{\alpha})$ is a safe choice: its high posterior robustly induces ideal action 1 for every $\alpha \in J$, with regret $\mu_0/\nu_{\mathcal{B}}(\alpha^\star) - \mu_0/\nu_{\mathcal{B}}(\underline{\alpha})$, which vanishes as $\underline{\alpha} \uparrow \alpha^\star$.

To achieve low regret, we shrink J by linking interval updates to realized actions. If the sender plays $\tau(m)$, then upon realizing the high posterior $\nu_{\mathcal{B}}(m)$, the receiver chooses action 1 iff $\nu_{\mathcal{B}}(m) \geq$

$\nu_{\mathcal{B}}(\alpha^\star) \Leftrightarrow m \le \alpha^\star$, Thus the sender can update J to either $[m, \overline{\alpha}]$ or $[\underline{\alpha}, m]$. A natural algorithm is binary search: probe $m = (\underline{\alpha} + \overline{\alpha})/2$ by playing $\tau(m)$ until $\nu_{\mathcal{B}}(m)$ is realized. Each informative realization halves J. After $O(\log T)$ such realizations, $|J| = O(1/T)$, so the sender commits to the safe scheme $\tau(\underline{\alpha})$ thereafter. The full algorithm is in Section B.2 and guarantees $O(\log T)$ regret.

Proposition 1. *For every binary instance $\mathcal{B} = (\mu_0, q_{\mathrm{th}})$ and every bias level $\alpha^\star \in (0, 1]$, $\mathrm{Reg}_T^{\mathcal{B}}\big(\Pi^{\mathrm{BS}}; \alpha^\star\big) = O(\log T)$ holds with a hidden constant $C_{\mathrm{BS}}(\mathcal{B})$.*

Throughout the paper, all asymptotic bounds are pointwise in the fixed learning instance, which means the constant is fixed once $I = (\Omega, A, \mu_0, u_S, u_R)$ and $\alpha^\star$ are fixed, and never depends on the horizon T. The constants in Sections 3 and 4 are collected in Sections B.1 and C.1 respectively.

Safe exploration (SE). Binary search is suboptimal because it ignores the asymmetric losses on the two sides of a probe, see Figure 1(b). For a candidate $m \in J$ with probing scheme $\tau(m)$, if $m \le \alpha^\star$, the regret is only the local gap $\mu_0/\nu_{\mathcal{B}}(\alpha^\star) - \mu_0/\nu_{\mathcal{B}}(m)$, which vanishes as $m \uparrow \alpha^\star$. If $m > \alpha^\star$, however, the high posterior $\nu_{\mathcal{B}}(m) < \nu^*$ no longer induces action 1, causing regret $\mu_0/\nu_{\mathcal{B}}(\alpha^\star)$. Thus, a conservative probe loses only a local utility gap, whereas an overly aggressive probe may lose the entire persuasive opportunity. This mirrors online posted-price auction [28], where a seller repeatedly posts a price to a buyer with unknown valuation. There, posting below the buyer's value only loses the price gap, whereas posting above it loses the sale entirely.

Our *Safe Exploration* algorithm, Algorithm 1, exploits this asymmetry by probing from the safe side with a small step. In each phase, starting from $J = [\underline{\alpha}, \overline{\alpha}]$, it scans $m = \underline{\alpha} + \epsilon, \underline{\alpha} + 2\epsilon, \ldots$ from left to right and plays $\tau(m)$, where ϵ is initialized to $1/2$ and squared after each interval update. The scan stops at the first informative realization for which the high posterior induces action 0, and updates J to the adjacent bracket $[m_{\mathrm{prev}}, m]$. If no switch occurs before the next probe exceeds $\overline{\alpha}$, it updates J to $[m_{\mathrm{prev}}, \overline{\alpha}]$. Figure 1(c) illustrates the safe-side nature of the algorithm: each phase scans upward from the conservative endpoint, accumulating many low-loss safe probes near $\alpha^\star$ before at most one aggressive probe closes the phase. Since the new interval length is at most the current step size and the next step size is squared, J shrinks doubly exponentially, so $O(\log\log T)$ phases suffice to reach length $1/T$. Algorithm 1 guarantees $O(\log\log T)$ regret in total. The simulation in Figure 1(d) also matches the $O(\log T)$ versus $O(\log\log T)$ comparison.

Geometrically, each probe enlarges the interval-safe region R_1^J. Starting from the safe scheme $\tau(\underline{\alpha})$, whose high posterior lies on the boundary, replacing $\nu(\underline{\alpha})$ with $\nu(m)$ pushes this boundary outward. As J shrinks, the interval-safe region expands toward the true action region $R_1^{\alpha^\star}$.

Proposition 2. *For every binary instance $\mathcal{B} = (\mu_0, q_{\mathrm{th}})$ and every bias level $\alpha^\star \in (0, 1]$, $\mathrm{Reg}_T^{\mathcal{B}}\big(\Pi^{\mathrm{SE}}; \alpha^\star\big) = O(\log\log T)$ holds with a hidden constant $C_{\mathrm{SE}}(\mathcal{B})$.*

Algorithm 1 Safe Exploration (SE)

Require: Horizon T, binary instance $\mathcal{B} = (\mu_0, q_{\mathrm{th}})$
$\underline{\alpha} \leftarrow \alpha_{\min}(\mathcal{B})$, $\overline{\alpha} \leftarrow 1$, $\epsilon \leftarrow 1/2$, $t \leftarrow 1$
Stage 1: Safe Exploration
while $\overline{\alpha} - \underline{\alpha} > T^{-1}$ and $t \le T$ **do**
  $m_{\mathrm{prev}} \leftarrow \underline{\alpha}$, $m \leftarrow \underline{\alpha} + \epsilon$
  while $m \le \overline{\alpha}$ and $t \le T$ **do**
    commit to and play $\tau(m)$
    if the realized posterior is not 0 **then**
      if the receiver chooses action 1 **then**
        $m_{\mathrm{prev}} \leftarrow m,\ m \leftarrow m + \epsilon$
      else
        $(\underline{\alpha}, \overline{\alpha}) \leftarrow (m_{\mathrm{prev}}, m),\ \epsilon \leftarrow \epsilon^2,\ t \leftarrow t + 1$
        break
    $t \leftarrow t + 1$
  if $m > \overline{\alpha}$ **then**
    $(\underline{\alpha}, \overline{\alpha}) \leftarrow (m_{\mathrm{prev}}, \overline{\alpha}),\ \epsilon \leftarrow \epsilon^2$
Stage 2: Commitment
Commit to $\tau(\underline{\alpha})$ for all remaining rounds.

Lower bound. We now prove a matching $\Omega(\log\log T)$ lower bound, showing that Algorithm 1 is asymptotically optimal. The proof is inspired by Kleinberg and Leighton [28], but must handle a new difficulty: information about the unknown threshold arrives only through informative signal realizations, while regret is incurred in every physical round. We apply Yao's principle [37] to reduce the minimax lower bound to deterministic sender strategies under a carefully chosen bias distribution, and then use martingale tools to translate the bound back to horizon T. This yields the following theorem, which also gives an $\Omega(\log\log T)$ lower bound for the general case in Section 4.

Theorem 1 (lower bound)**.** *For any binary instance $\mathcal{B} = (\mu_0, q_{\text{th}})$ and any randomized sender strategy Π, $\sup_{\alpha^\star\in[0,1]} \mathrm{Reg}_T^{\mathcal{B}}(\Pi;\alpha^\star) = \Omega(\log\log T)$ holds with a hidden constant $c_{\text{LB}}(\mathcal{B})$.*

4 General setting: safe exploration in moving posterior geometry

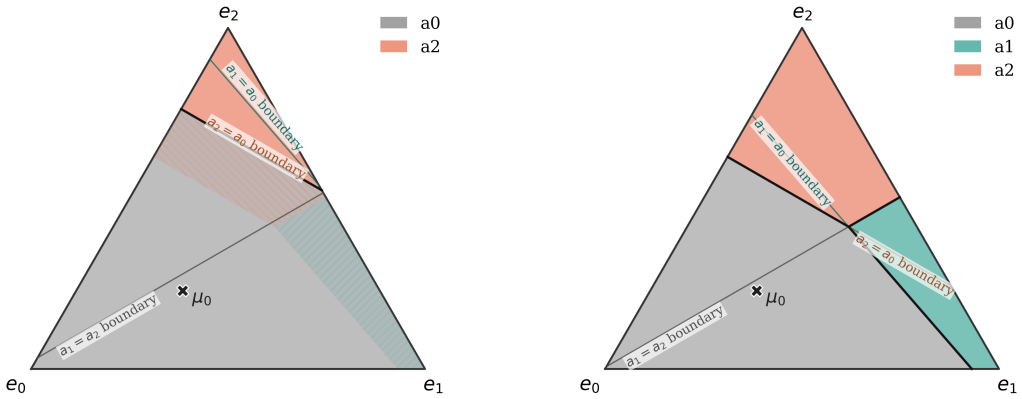
The binary case shows why safe exploration beats binary search: probing losses are asymmetric. We now similarly define interval-safe action region in the general setting. For any action $a \in A$ and bias $\alpha > 0$, let R_a^α be the region of Bayesian posteriors that weakly induce action a:

$$R_a^\alpha = \left\{\nu \in \Delta(\Omega) : \Delta u_{a,a'}^\top \nu \geq \frac{\alpha-1}{\alpha}\Delta u_{a,a'}^\top \mu_0 = b_{a,a'}(\alpha), \quad \forall a' \neq a\right\}.$$

Here $\Delta u_{a,a'} \in \mathbb{R}^{|\Omega|}$ is defined component-wise by $\big(\Delta u_{a,a'}\big)(\omega) = u_R(a,\omega) - u_R(a',\omega)$. Each inequality is a weak incentive-compatibility (IC) constraint, ensuring that α-biased receiver weakly prefers a to a'. For an interval $J = [\underline{\alpha}, \overline{\alpha}]$, define the interval-safe region, i.e., the set of Bayesian posteriors under which action a remains weakly IC for all $\alpha \in J$:

$$R_a^J = \bigcap_{\alpha\in J} R_a^\alpha = \left\{\nu \in \Delta(\Omega) : \Delta u_{a,a'}^\top \nu \geq b_{a,a'}(J) = \max\left\{b_{a,a'}(\underline{\alpha}), b_{a,a'}(\overline{\alpha})\right\}, \ \forall a' \neq a\right\} \quad (1)$$

The last equality follows because $b_{a,a'}(\alpha)$ is monotone in α, with derivative $\Delta u_{a,a'}^\top \mu_0/\alpha^2$. Thus the binding endpoint can differ across constraints: if $\Delta u_{a,a'}^\top \mu_0 > 0$, the safe constraint is determined by $\overline{\alpha}$, while if $\Delta u_{a,a'}^\top \mu_0 < 0$, it is determined by $\underline{\alpha}$. This is the main geometric difference from the binary case. In Figure 2, for example, the default-action region $R_{a_0}^\alpha$ is smaller in the high-bias panel, while the region for a_2 moves in the opposite direction.



(a) $\alpha_{\text{low}} = 0.55$: $R_{a_1}^{\alpha_{\text{low}}} = \emptyset$ (b) $\alpha_{\text{high}} = 0.85$: $R_{a_1}^{\alpha_{\text{high}}} \neq \emptyset$

Figure 2: Action regions for $\mu_0 = (0.50, 0.27, 0.23)$. In state order (e_0, e_1, e_2), $u_R(a_0,\cdot) = (0,0,0)$, $u_R(a_1,\cdot) = (-2.1, 0.3, 0.9)$, and $u_R(a_2,\cdot) = (-2.1, -0.3, 1.5)$.

We now specify the condition for tie-breaking rules. Define the set of relevant actions:

$$A_{\text{rel}} = \left\{a \in A : \exists \nu \in \Delta(\Omega) \text{ that } a \text{ uniquely maximizes } \sum_{\omega\in\Omega}(\alpha^\star\nu(\omega) + (1-\alpha^\star)\mu_0(\omega))u_R(a,\omega)\right\}.$$

Actions outside A_{rel} are weakly dominated for the receiver: they are never unique best responses under the true bias. We impose the following relevant-action tie-breaking convention: the receiver's

tie-breaking rule never selects any action in $A \setminus A_{\mathrm{rel}}$. This restriction can be understood as a robustness requirement on the regret benchmark. If tie-only actions were allowed, the full-information benchmark could target a posterior lying exactly on the corresponding indifference set, while a learner who does not perfectly know $\alpha^\star$ could not robustly induce the same action. In Section C.3, we show that without this restriction, low regret cannot be guaranteed. Also, note that although A_{rel} depends on the unknown $\alpha^\star$, the sender does not need to know this specific set as Algorithm 2 is run over the original action set A. With additional strict-feasibility check defined in Section 4.2.1, the sender can identify A_{rel} without knowing $\alpha^\star$. Among relevant actions, the receiver breaks ties in favor of the sender, as a standard rule in Bayesian persuasion [17, 27].

We also keep one standard assumption from Chen et al. [11] as below.

Assumption 1 (Unique Default Action [11])**.** *There is a unique action a_0 that maximizes the expected receiver utility based on the prior μ_0 : $|\arg\max_{a\in A}\{\sum_{\omega\in\Omega}\mu_0(\omega)u_R(a,\omega)\}| = 1$.*

We propose the *General Safe Exploration (GSE)* algorithm. The algorithm proceeds in three stages, corresponding to pure exploration, a balanced exploration–exploitation stage, and pure exploitation. First, the localization stage quickly shrinks the bias interval to length $O(1/\log T)$ so that interval-safe regions are not overly conservative. Second, the safe exploration stage follows the same phase structure as Algorithm 1, to shrink the interval quadratically until its length reaches $O(1/T)$. Third, the commitment stage plays an interval-safe scheme. The resulting regret is $O(\log\log T)$ in total.

Algorithm 2 General Safe Exploration (GSE)

Require: Horizon T, general instance $\mathcal{I} = (\Omega, A, \mu_0, u_S, u_R)$
Initialize the bias interval $J \leftarrow [\alpha_{\min}(\mathcal{I}), 1]$.
Stage 1: Localization
while $|J| > 1/\log T$ **do**
    Run ThresholdTest(J, m) at the midpoint m and update J. ▷ via Algorithm 5
Stage 2: Safe exploration
Set $J_0 \leftarrow J, r \leftarrow 0$.
while $|J_r| > T^{-1}$ **do**
    Run SafeExplore(J_r) to obtain (J_{r+1}, τ^{J_r}). ▷ via Algorithm 3
    if $J_{r+1} = J_r$ **then**
        break ▷ The current safe scheme is already optimal
    $r \leftarrow r + 1$
Stage 3: Commitment
Compute and commit to the current interval-safe scheme for all remaining rounds.

Theorem 2 (General safe exploration)**.** *For every instance $\mathcal{I} = (\Omega, A, \mu_0, u_S, u_R)$ satisfying Assumption 1 and every $\alpha^\star \in (0, 1]$, Algorithm 2 runs in polynomial time with respect to $\mathcal{I}$ and T. Moreover, $\mathrm{Reg}_T^{\mathcal{I}}\big(\Pi^{\mathrm{GSE}}; \alpha^\star\big) = O(\log\log T)$ holds with a hidden constant $C_{\mathrm{GSE}}(\mathcal{I}, \alpha^\star)$.*

4.1 Localization stage

Figure 2 shows why starting safe exploration on a wide interval can be overly conservative. The interval-safe region $R_{a_1}^J$ over a wide interval J is empty, even though a_1 may be relevant and utility-improving at the true high bias. The localization stage first shrinks the interval so that interval safety does not discard such actions. Specifically, the goal in this stage is localizing $\alpha^\star$ to an $O(1/\log T)$-length interval as quickly as possible, without caring about utility loss. We use the threshold test of Chen et al. [11] (see Algorithm 5): for each candidate bias threshold β, it solves LP (see equation 2) to construct the signaling scheme, where most posteriors lie on the default/non-default indifference boundary at β, so the realized action reveals whether $\alpha^\star \geq \beta$ or $\alpha^\star < \beta$. We show that each test is informative with constant probability, so binary search reaches length $O(1/\log T)$ after $O(\log\log T)$ informative comparisons and incurs $O(\log\log T)$ total regret. Details are deferred to Section C.2.

4.2 Safe exploration stage

This stage adapts the safe exploration principle of Algorithm 1 to the general moving-polytope geometry. We give a short version in Algorithm 3 and defer the full implementation to Algorithm 6.

First, Section 4.2.1 explains how, on the current uncertainty interval J, we compute an interval-safe vertex-supported scheme, i.e., a Bayes-plausible utility-maximizing scheme supported on vertices of the safe polytopes $\{R_a^J\}_{a\in A_{\rm rel}}$, so that every support posterior remains IC for all $\alpha \in J$. Second, Section 4.2.2 explains how vertex posteriors of this scheme are converted into local probes by perturbing movable binding IC constraints. Similar to Algorithm 1, each informative realization either safely advances an endpoint of J by $\eta = |J|^2$ or localizes $\alpha^\star$ to the adjacent length-η interval. Thus each phase reduces $|J|$ to at most $|J|^2$, and Section 4.2.3 bounds waiting time and total regret.

Algorithm 3 Fixed-Interval Safe Exploration: Short Version

Require: Current interval $J = [\underline{\alpha}, \overline{\alpha}]$, instance $\mathcal{I} = (\Omega, A, \mu_0, u_S, u_R)$.
$L \leftarrow |J|$, $\eta \leftarrow L^2$, $(\ell, r) \leftarrow (\underline{\alpha}, \overline{\alpha})$
$\tau^J_{\rm vtx} \leftarrow$ VERTEXSAFESCHEME$(J, \mathcal{I})$
if $P_{\rm info}(\tau^J_{\rm vtx}) = 0$ **then**
  return $(J, \tau^J_{\rm vtx})$
for each informative index $i \in {\rm MOV}(\tau^J_{\rm vtx})$ **do**
  $\mathsf{ori}_i \leftarrow$ CHOOSEMOVABLEBINDINGCONSTRAINT$(i, \tau^J_{\rm vtx}, J)$
$\tau^{\rm last} \leftarrow \tau^J_{\rm vtx}$
while $r - \ell > \eta$ **do**
  $\tau^{\rm last} \leftarrow$ BUILDPROBE $\left(\tau^J_{\rm vtx}, \{\mathsf{ori}_i\}_{i\in{\rm MOV}(\tau^J_{\rm vtx})}, \ell, r, \eta\right)$
  $(i, a_t) \leftarrow$ RUNUNTILINFORMATIVE $\left(\tau^{\rm last}, {\rm MOV}(\tau^J_{\rm vtx})\right)$
  $(\ell, r, \mathsf{status}) \leftarrow$ REFINEINTERVAL$(i, a_t, \mathsf{ori}_i, \ell, r, \eta)$
return $([\ell, r], \tau^{\rm last})$

4.2.1 Interval-safe vertex-supported schemes

The interval-safe regions $\{R_a^J\}_{a\in A}$ were defined in Equation (1). For the current localized interval $J = [\underline{\alpha}, \overline{\alpha}]$, also define the strict interval-safe region ${\rm int}_{\rm IC}(R_a^J) = \{\nu \in \Delta(\Omega) : \Delta u_{a,a'}^\top \nu > b_{a,a'}(J),\ \forall a' \neq a\}$. We additionally remove all actions with ${\rm int}_{\rm IC}(R_a^J) = \emptyset$. This feasibility check exactly recovers the relevant action set $A_{\rm rel}$, guaranteed once J is sufficiently small after localization, see Lemma 2 in Section C.3. Note that this is an action-level filter designed to exclude $A \setminus A_{\rm rel}$. For $a \in A_{\rm rel}$, we still consider the closed polytope R_a^J, as the boundary cases are addressed by the sender-favor tie-breaking rule.

View only posteriors in the interval-safe regions $\bigcup_{a\in A_{\rm rel}} R_a^J$ as feasible. Using the corresponding tighter IC constraints, we solve a LP in equation 3 to obtain an interval-safe scheme, i.e., a Bayes-plausible utility-maximizing scheme supported in the interval-safe regions $\bigcup_{a\in A_{\rm rel}} R_a^J$.

Unlike the binary case, where the optimal scheme $\tau(m)$ places its high posterior on the boundary of the interval-safe region R_1^J and is therefore easy to probe, an arbitrary interval-safe optimizer in the general case may obscure which posteriors are informative for probing. We therefore replace it by an equivalent vertex-supported scheme. Specifically, since each R_a^J is a polytope with finitely many vertices, each posterior in the interval-safe scheme can be decomposed into a distribution over vertices of the corresponding polytope. Such decomposition is without loss as it preserves utility and Bayes plausibility. Applying this to all posteriors gives an interval-safe vertex-supported scheme $\tau^J_{\rm vtx}$ on J, see Section C.4.1 for details. When J is small, its utility gap from optimal $\tau_{\rm opt}$ is also small.

Proposition 3 (Safe–optimal gap)**.** *For every interval $J \ni \alpha^\star$, $U_S(\tau_{\rm opt}) - U_S(\tau^J_{\rm vtx}) = O(|J|)$.*

This vertex view is the higher-dimensional analogue of the binary safe scheme $\tau(\underline{\alpha})$ where one posterior lies on the moving boundary of R_1^J. Here informative vertices are similarly identified by binding IC constraints, whose movement in α gives the local one-dimensional probes below.

4.2.2 Probing the bias via movable binding constraints

Based on the current interval-safe vertex-supported scheme $\tau^J_{\rm vtx}$, we construct the following probing scheme. Call a vertex posterior $\nu \in R_a^J$ informative if it has a movable binding IC constraint: $\exists a' \neq a$, $\Delta u_{a,a'}^\top \nu = b_{a,a'}(J) = b_{a,a'}(\beta)$ with $\beta \in \{\underline{\alpha}, \overline{\alpha}\}$ where the right-hand side varies with α, i.e., $\Delta u_{a_i,a_i'}^\top \mu_0 \neq 0$. Such a constraint plays the role of the binary cutoff. To probe a candidate value

$m \in J$, we move ν slightly across this one boundary, while leaving the other IC constraints satisfied. Denote the resulting posterior by $\nu^{\mathrm{pr}}(m)$. The formal construction is deferred to Section C.4.1. This local probe has exactly the same threshold interpretation as in the binary case. If the selected binding constraint is determined by the lower endpoint $\beta = \underline{\alpha}$, then

$$a^\star(\nu^{\mathrm{pr}}(m); \alpha^\star) = a \iff \alpha^\star \geq m, \qquad a^\star(\nu^{\mathrm{pr}}(m); \alpha^\star) \neq a \iff \alpha^\star < m$$

If it binds at the upper endpoint, $\beta = \overline{\alpha}$, the inequalities reverse:

$$a^\star(\nu^{\mathrm{pr}}(m); \alpha^\star) = a \iff \alpha^\star \leq m, \qquad a^\star(\nu^{\mathrm{pr}}(m); \alpha^\star) \neq a \iff \alpha^\star > m$$

Thus each informative vertex reduces the general problem to a local binary comparison.

We construct the probing scheme $\tau^{J,\eta}_{\mathrm{probe}}(\ell, r)$ by perturbing all informative vertices of τ^J_{vtx} simultaneously. Maintaining a current subinterval $(\ell, r) \subseteq J$ and step size $\eta = |J|^2$, a lower-endpoint vertex probes $m = \ell + \eta$, while an upper-endpoint vertex probes $m = r - \eta$. When an informative signal is realized, the observed action either moves the corresponding endpoint inward by η, or localizes $\alpha^\star$ to the adjacent length-η interval. This is the same safe-side logic as in Algorithm 1, except that different vertices may probe different sides of the interval.

We then bound the utility gap between $\tau^{J,\eta}_{\mathrm{probe}}(\ell, r)$ and τ^J_{vtx}. Note that the following property only addresses safe probes, and the wrong-action loss caused by unsafe probes is handled separately later.

Proposition 4 (Safe–probe gap)**.** *For every (ℓ, r), $U_S(\tau^J_{\mathrm{vtx}}) - U_S(\tau^{J,\eta}_{\mathrm{probe}}(\ell, r)) = O(P_{\mathrm{info}}(\tau^J_{\mathrm{vtx}})|J|)$.*

4.2.3 A sample complexity dichotomy

Propositions 3 and 4 control the per-round loss of a fixed probe. The only remaining issue is the waiting time for informative signals. Define the informative probability $P_{\mathrm{info}}\left(\tau^J_{\mathrm{vtx}}\right)$ of an interval-safe vertex-supported scheme τ^J_{vtx} as the total probability mass on informative vertex posteriors. The following dichotomy rules out slow learning.

Proposition 5. *There exists an instance-dependent constant $\epsilon_{\mathrm{dich}}(\mathcal{I}) > 0$, such that for every interval $J \ni \alpha^\star$ with $|J| \leq \epsilon_{\mathrm{dich}}(\mathcal{I})$, one of the following holds:*

1. *$\exists$ constant $c > 0$, every interval-safe vertex-supported optimizer satisfies $P_{\mathrm{info}}(\tau^J_{\mathrm{vtx}}) \geq c$.*
2. *There exists an optimal scheme τ_{opt} with $P_{\mathrm{info}}(\tau_{\mathrm{opt}}) = 0$.*

After the localization stage, $|J| = O(1/\log T) \leq \epsilon_{\mathrm{dich}}(\mathcal{I})$ for all sufficiently large T. In the first case, informative signals arrive after $O(1)$ rounds in expectation. A phase with interval J and step size $\eta = |J|^2$ has at most $O(|J|/\eta) = O(1/|J|)$ safe probes and one unsafe probe. Given Propositions 3 and 4, its expected regret is $O(1)$. In the second case, the optimal scheme is also safe within J, then there is no safe-optimal gap, only the safe-probe gap $O(P_{\mathrm{info}}|J|)$ remains. After multiplying by the $O(1/P_{\mathrm{info}})$ waiting time, each safe probe costs $O(|J|)$, hence the whole phase again costs $O(1)$. Finally, since every completed phase satisfies $|J_{r+1}| \leq |J_r|^2$, only $O(\log\log T)$ phases are needed.

Proposition 6. *The total expected regret during the safe exploration stage is $O(\log\log T)$.*

4.3 Commitment stage

After the safe-exploration stage, the sender commits to the current interval-safe vertex-supported optimizer τ^J_{vtx}. If the final interval satisfies $|J| \leq T^{-1}$, Proposition 3 implies that the per-round regret from the full-information benchmark is $O(T^{-1})$, so the total commitment regret is $O(1)$. If instead the safe-exploration routine stops earlier because $P_{\mathrm{info}}(\tau^J_{\mathrm{vtx}}) = 0$, then by the second branch of Proposition 5, committing to it incurs no additional regret. Combining the $O(\log\log T)$ localization cost, the $O(\log\log T)$ safe-exploration cost, and this $O(1)$ commitment cost proves Theorem 2.

5 Extensions and open problems

We consider three extensions beyond the baseline model.

For receiver welfare, we give two comparisons. First, relative to no persuasion, persuasion never hurts the biased receiver: her true expected utility under any feasible scheme is at least her no-persuasion

utility. Second, comparing a sender who knows the true bias with one who wrongly assumes a Bayesian receiver (both persuading), the effect on the receiver can go either way, so knowing the bias can leave the receiver better or worse off (see Section D).

We also allow priors and utilities to vary across rounds while the bias remains fixed. The sender observes the current prior and utilities before choosing a scheme, and the context sequence can be arbitrary. Under uniform contextual regularity, the contextual version of our safe exploration algorithm still achieves $O(\log \log T)$ regret (see Section E).

Finally, when both the prior and the bias are unknown, a prior-independent signaling family reduces the binary model to learning one implementable threshold, yielding $O(\log \log T)$ regret. In the general setting, the unknown-prior lower bound of Li and Lin [30] yields $\Omega(\log T)$ regret. A matching upper bound remains open (see Section F).

Further directions include extending safe exploration to other biased-updating models and understanding how the guarantees change under noisy or partial action feedback.

Acknowledgments and Disclosure of Funding

We thank Tao Lin for in-depth discussions on the setting with jointly unknown prior and bias and for sharing important intuition that informed Section F. We also thank the anonymous reviewers for their detailed feedback and discussions. We have no competing interests to disclose.

References

[1] Robert J Aumann, Michael Maschler, and Richard E Stearns. *Repeated games with incomplete information*. MIT press, 1995.

[2] Cuimin Ba, J Aislinn Bohren, and Alex Imas. Over-and underreaction to information: Belief updating with cognitive constraints. Technical report, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2025.

[3] Francesco Bacchiocchi, Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Markov persuasion processes: Learning to persuade from scratch. *arXiv preprint arXiv:2402.03077*, 2024.

[4] Phoebe E Bailey, Tarren Leon, Natalie C Ebner, Ahmed A Moustafa, and Gabrielle Weidemann. A meta-analysis of the weight of advice in decision-making. *Current Psychology*, 42(28): 24516–24541, 2023.

[5] Maya Bar-Hillel. The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3): 211–233, 1980.

[6] Daniel J Benjamin. Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1*, 2:69–186, 2019.

[7] Dirk Bergemann and Stephen Morris. Information design, bayesian persuasion, and bayes correlated equilibrium. *American Economic Review*, 106(5):586–591, 2016.

[8] David Blackwell. Equivalent comparisons of experiments. *The annals of mathematical statistics*, pages 265–272, 1953.

[9] Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Online bayesian persuasion. *Advances in neural information processing systems*, 33:16188–16198, 2020.

[10] Matteo Castiglioni, Alberto Marchesi, Andrea Celli, and Nicola Gatti. Multi-receiver online bayesian persuasion. In *International Conference on Machine Learning*, pages 1314–1323. PMLR, 2021.

[11] Yiling Chen, Tao Lin, Ariel D Procaccia, Aaditya Ramdas, and Itai Shapira. Bias detection via signaling. *Advances in Neural Information Processing Systems*, 37:69120–69143, 2024.

[12] Sarah Cooney, Kai Wang, Elizabeth Bondi, Thanh Nguyen, Phebe Vayanos, Hailey Winetrobe, Edward A Cranford, Cleotilde Gonzalez, Christian Lebiere, and Milind Tambe. Learning to signal in the goldilocks zone: Improving adversary compliance in security games. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 725–740. Springer, 2019.

[13] Edward A Cranford, Cleotilde Gonzalez, Palvi Aggarwal, Sarah Cooney, Milind Tambe, and Christian Lebiere. Toward personalized deceptive signaling for cyber defense using cognitive models. *Topics in Cognitive Science*, 12(3):992–1011, 2020.

[14] Ishita Dasgupta, Eric Schulz, Joshua B Tenenbaum, and Samuel J Gershman. A theory of learning to infer. *Psychological review*, 127(3):412, 2020.

[15] Geoffroy De Clippel and Xu Zhang. Non-bayesian persuasion. *Journal of Political Economy*, 130(10):2594–2642, 2022.

[16] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1):114, 2015.

[17] Shaddin Dughmi and Haifeng Xu. Algorithmic bayesian persuasion. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 412–425, 2016.

[18] Shaddin Dughmi and Haifeng Xu. Algorithmic persuasion with no externalities. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 351–368, 2017.

[19] Ward Edwards. Conservatism in human information processing. *Formal representation of human judgment*, 1968.

[20] Larry G Epstein, Jawwad Noor, Alvaro Sandroni, et al. Non-bayesian learning. *The BE Journal of Theoretical Economics*, 10(1):1–20, 2010.

[21] Yiding Feng, Wei Tang, and Haifeng Xu. Online bayesian recommendation with no regret. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 818–819, 2022.

[22] Yiding Feng, Chien-Ju Ho, and Wei Tang. Rationality-robust information design: Bayesian persuasion under quantal response. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 501–546. SIAM, 2024.

[23] Matthew Gentzkow and Emir Kamenica. A rothschild-stiglitz approach to bayesian persuasion. *American Economic Review*, 106(5):597–601, 2016.

[24] David M Grether. Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly journal of economics*, 95(3):537–557, 1980.

[25] David Hagmann and George Loewenstein. Persuasion with motivated beliefs. In *Opinion Dynamics & Collective Decisions Workshop*, 2017.

[26] Keegan Harris, Nicole Immorlica, Brendan Lucier, and Aleksandrs Slivkins. Algorithmic persuasion through simulation. *arXiv preprint arXiv:2311.18138*, 2023.

[27] Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

[28] Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605. IEEE, 2003.

[29] Masanori Kobayashi. Dynamic non-bayesian persuasion. *arXiv preprint arXiv:2508.12328*, 2025.

[30] Ce Li and Tao Lin. Information design with unknown prior. *arXiv preprint arXiv:2410.05533*, 2024.

[31] Hasan Mahmud, AKM Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander. What influences algorithmic decision-making? a systematic literature review on algorithm aversion. *Technological forecasting and social change*, 175:121390, 2022.

[32] Luis Rayo and Ilya Segal. Optimal information disclosure. *Journal of political Economy*, 118 (5):949–987, 2010.

[33] Charles S Taber and Milton Lodge. Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3):755–769, 2006.

[34] Wei Tang and Chien-Ju Ho. On the bayesian rational assumption in information design. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pages 120–130, 2021.

[35] Ben M Tappin, Gordon Pennycook, and David G Rand. Bayesian or biased? analytic thinking and political belief updating. *Cognition*, 204:104375, 2020.

[36] Jibang Wu, Zixuan Zhang, Zhe Feng, Zhaoran Wang, Zhuoran Yang, Michael I Jordan, and Haifeng Xu. Sequential information design: Markov persuasion process and its efficient reinforcement learning. *arXiv preprint arXiv:2202.10678*, 2022.

[37] Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pages 222–227. IEEE Computer Society, 1977.

[38] You Zu, Krishnamurthy Iyer, and Haifeng Xu. Learning to persuade on the fly: Robustness against ignorance. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 927–928, 2021.

A Related Work (Detailed Discussion)

A large body of empirical and theoretical work shows that Bayesian updating is a useful normative benchmark but often fails as a descriptive model of human responses to information. Early evidence on conservative updating finds that people revise beliefs in the Bayesian direction but insufficiently, suggesting that new information is only partially incorporated into posterior beliefs [19]. Related experimental work documents distorted uses of priors and likelihoods, including base-rate neglect [5], representativeness-based inference [24], and broader errors in probabilistic reasoning [6]. More recent work develops tests of belief updating and documents systematic over- and underreaction to information [2]. A separate line of work, mostly in political belief updating, studies motivated reasoning and belief polarization, showing that belief revisions may depend on prior commitments rather than only on statistical evidence [33, 35]. Complementing these findings, recent work in security games explicitly relaxes the assumption that adversaries interpret signals as perfectly Bayesian decision makers: [12] combines machine learning with cognitive models to account for boundedly rational responses to warnings, and [13] uses Instance-Based Learning Theory to model attackers who rely on past experience when interpreting deceptive signals. Together, these studies motivate a tractable model of receivers whose responses to information may be systematically attenuated or distorted rather than fully Bayesian.

Classical Bayesian persuasion studies how an informed sender strategically discloses information to influence a Bayesian receiver's action. The posterior-based formulation used in our paper follows the comparison-of-experiments tradition of Blackwell [8] and the splitting/concavification ideas developed in repeated games with incomplete information [1]. Kamenica and Gentzkow [27] formulate the canonical Bayesian-persuasion model, where the sender chooses a Bayes-plausible distribution over posteriors. Related early disclosure models include optimal information disclosure [32]. Subsequent work gives alternative characterizations and broader information-design formulations: Gentzkow and Kamenica [23] use a Rothschild–Stiglitz perspective, and Bergemann and Morris [7] connect information design to Bayes correlated equilibrium. A parallel algorithmic literature studies the computational complexity of finding optimal schemes when the primitives are known [17, 18].

Our work builds on the literature that relaxes this Bayesian-rationality assumption and is closest to work on persuasion with biased or boundedly rational receivers. [15, 29] study static and dynamic persuasion problems in which the receiver follows known non-Bayesian updating rules and the sender designs optimal information accordingly. Relatedly, [34] provides experimental evidence that behavior in information-design environments can deviate from the Bayesian benchmark and be approximated by linear updating, while [22] studies robust information design when receivers follow quantal-response or approximately rational behavior. However, papers above either take the receiver's non-Bayesian rule or bias level as known when designing signals, or focus on a static design problem. The main exception is [11], which also considers a linear belief distortion, but the objective is diagnostic: to infer whether the receiver's bias is above or below a threshold. In contrast, our sender must learn the unknown bias while simultaneously maximizing persuasion utility.

Our work also relates to the growing literature on learning in information design. One strand studies online Bayesian persuasion when the sender does not know the receiver's utility function or type, and designs no-regret signaling policies relative to a benchmark that knows these primitives [9, 10, 21, 26]. Another strand focuses on uncertainty about the receiver's prior or the state distribution [30, 38]. Related work also studies Markovian information-design problems in which the sender learns transition or event probabilities while maintaining persuasiveness [3, 36]. These papers study learning about utility, type, prior, state distribution, or dynamic environment primitives, typically while maintaining Bayesian receiver behavior. By contrast, in our model, the prior and utility functions are known, but the receiver's responsiveness to information is unknown.

B Missing Details in Section 3

B.1 Constants and Notation

Fix a binary instance $\mathcal{B} = (\mu_0, q_{\mathrm{th}})$, where $\mu_0 \in (0,1)$ and $q_{\mathrm{th}} \in (\mu_0, 1)$. Recall

$$\alpha_{\min}(\mathcal{B}) := \frac{q_{\mathrm{th}} - \mu_0}{1 - \mu_0}, \qquad \nu_{\mathcal{B}}(\alpha) := \mu_0 + \frac{q_{\mathrm{th}} - \mu_0}{\alpha}, \qquad W_{\mathcal{B}}(\alpha) := \frac{\mu_0}{\nu_{\mathcal{B}}(\alpha)}.$$

For every $\alpha \in [\alpha_{\min}(\mathcal{B}), 1]$, we have

$$\nu_{\mathcal{B}}(\alpha) \in [q_{\text{th}}, 1].$$

Define the two Lipschitz constants

$$L_\nu(\mathcal{B}) := \sup_{\alpha \in [\alpha_{\min}(\mathcal{B}),1]} \left| \frac{d}{d\alpha} \nu_{\mathcal{B}}(\alpha) \right| = \frac{q_{\text{th}} - \mu_0}{\alpha_{\min}(\mathcal{B})^2} = \frac{(1-\mu_0)^2}{q_{\text{th}} - \mu_0},$$

and

$$L_W(\mathcal{B}) := \sup_{\alpha \in [\alpha_{\min}(\mathcal{B}),1]} \left| \frac{d}{d\alpha} W_{\mathcal{B}}(\alpha) \right| = \frac{\mu_0(1-\mu_0)^2}{q_{\text{th}} - \mu_0}.$$

The constants used in Propositions 1 and 2 and theorem 1 can be taken as

$$C_{\text{BS}}(\mathcal{B}) := \big(1 - \alpha_{\min}(\mathcal{B})\big) \left(\frac{L_\nu(\mathcal{B})}{q_{\text{th}}} + L_W(\mathcal{B}) \right),$$

$$C_{\text{SE}}(\mathcal{B}) := 2\frac{L_\nu(\mathcal{B})}{q_{\text{th}}} + 1,$$

and

$$c_{\text{LB}}(\mathcal{B}) := \frac{1}{3} \min \left\{ \frac{\mu_0^2}{2}, \frac{1 - q_{\text{th}}}{16} \right\}.$$

All these constants depend only on $\mathcal{B}$ and are independent of T and $\alpha^\star$.

B.2 Missing Details in Binary Search

Recall that $\alpha_{\min}(\mathcal{B}) = (q_{\text{th}} - \mu_0)/(1 - \mu_0)$, and for any $\alpha \in [\alpha_{\min}(\mathcal{B}), 1]$

$$\tau(\alpha) = \left\{ \left(1 - \frac{\mu_0}{\nu_{\mathcal{B}}(\alpha)}, 0\right), \left(\frac{\mu_0}{\nu_{\mathcal{B}}(\alpha)}, \nu_{\mathcal{B}}(\alpha)\right) \right\}.$$

Algorithm 4 Binary Search (BS)

Require: Horizon T, binary instance $\mathcal{B} = (\mu_0, q_{\text{th}})$

$(\underline{\alpha}_0, \overline{\alpha}_0) \leftarrow (\alpha_{\min}(\mathcal{B}), 1)$, $k \leftarrow 0$, $t \leftarrow 1$, $M \leftarrow \lceil 2 \log_2 T \rceil$

Phase 1: Exploration

while $k < M$ and $t \le T$ **do**

  $m \leftarrow (\underline{\alpha}_k + \overline{\alpha}_k)/2$

  commit to and play $\tau(m)$

  if the realized posterior is not 0 **then**

    if the receiver chooses action 1 **then**

      $(\underline{\alpha}_{k+1}, \overline{\alpha}_{k+1}) \leftarrow (m, \overline{\alpha}_k)$

    else

      $(\underline{\alpha}_{k+1}, \overline{\alpha}_{k+1}) \leftarrow (\underline{\alpha}_k, m)$

    $k \leftarrow k + 1$

  $t \leftarrow t + 1$

Phase 2: Commitment

Commit to $\tau(\underline{\alpha}_k)$ for all remaining rounds.

Proof of Proposition 1. If the horizon ends before the algorithm collects M informative realizations, then the final epoch is truncated; this can only decrease the regret. It therefore suffices to analyze the hypothetical process in which the first M informative epochs are completed, and the policy then commits.

Let

$$\Delta_k := \overline{\alpha}_k - \underline{\alpha}_k.$$

By construction,

$$\Delta_k = (1 - \alpha_{\min}(\mathcal{B}))2^{-k}, \qquad k = 0, 1, \ldots, M.$$

For each $k = 1, \dots, M$, define

$$m_k := \frac{\underline{\alpha}_{k-1} + \overline{\alpha}_{k-1}}{2}.$$

Epoch k consists of all rounds in which the algorithm keeps probing m_k before the k-th informative realization occurs. In each such round, the high posterior $\nu_{\mathcal{B}}(m_k)$ appears with probability $\mu_0/\nu_{\mathcal{B}}(m_k)$, so the epoch length N_k is geometric with

$$\mathbb{E}[N_k] = \frac{\nu_{\mathcal{B}}(m_k)}{\mu_0}.$$

We first bound the exploration regret. Fix an epoch k.

If $m_k \le \alpha^\star$, then $\nu_{\mathcal{B}}(m_k) \ge \nu^\star$, so the high posterior induces action 1. Along every sample path of this epoch, the sender obtains utility 0 in the first $N_k - 1$ rounds and utility 1 in the last round, while the benchmark obtains $\mu_0/\nu^\star$ in every round. Hence the total regret of epoch k is

$$N_k \frac{\mu_0}{\nu^\star} - 1.$$

Taking expectations gives

$$\mathbb{E}[\text{regret in epoch } k] = \mathbb{E}[N_k]\frac{\mu_0}{\nu^\star} - 1 = \frac{\nu_{\mathcal{B}}(m_k)}{\nu^\star} - 1 = \frac{\nu_{\mathcal{B}}(m_k) - \nu^\star}{\nu^\star}.$$

Moreover,

$$\left|\frac{d}{d\alpha}\nu_{\mathcal{B}}(\alpha)\right| = \frac{q_{\text{th}} - \mu_0}{\alpha^2},$$

so $\nu_{\mathcal{B}}(\alpha)$ is $L_\nu(\mathcal{B})$-Lipschitz on $[\alpha_{\min}(\mathcal{B}), 1]$ with

$$L_\nu(\mathcal{B}) := \frac{q_{\text{th}} - \mu_0}{\alpha_{\min}(\mathcal{B})^2}.$$

Since m_k is the midpoint of $[\underline{\alpha}_{k-1}, \overline{\alpha}_{k-1}]$,

$$|m_k - \alpha^\star| \le \frac{\Delta_{k-1}}{2},$$

and therefore

$$\frac{\nu_{\mathcal{B}}(m_k) - \nu^\star}{\nu^\star} \le \frac{L_\nu(\mathcal{B})|m_k - \alpha^\star|}{\nu^\star} \le \frac{L_\nu(\mathcal{B})\Delta_{k-1}}{2\nu^\star}.$$

Summing over all such epochs gives

$$\sum_{k:\, m_k \le \alpha^\star} \mathbb{E}[\text{regret in epoch } k] \le \frac{L_\nu(\mathcal{B})}{2\nu^\star}\sum_{k=1}^{\infty}\Delta_{k-1} = \frac{L_\nu(\mathcal{B})(1 - \alpha_{\min}(\mathcal{B}))}{\nu^\star}.$$

If $m_k > \alpha^\star$, then $\nu_{\mathcal{B}}(m_k) < \nu^\star$, so even the high posterior induces action 0. Along every sample path of this epoch, the sender's utility is 0 in all N_k rounds, while the benchmark still obtains $\mu_0/\nu^\star$ per round. Hence the total regret of epoch k is

$$N_k \frac{\mu_0}{\nu^\star}.$$

Taking expectations gives

$$\mathbb{E}[\text{regret in epoch } k] = \mathbb{E}[N_k]\frac{\mu_0}{\nu^\star} = \frac{\nu_{\mathcal{B}}(m_k)}{\nu^\star} \le 1,$$

where the last inequality uses $\nu_{\mathcal{B}}(m_k) \le \nu^\star$. There are at most M such epochs. Hence

$$\text{Reg}_T^{\text{explore}} \le M + \frac{L_\nu(\mathcal{B})(1 - \alpha_{\min}(\mathcal{B}))}{\nu^\star}.$$

We next bound the regret after commitment. The policy commits to $\tau(\underline{\alpha}_M)$. Since $\underline{\alpha}_M \le \alpha^\star$, this scheme is conservative and therefore always persuasive. Also,

$$0 \le \alpha^\star - \underline{\alpha}_M \le \Delta_M = (1 - \alpha_{\min}(\mathcal{B}))2^{-M}.$$

Define

$$W(\alpha) := \frac{\mu_0}{\nu_{\mathcal{B}}(\alpha)} = \frac{\mu_0 \alpha}{q_{\mathrm{th}} - \mu_0 + \mu_0 \alpha}.$$

Then the benchmark utility is $W(\alpha^\star)$, while the committed scheme yields $W(\underline{\alpha}_M)$. Since

$$W'(\alpha) = \frac{\mu_0(q_{\mathrm{th}} - \mu_0)}{(q_{\mathrm{th}} - \mu_0 + \mu_0 \alpha)^2},$$

we have the uniform bound

$$|W'(\alpha)| \le L_W(\mathcal{B}) := \frac{\mu_0(1-\mu_0)^2}{q_{\mathrm{th}} - \mu_0}, \qquad \alpha \in [\alpha_{\min}(\mathcal{B}), 1].$$

Therefore the total regret in the commitment phase is at most

$$T\big(W(\alpha^\star) - W(\underline{\alpha}_M)\big) \le T L_W(\mathcal{B})(\alpha^\star - \underline{\alpha}_M) \le T L_W(\mathcal{B})(1-\alpha_{\min}(\mathcal{B}))2^{-M} \le L_W(\mathcal{B})(1-\alpha_{\min}(\mathcal{B})),$$

where the last step uses $T2^{-M} \le 1$.

Combining the exploration and commitment bounds yields

$$\mathrm{Reg}_T(\Pi^{\mathrm{BS}}; \alpha^\star) \le \lceil 2 \log_2 T \rceil + \frac{L_\nu(\mathcal{B})(1-\alpha_{\min}(\mathcal{B}))}{\nu^\star} + L_W(\mathcal{B})(1-\alpha_{\min}(\mathcal{B})).$$

□

B.3 Proof of Proposition 2

For a probe value $m \in [\alpha_{\min}(\mathcal{B}), 1]$, the algorithm uses the two-point scheme supported on $\{0, \nu_{\mathcal{B}}(m)\}$:

$$\tau(m) = \left\{ \left(1 - \frac{\mu_0}{\nu_{\mathcal{B}}(m)},\, 0\right), \left(\frac{\mu_0}{\nu_{\mathcal{B}}(m)},\, \nu_{\mathcal{B}}(m)\right) \right\}.$$

The receiver takes action 1 if and only if

$$(1-\alpha^\star)\mu_0 + \alpha^\star \nu_{\mathcal{B}}(m) \ge q_{\mathrm{th}} \iff m \le \alpha^\star.$$

Hence the sender's one-round regret under probe m is

$$r(m) = \begin{cases} \dfrac{\mu_0}{\nu^\star} - \dfrac{\mu_0}{\nu_{\mathcal{B}}(m)}, & m \le \alpha^\star, \\ \dfrac{\mu_0}{\nu^\star}, & m > \alpha^\star. \end{cases}$$

The algorithm repeats the same probe until the first High signal appears, so each probe corresponds to one epoch. Let $N(m)$ be the number of physical rounds spent in the epoch for probe m. Then $N(m) \sim \mathrm{Geom}(p_H(m))$ and

$$\mathbb{E}[N(m)] = \frac{1}{p_H(m)} = \frac{\nu_{\mathcal{B}}(m)}{\mu_0}.$$

Since the probe m is repeated until the first High signal appears, the total regret of one epoch is

$$N(m)\frac{\mu_0}{\nu^\star} - \mathbf{1}\{m \le \alpha^\star\}.$$

Therefore,

$$R_{\mathrm{ep}}(m) = \mathbb{E}\Big[N(m)\frac{\mu_0}{\nu^\star} - \mathbf{1}\{m \le \alpha^\star\}\Big],$$

which gives

$$R_{\mathrm{ep}}(m) = \begin{cases} \dfrac{\nu_{\mathcal{B}}(m) - \nu^\star}{\nu^\star}, & m \le \alpha^\star, \\ \dfrac{\nu_{\mathcal{B}}(m)}{\nu^\star}, & m > \alpha^\star. \end{cases}$$

In particular, if $m > \alpha^\star$, then $\nu_{\mathcal{B}}(m) \le \nu^\star$ and hence

$$R_{\mathrm{ep}}(m) \le 1.$$

Next define

$$C(\mathcal{B}) := \sup_{x\in[\alpha_{\min}(\mathcal{B}),1]} \left|\frac{d}{dx}\nu_{\mathcal{B}}(x)\right| = \sup_{x\in[\alpha_{\min}(\mathcal{B}),1]} \frac{q_{\text{th}}-\mu_0}{x^2} = \frac{q_{\text{th}}-\mu_0}{\alpha_{\min}(\mathcal{B})^2}.$$

For every $m \le \alpha^\star$,

$$0 \le \nu_{\mathcal{B}}(m) - \nu^\star \le C(\mathcal{B})(\alpha^\star - m),$$

so

$$R_{\text{ep}}(m) \le \frac{C(\mathcal{B})}{\nu^\star}(\alpha^\star - m).$$

We now bound the exploration regret phase by phase. In a given phase, let the current interval be $[a, b]$ with length $L := b - a$, and let the current step size be ϵ. The algorithm probes the points

$$m_j = a + j\epsilon, \qquad j = 1, \dots, J,$$

,then either stops at the first point m_J with high signal inducing action 0 or exits the interval and updates to $[m_{J-1}, b]$. In both cases,

$$\alpha^\star - m_j \le (J-j)\epsilon, \qquad J \le \frac{L}{\epsilon} + 1.$$

There is at most one epoch with high posterior inducing action 0 in each phase, and its regret is at most 1. Therefore the total regret except the above special epoch in this phase is bounded by

$$\begin{aligned}\sum_{j=1}^{J-1} R_{\text{ep}}(m_j) &\le \frac{C(\mathcal{B})}{\nu^\star}\sum_{j=1}^{J-1}(\alpha^\star - m_j) \\ &\le \frac{C(\mathcal{B})}{\nu^\star}\,\epsilon\sum_{r=1}^{J-1} r \le \frac{C(\mathcal{B})}{\nu^\star}\left(\frac{L^2}{2\epsilon} + \frac{L}{2}\right).\end{aligned}$$

It remains to count phases. Let $\epsilon_0 = 1/2$ be the initial step size. After each phase, the new interval length is at most the previous step size, and the next step size is squared. Thus if L_r and ϵ_r denote the interval length and step size at the beginning of phase r, then

$$L_{r+1} \le \epsilon_r, \qquad \epsilon_{r+1} = \epsilon_r^2.$$

Hence

$$\epsilon_r = 2^{-2^r}, \qquad L_r \le \epsilon_{r-1} \le 1 \quad (r \ge 1),$$

and in particular $\epsilon_r \ge L_r^2$ for every $r \ge 1$. Therefore every phase after the first incurs regret at most $C(\mathcal{B})/\nu^\star + 1$. The first phase also has $L_0 \le 1$ and $\epsilon_0 = 1/2$, so it contributes the same order.

Exploration terminates once the interval length becomes at most $1/T$, which is implied by $\epsilon_r \le 1/T$. Since $\epsilon_r = 2^{-2^r}$, the number of phases is at most

$$P \le 1 + \lceil \log_2 \log_2 T \rceil.$$

Therefore the total exploration regret is

$$O\left(\left(\frac{C(\mathcal{B})}{\nu^\star} + 1\right)\log\log T\right).$$

In the commitment stage, the algorithm commits to the left endpoint $\widehat{\alpha}$ of the final interval. Because $\widehat{\alpha} \le \alpha^\star$ and $|\alpha^\star - \widehat{\alpha}| \le 1/T$, its one-round regret is

$$\frac{\mu_0}{\nu^\star} - \frac{\mu_0}{\nu_{\mathcal{B}}(\widehat{\alpha})} \le \frac{\nu_{\mathcal{B}}(\widehat{\alpha}) - \nu^\star}{\nu^\star} \le \frac{C}{\nu^\star}\,|\alpha^\star - \widehat{\alpha}| \le \frac{C}{T\nu^\star}.$$

Summing over the remaining T rounds gives commitment regret $O(C(\mathcal{B})/\nu^\star)$. Combining the exploration and commitment parts proves

$$\text{Reg}_T\big(\Pi^{\text{SE}}; \alpha^\star\big) = O\left(\left(\frac{2C(\mathcal{B})}{\nu^\star} + 1\right)\log\log T\right).$$

Finally, since

$$\nu^\star = \frac{q_{\text{th}} - (1-\alpha^\star)\mu_0}{\alpha^\star}, \qquad C(\mathcal{B}) = \frac{(1-\mu_0)^2}{q_{\text{th}} - \mu_0},$$

we obtain

$$\frac{C(\mathcal{B})}{\nu^\star} \le \frac{C(\mathcal{B})}{q_{\text{th}}} = \frac{(1-\mu_0)^2}{(q_{\text{th}} - \mu_0)q_{\text{th}}}.$$

B.4 Proof of Theorem 1

As discussed above, if $\alpha^\star < \alpha_{\min}(\mathcal{B})$ then persuasion is impossible and the regret is identically zero. Hence it suffices to prove the lower bound on the subfamily $\alpha^\star \in [\alpha_{\min}(\mathcal{B}), 1]$. Under the change of variables

$$\nu^\star = \mu_0 + \frac{q_{\text{th}} - \mu_0}{\alpha^\star},$$

this is equivalent to proving a lower bound for an unknown threshold $\nu^\star \in [q_{\text{th}}, 1]$.

By Yao's principle [37], it is enough to exhibit a distribution over $\nu^\star$ such that every deterministic strategy suffers expected regret $\Omega(\log \log T)$. We choose $\nu^\star$ uniformly on $[q_{\text{th}}, 1]$.

It is without loss of generality to restrict attention to two-point schemes supported on $\{0, \nu\}$ with $\nu \in [q_{\text{th}}, 1]$. Posteriors below μ_0 can be replaced by 0. This modification strictly improves performance, as posteriors below μ_0 yield zero utility. Moreover, shifting these values to 0 minimizes the probability mass required for the lower signals, thereby allowing for a higher probability on the utility-generating posteriors. For posteriors above μ_0, denoted by $\overline{\nu}_1 < \overline{\nu}_2 < \cdots$, they can be decomposed into a convex combination of binary signal schemes $\{0, \overline{\nu}_1\}, \{0, \overline{\nu}_2\}, \ldots$. Since we restrict our attention to deterministic algorithms, it suffices to consider a binary signal scheme with support $\{0, \nu\}$, which induces the useful posterior ν with probability $\frac{\mu_0}{\nu}$. We use High signal to denote this useful posterior ν in each signal scheme and Low signal to denote the posterior 0.

We now work in *event time*. Let $\tau_1 < \tau_2 < \cdots$ be the physical rounds in which a High signal is realized, and let

$$N := \max\{k : \tau_k \le T\}$$

be the total number of High signals by time T. Under a deterministic strategy, the history of High realizations only reveals comparisons of $\nu^\star$ with queried values, so conditional on the current history the posterior distribution of $\nu^\star$ remains uniform on some feasible interval $I_k = [a_k, b_k]$. Within each phase, the strategy may therefore be viewed as choosing a deterministic descending sequence of probes until the first High signal that inducing action 0 arrives.

Following Kleinberg and Leighton [28], define phase $k \ge 0$ to end after either

$$M_k := 2^{2^k} - 1$$

High signals inducing action 1 or the first High signal inducing action 0, whichever comes first. Let $I_k = [a_k, b_k]$ denote the feasible interval at the beginning of phase k, and let

$$E_k := \left\{ |I_k| \ge \frac{1 - q_{\text{th}}}{2}\, 2^{-2^k} \right\}.$$

The same feasible-interval argument as in Claims 2.3–2.4 of Kleinberg and Leighton [28] gives

$$\mathbb{P}(E_k) \ge \frac{1}{2}.$$

We will show that every phase contributes a constant expected regret.

Fix a phase k and condition on E_k. Let

$$c_k := \frac{a_k + b_k}{2}$$

be the midpoint of the feasible interval.

Case 1: Some probe $\nu < c_k$ is used during phase k. Since $\nu \le 1$, the corresponding High signal occurs with probability at least μ_0. Also, conditional on I_k, the true threshold is uniform on I_k, so with probability $1/2$ we have $\nu^\star \ge c_k > \nu$. Whenever both events occur, the realized High signal induces action 0 and the sender incurs instantaneous regret at least

$$\frac{\mu_0}{\nu^\star} \ge \mu_0.$$

Therefore, conditional on E_k, the expected regret in this phase is at least

$$\frac{\mu_0}{2} \cdot \mu_0 = \frac{\mu_0^2}{2}.$$

Case 2: Every probe used during phase k satisfies $\nu \geq c_k$. Conditional on $\nu^\star \leq c_k$ (which has probability $1/2$ given I_k is uniform), every realized High signal in this phase induces action 1. Hence the phase ends only when it accumulates exactly M_k High signals.

Let S_k be the number of physical rounds in phase k, and for each round t in this phase let $H_t \in \{0, 1\}$ be the indicator of a High realization. If $\mathcal{F}_t$ is the natural filtration, then conditional on $\mathcal{F}_{t-1}$ the strategy's choice ν_t is fixed and

$$\mathbb{E}[H_t \mid \mathcal{F}_{t-1}] = \frac{\mu_0}{\nu_t} \leq \frac{\mu_0}{c_k}.$$

Thus $Z_t := H_t - \mathbb{E}[H_t \mid \mathcal{F}_{t-1}]$ is a bounded martingale difference sequence, and optional stopping gives

$$M_k = \mathbb{E}\left[\sum_{t=1}^{S_k} H_t\right] = \mathbb{E}\left[\sum_{t=1}^{S_k} \frac{\mu_0}{\nu_t}\right] \leq \frac{\mu_0}{c_k}\,\mathbb{E}[S_k].$$

Hence

$$\mathbb{E}[S_k] \geq M_k \frac{c_k}{\mu_0}.$$

On every physical round of this phase, because $\nu_t \geq c_k$, the sender's one-round regret is at least

$$\frac{\mu_0}{\nu^\star} - \frac{\mu_0}{c_k} = \frac{\mu_0(c_k - \nu^\star)}{c_k \nu^\star}.$$

Multiplying by the lower bound on $\mathbb{E}[S_k]$ yields

$$\mathbb{E}[\text{Regret in phase } k \mid \nu^\star \leq c_k, E_k] \geq M_k\,\mathbb{E}\left[\frac{c_k - \nu^\star}{\nu^\star} \mid \nu^\star \leq c_k, E_k\right] \geq M_k\,\mathbb{E}[c_k - \nu^\star \mid \nu^\star \leq c_k, E_k],$$

where the last inequality uses $\nu^\star \leq 1$. Since $\nu^\star$ is uniform on $[a_k, c_k]$ under the conditioning,

$$\mathbb{E}[c_k - \nu^\star \mid \nu^\star \leq c_k, E_k] = \frac{|I_k|}{4}.$$

Therefore,

$$\mathbb{E}[\text{Regret in phase } k \mid E_k] \geq \frac{1}{2}\,M_k \frac{|I_k|}{4} \geq \frac{1 - q_{\text{th}}}{16}\,(1 - 2^{-2^k}) \geq \frac{1 - q_{\text{th}}}{16} \cdot \frac{1}{2},$$

so in particular the phase contributes at least a positive constant.

Combining the two cases, there exists a universal constant

$$c_0(\mathcal{B}) := \frac{1}{2} \min\left\{\frac{\mu_0^2}{2}, \frac{1 - q_{\text{th}}}{16}\right\} > 0$$

such that every phase contributes at least $c_0(\mathcal{B})$ expected regret. Since the number of completed phases is $\Omega(\log \log N)$, we obtain

$$\text{Reg}_T \geq c_0(\mathcal{B})\,\Omega(\log \log N).$$

It remains to relate event time to physical time. For each physical round t, the chosen probe satisfies $\nu_t \leq 1$, hence the High realization probability obeys

$$\mathbb{E}[H_t \mid \mathcal{F}_{t-1}] = \frac{\mu_0}{\nu_t} \geq \mu_0.$$

Therefore $X_t := H_t - \mathbb{E}[H_t \mid \mathcal{F}_{t-1}]$ is a bounded martingale difference sequence, and Azuma–Hoeffding implies

$$\mathbb{P}\left(N \leq \frac{\mu_0 T}{2}\right) \leq \exp\left(-\frac{\mu_0^2 T}{8}\right).$$

Thus with probability at least $1 - e^{-\mu_0^2 T/8}$ we have $N \geq \mu_0 T/2$, and consequently

$$\text{Reg}_T \geq \left(1 - e^{-\mu_0^2 T/8}\right) c_0(\mathcal{B}) \log_2 \log_2\left(\frac{\mu_0 T}{2}\right) = \Omega(\log \log T).$$

This completes the proof.

C Missing Details in Section 4

C.1 Constants and Notation

We collect the constants used in Section 4. All of them are independent of the horizon T.

Because states with zero prior probability are never reached, μ_0 can be viewed as full support on Ω without loss of generality. Then define

$$\delta_{\mu_0} = \min\left\{\operatorname{dist}\left(\mu_0, \partial R^1_{a_0}\right), \operatorname{dist}\left(\mu_0, \partial\Delta(\Omega)\right)\right\}.$$

Here ∂ denote the boundary of one region. Moreover, $R^1_{a_0}$ is the smallest default-action region over $\alpha \in (0,1]$, so δ_{μ_0} is a uniform lower bound on the distance from μ_0 to the nearest default-action region boundary. By Assumption 1, $\delta_{\mu_0} > 0$. Also, Define

$$U_{\max} = \max_{a\in A, \omega\in\Omega} |u_S(a,\omega)|, \qquad \Delta U_{\max} = 2U_{\max}.$$

Perturbation constants Fix an action a. Let $n := |\Omega|$ and $d := n-1$. Let

$$H_0 := \{x \in \mathbb{R}^n : \mathbf{1}^\top x = 0\},$$

and let $Q \in \mathbb{R}^{n\times d}$ have orthonormal columns spanning H_0. Fix any $\bar{\nu}$ with $\mathbf{1}^\top\bar{\nu} = 1$. Every vector ν with $\mathbf{1}^\top\nu = 1$ can be written uniquely as

$$\nu = \bar{\nu} + Qz, \qquad z \in \mathbb{R}^d,$$

and Euclidean distances are preserved:

$$\|\nu - \nu'\|_2 = \|z - z'\|_2.$$

Collect all inequality constraints defining R^J_a: the IC constraints $\Delta u^\top_{a,a'}\nu \geq b^J_{a,a'}$ and the simplex nonnegativity constraints $\nu(\omega) \geq 0$. Stack their normals as rows of a matrix $A_a \in \mathbb{R}^{m_a\times n}$ and write

$$R^J_a = \{\nu \in \mathbb{R}^n : \mathbf{1}^\top\nu = 1,\ A_a\nu \geq b_a(J)\}.$$

Define

$$\widetilde{A}_a := A_aQ, \qquad \widetilde{b}_a(J) := b_a(J) - A_a\bar{\nu},$$

and the reduced polytope

$$\widetilde{R}^J_a := \{z \in \mathbb{R}^d : \widetilde{A}_az \geq \widetilde{b}_a(J)\}.$$

Then $\nu = \bar{\nu} + Qz \in R^J_a$ if and only if $z \in \widetilde{R}^J_a$, and

$$\operatorname{dist}(\nu, R^J_a) = \operatorname{dist}(z, \widetilde{R}^J_a).$$

For $B \subseteq [m_a]$, let $\widetilde{A}_{a,B}$ be the submatrix of $\widetilde{A}_a$ with rows in B. Define

$$\mathcal{B}_a := \left\{B \subseteq [m_a] : 1 \leq |B| \leq d,\ \operatorname{rank}(\widetilde{A}_{a,B}) = |B|\right\}.$$

For $B \in \mathcal{B}_a$, define the right pseudoinverse

$$\widetilde{A}^\dagger_{a,B} := \widetilde{A}^\top_{a,B}\left(\widetilde{A}_{a,B}\widetilde{A}^\top_{a,B}\right)^{-1}.$$

Finally we can define

$$\kappa_a := \max_{B\in\mathcal{B}_a} \|\widetilde{A}^\dagger_{a,B}\|_2, \qquad \kappa := \max_{a\in A}\kappa_a.$$

Geometric constants We then define three new geometric constants:

$$G_{\max} = \max_{a\neq a'} |\Delta u^\top_{a,a'}\mu_0|, \qquad L_b := \max_{a\in A}\sup_{\alpha\in[\alpha_{\min}(\mathcal{I}),1]}\left\|\frac{d}{d\alpha}\widetilde{b}_a(\alpha)\right\|_2.$$

Both of them are finite and depends only on $\mathcal{I}$. For L_b, using $\alpha_{\min}(\mathcal{I}) \geq \delta_{\mu_0}/\sqrt{2}$, it can be upper bounded by an instance-dependent multiple of $G_{\max}/\delta^2_{\mu_0}$.

Regret constants The hidden constant in Proposition 8 is

$$C_{\mathrm{loc}} := \frac{\Delta U_{\max}}{p_{\min}} = \frac{\sqrt{2}\Delta U_{\max}}{\delta_{\mu_0}}.$$

The posterior-distance constant is

$$K_{\mathrm{dist}} := \kappa L_b.$$

The Bayes-plausibility repair constant from Lemma 4 is

$$C_{\mathrm{rep}} := U_{\max}\left(\sqrt{|\Omega|} + \frac{4}{\delta_{\mu_0}}\right).$$

We use

$$K_{\mathrm{safe}} := C_{\mathrm{rep}} K_{\mathrm{dist}}, \qquad K_{\mathrm{probe}} := C_{\mathrm{rep}} K_{\mathrm{dist}}$$

for Propositions 3 and 4, respectively.

Let c_{info} denote the constant c in the first branch of Proposition 5. One valid phase-level hidden constant is

$$C_{\mathrm{phase}} = \max\left\{\frac{2(K_{\mathrm{safe}} + K_{\mathrm{probe}} + \Delta U_{\max})}{c_{\mathrm{info}}}, 2K_{\mathrm{probe}} + \Delta U_{\max}\right\}.$$

Thus the general-case constant in Theorem 2 may be chosen as any finite constant satisfying

$$C_{\mathrm{GSE}}(\mathcal{I}, \alpha^\star) \geq C_{\mathrm{loc}} + C_{\mathrm{phase}} + 1.$$

C.2 Missing Details in the Localization Stage

Our main goal in the first stage is to conduct pure exploration and shrink the uncertainty interval as quickly as possible. This objective is aligned with Chen et al. [11], whose algorithm is designed to test a single threshold with minimum sample complexity. Therefore, we adopt the same threshold-testing construction, see Algorithm 5. It designs a signaling scheme for threshold β by solving a carefully designed LP, which constructs a direct-recommendation scheme whose non-default recommendations lie exactly on the receiver's decision boundary between non-default action and default action at β. Conditional on such a non-default recommendation being realized, the observed receiver action reveals whether $\alpha^\star \geq \beta$ or $\alpha^\star < \beta$. The LP objective maximizes the probability of these informative non-default recommendations, equivalently minimizing the expected waiting time until the comparison is revealed.

The threshold-test LP and algorithm. Fix a test point β. We use a direct-recommendation scheme $\pi(a \mid \omega)$ by solving the following LP:

$$\begin{aligned}
\text{Maximize} \quad & \sum_{a \in A \setminus \{a_0\}} \sum_{\omega \in \Omega} \pi(a \mid \omega)\, \mu_0(\omega) \\
\text{subject to:} \quad & \textbf{Optimality of } a \textbf{ over other actions: } \forall a \in A,\ \forall a' \in A \setminus \{a\} \\
& \sum_{\omega \in \Omega} \pi(a \mid \omega)\, \mu_0(\omega) \Big[\beta \Delta u_{a,a'}(\omega) + (1-\beta) \sum_{\omega' \in \Omega} \mu_0(\omega') \Delta u_{a,a'}(\omega')\Big] \geq 0; \\
& \textbf{Indifference between } a \textbf{ and } a_0\textbf{: } \forall a \in A \setminus \{a_0\} \\
& \sum_{\omega \in \Omega} \pi(a \mid \omega)\, \mu_0(\omega) \Big[\beta \Delta u_{a,a_0}(\omega) + (1-\beta) \sum_{\omega' \in \Omega} \mu_0(\omega') \Delta u_{a,a_0}(\omega')\Big] = 0; \\
& \textbf{Probability distribution constraints: } \forall \omega \in \Omega \\
& \sum_{a \in A} \pi(a \mid \omega) = 1 \quad \text{and} \quad \forall a \in A,\ \pi(a \mid \omega) \geq 0.
\end{aligned} \tag{2}$$

Algorithm 5 THRESHOLDTEST at candidate bias β

Require: Current interval $J = [\underline{\alpha}, \overline{\alpha}]$, test point $\beta \in J$, general instance $\mathcal{I} = (\Omega, A, \mu_0, u_S, u_R)$.

Solve the threshold-test LP at β, and let π^β be an optimal solution.
repeat
    Play π^β for one round.
    Observe the realized recommendation action a_{rec} and the receiver's actual action.
until $a_{\mathrm{rec}} \neq a_0$
if the receiver takes the default action a_0 under recommendation $a_{\mathrm{rec}} \neq a_0$ **then**
    return $[\underline{\alpha}, \beta]$.
else
    return $[\beta, \overline{\alpha}]$.

Note, however, that as in the binary case, not every candidate bias is testable. When the bias is very small, the receiver may remain too close to the prior for any signal to affect her decision. We establish below the existence of an instance-dependent detectable threshold $\alpha_{\min}(\mathcal{I})$. Since the absence of persuasion implies zero regret, and one additional test suffices to detect whether persuasion is possible, we focus on $\alpha^\star \in [\alpha_{\min}(\mathcal{I}), 1]$ in our analysis.

Proposition 7. *There exists an instance-dependent threshold $\alpha_{\min}(\mathcal{I}) \in [\delta_{\mu_0}/\sqrt{2}, 1]$ such that threshold testing is feasible if and only if $\beta \geq \alpha_{\min}(\mathcal{I})$. Moreover, for every feasible candidate bias, the testing scheme produces an informative realization with probability at least $\delta_{\mu_0}/\sqrt{2}$.*

Proof. For every non-default action $a \neq a_0$, define the indifference hyperplane in distorted-belief space

$$\mathcal{I}_a := \left\{ \widehat{\nu} \in \Delta(\Omega) : \Delta u_{a,a_0}^\top \widehat{\nu} = 0 \right\}.$$

For a candidate bias level $\beta \in [0, 1]$, define its pullback to Bayesian-posterior space:

$$\mathcal{I}_{a,\beta} := \{ \nu \in \Delta(\Omega) : (1-\beta)\mu_0 + \beta\nu \in \mathcal{I}_a \}.$$

From Theorem 5.2 in [11], we have that For a candidate $\beta \in [0, 1]$, threshold testing between $\alpha^\star \geq \beta$ and $\alpha^\star < \beta$ is feasible if and only if $\mathcal{I}_{a,\beta} \neq \emptyset$ for some $a \neq a_0$. Namely, a threshold test at level β exists exactly when some non-default action can be placed on an indifference boundary with the default action at the distorted belief generated by β.

We now show that the set of feasible thresholds is an upper interval: there exists $\alpha_{\min}(I) \in [\delta_{\mu_0}/\sqrt{2}, 1]$ such that β is testable if and only if $\beta \geq \alpha_{\min}(I)$.

Let

$$\mathcal{S} := \{ \beta \in [0, 1] : \exists a \neq a_0 \text{ such that } \mathcal{I}_{a,\beta} \neq \emptyset \}.$$

If $\beta \in \mathcal{S}$ and $\beta' \geq \beta$, choose $\nu \in \mathcal{I}_{a,\beta}$. Define

$$\nu' := \frac{\beta}{\beta'}\nu + \left(1 - \frac{\beta}{\beta'}\right)\mu_0.$$

Then $\nu' \in \Delta(\Omega)$, and

$$(1-\beta')\mu_0 + \beta'\nu' = (1-\beta)\mu_0 + \beta\nu \in \mathcal{I}_a.$$

Hence $\nu' \in \mathcal{I}_{a,\beta'}$, so $\beta' \in \mathcal{S}$. Therefore $\mathcal{S}$ is an upper interval, and we may set

$$\alpha_{\min}(I) := \inf \mathcal{S}.$$

Finally, suppose $\beta \in \mathcal{S}$. Then for some $a \neq a_0$ there exists $\nu \in \Delta(\Omega)$ such that

$$\widehat{\nu} := (1-\beta)\mu_0 + \beta\nu \in \mathcal{I}_a.$$

Since $\mathcal{I}_a$ is contained in the boundary of the Bayesian default-action region, the definition of δ_{μ_0} gives

$$\|\widehat{\nu} - \mu_0\|_2 \geq \delta_{\mu_0}.$$

On the other hand,

$$\|\widehat{\nu} - \mu_0\|_2 = \beta\|\nu - \mu_0\|_2 \leq \beta\sqrt{2},$$

because both ν and μ_0 lie in the simplex. Thus $\beta \geq \delta_{\mu_0}/\sqrt{2}$, and in particular $\alpha_{\min}(I) \geq \delta_{\mu_0}/\sqrt{2}$.

Now it remains to lower bound the useful-signal probability. Let p_β denote the optimal objective value of the threshold-test LP, i.e., the total probability of recommending a non-default action. Let ν_{use} be the probability-weighted average posterior conditional on a non-default recommendation, and let ν_{def} be the posterior conditional on the default recommendation. Bayes plausibility gives

$$\mu_0 = p_\beta \nu_{\text{use}} + (1 - p_\beta)\nu_{\text{def}}.$$

Equivalently,

$$p_\beta = \frac{\|\mu_0 - \nu_{\text{def}}\|_2}{\|\nu_{\text{use}} - \nu_{\text{def}}\|_2}.$$

For an optimal threshold-test scheme, the default posterior can be chosen on the boundary of the Bayesian default-action region. Otherwise, if ν_{def} were strictly inside that region, moving it farther away from μ_0 along the ray passing through ν_{def} would preserve default optimality and would strictly increase the feasible probability of non-default recommendations, contradicting optimality of the LP objective. Hence

$$\|\mu_0 - \nu_{\text{def}}\|_2 \geq \delta_{\mu_0}.$$

Since any two points in the simplex are at Euclidean distance at most $\sqrt{2}$,

$$p_\beta \geq \frac{\delta_{\mu_0}}{\sqrt{2}}.$$

Thus every feasible threshold-test scheme emits a useful non-default recommendation with probability at least $\delta_{\mu_0}/\sqrt{2}$. □

In this stage, our algorithm stops once the uncertainty interval satisfies $|J| = O(1/\log T)$, which requires $O(\log\log T)$ threshold tests. By Proposition 7, the probability of obtaining an informative realization is bounded below by a constant. Hence, each threshold test has constant sample complexity, and the total regret incurred in this stage is $O(\log\log T)$.

Proposition 8. *Given the target number of informative signals in the localization stage is $O(\log\log T)$, the expected regret during the localization stage is $O(\log\log T)$.*

Proof. The targeted number of informative signals are $N = O(\log\log T)$. Let X_k be the number of rounds between the $(k-1)$-st and k-th useful recommendation. By Proposition 7, each X_k is stochastically dominated by a geometric random variable with success probability $\delta_{\mu_0}/\sqrt{2}$. Therefore

$$\mathbb{E}[X_k] \leq \frac{\sqrt{2}}{\delta_{\mu_0}}, \qquad \mathbb{E}\left[\sum_{k=1}^{N} X_k\right] \leq \frac{\sqrt{2}N}{\delta_{\mu_0}} = O(\log\log T).$$

Each physical round contributes at most $\Delta U_{\max}$ regret. The hidden constant here is $\sqrt{2}\Delta U_{\max}/\delta_{\mu_0}$. □

C.3 Relevant Actions

Recall the relevant-action set

$$A_{\text{rel}} = \Big\{a \in A : \exists \nu \in \Delta(\Omega) \text{ that } a \text{ uniquely maximizes } \sum_{\omega\in\Omega} (\alpha^\star \nu(\omega) + (1-\alpha^\star)\mu_0(\omega)) u_R(a,\omega)\Big\}.$$

The following discussion shows that without excluding those actions, it is impossible to guarantee low regret.

Remark 1 (Why tie-only actions are excluded)**.** *Suppose $a^\circ \notin A_{\text{rel}}$. Then the set of distorted beliefs at which a° can be selected is contained in a finite union of indifference hyperplanes. Consequently, for any non-degenerate interval $J = [\underline{\alpha}, \overline{\alpha}]$, the set of Bayesian posteriors that induce a° for every $\alpha \in J$ has empty relative interior in $\Delta(\Omega)$. Thus any scheme that relies on such an action is not stable to arbitrarily small perturbations of the unknown bias parameter.*

This is not only a technical inconvenience. Consider a binary-state example in which a° is selected by tie-breaking only at the distorted belief $\widehat{\nu}(1) = 1/2$, and suppose the sender receives utility 1 from

a° and utility 0 from all other actions. For each known α, a full-information sender can target the unique Bayesian posterior

$$\nu_\alpha(1) = \frac{1/2 - (1-\alpha)\mu_0(1)}{\alpha}$$

whenever it is feasible. But if α is unknown, inducing a° requires hitting this posterior exactly. Any interval-safe learning rule fails to induce a° on a whole open interval of possible bias values, while the full-information benchmark obtains a positive value. Hence including tie-only actions can create linear regret against the full-information benchmark.

We also recall the definition of interval-safe action region within J and strict interval-safe region.

$$R_a^J = \bigcap_{\alpha \in J} R_a^\alpha = \Big\{\nu \in \Delta(\Omega) : \Delta u_{a,a'}^\top \nu \geq b_{a,a'}(J) = \max\left\{b_{a,a'}(\underline{\alpha}), b_{a,a'}(\overline{\alpha})\right\},\ \forall a' \neq a\Big\}.$$

$$\mathrm{int}_{\mathrm{IC}}\big(R_a^J\big) = \left\{\nu \in \Delta(\Omega) : \Delta u_{a,a'}^\top \nu > b_{a,a'}(J),\ \forall a' \neq a\right\}.$$

We further show that for sufficiently small intervals $J \ni \alpha^\star$, we can identify A_{rel} through checking whether $\mathrm{int}_{\mathrm{IC}}\big(R_a^J\big) \neq \emptyset$. After the localization stage returns an interval $J = [\underline{\alpha}, \overline{\alpha}]$ of length $O(1/\log T)$, for all sufficiently large T, J satisfies this length requirement.

Lemma 2 (Local equivalence of relevant actions)**.** *Fix the instance I and the true bias $\alpha^\star$. There exists $\epsilon_{\mathrm{rel}}(I, \alpha^\star) > 0$ such that for every interval $J \ni \alpha^\star$ with $|J| \leq \epsilon_{\mathrm{rel}}(I, \alpha^\star)$, and every action $a \in A$,*

$$a \in A_{\mathrm{rel}} \quad \Longleftrightarrow \quad \mathrm{int}_{\mathrm{IC}}\big(R_a^J\big) \neq \emptyset$$

Consequently, on sufficiently small localized intervals, the computable strict-feasibility check in Section 4.2.1 keeps exactly the relevant actions.

Proof. Define the single-bias maximal slack

$$s_a(\alpha^\star) := \max_{\nu \in \Delta(\Omega),\, s \in \mathbb{R}} s \quad \text{s.t.} \quad \Delta u_{a,a'}^\top \nu \geq b_{a,a'}(\alpha^\star) + s,\ \forall a' \neq a.$$

By definition, $a \in A_{\mathrm{rel}}$ if and only if $s_a(\alpha^\star) > 0$.

First suppose $\mathrm{int}_{\mathrm{IC}}\big(R_a^J\big) \neq \emptyset$. Then there exists $\nu \in \Delta(\Omega)$ such that

$$\Delta u_{a,a'}^\top \nu > b_{a,a'}(J), \qquad \forall a' \neq a.$$

Since $J \ni \alpha^\star$, we have

$$b_{a,a'}(J) \geq b_{a,a'}(\alpha^\star)$$

for every $a' \neq a$. Hence the same posterior has strict IC slack at $\alpha^\star$, so $s_a(\alpha^\star) > 0$, and therefore $a \in A_{\mathrm{rel}}$.

Conversely, suppose $a \in A_{\mathrm{rel}}$. Then $s_a(\alpha^\star) > 0$. Let

$$\bar{s}_a := s_a(\alpha^\star) > 0.$$

Choose $\nu_a \in \Delta(\Omega)$ such that

$$\Delta u_{a,a'}^\top \nu_a \geq b_{a,a'}(\alpha^\star) + \bar{s}_a/2, \qquad \forall a' \neq a.$$

The functions $b_{a,a'}(\alpha)$ are Lipschitz on $[\alpha_{\min}(I), 1]$. Hence there exists $\epsilon_a > 0$ such that for every interval $J \ni \alpha^\star$ with $|J| \leq \epsilon_a$,

$$b_{a,a'}(J) \leq b_{a,a'}(\alpha^\star) + \bar{s}_a/4, \qquad \forall a' \neq a.$$

Therefore

$$\Delta u_{a,a'}^\top \nu_a \geq b_{a,a'}(J) + \bar{s}_a/4, \qquad \forall a' \neq a,$$

so $s_a(J) > 0$ and $\mathrm{int}_{\mathrm{IC}}\big(R_a^J\big) \neq \emptyset$.

Taking the minimum of ϵ_a over the finite set A_{rel} gives $\epsilon_{\mathrm{rel}}(I, \alpha^\star) > 0$. This proves the equivalence. □

Algorithm 6 SAFEEXPLORE on a fixed interval

Require: Current interval $J = [\underline{\alpha}, \overline{\alpha}]$, instance I.

$L \leftarrow |J|$, $\eta \leftarrow L^2$.
Compute $\tau^J_{\mathrm{vtx}} = \{(p_i, v_i^J, a_i)\}_{i=1}^{M_J}$ using Algorithm 7. ▷ VERTEXSAFESCHEME.
if $P_{\mathrm{info}}(\tau^J_{\mathrm{vtx}}) = 0$ **then**
    return $(J, \tau^J_{\mathrm{vtx}})$.
for each $i \in \mathrm{MOV}(\tau^J_{\mathrm{vtx}})$ **do** ▷ CHOOSEMOVABLEBINDINGCONSTRAINT
    Choose one movable binding IC constraint $\Delta u_{a_i,a_i'}^\top \nu = b_{a_i,a_i'}(\beta_i), \qquad \beta_i \in \{\underline{\alpha}, \overline{\alpha}\}$.
    if $\beta_i = \underline{\alpha}$ **then**
        Mark i as LOWER.
    else
        Mark i as UPPER.
$(\ell, r) \leftarrow (\underline{\alpha}, \overline{\alpha})$.
$\tau_{\mathrm{last}} \leftarrow \tau^J_{\mathrm{vtx}}$.
while $r - \ell > \eta$ **do**
    for each $i \in \mathrm{MOV}(\tau^J_{\mathrm{vtx}})$ **do** ▷ BUILDPROBE
        if i is LOWER **then**
            $m_i \leftarrow \ell + \eta$.
        else
            $m_i \leftarrow r - \eta$.
        Replace v_i^J by the probe posterior $v_i^{\mathrm{pr}}(m_i)$ using Euclidean projection.
    Leave non-informative posteriors unchanged.
    Add at most one correction posterior, if needed, to restore Bayes plausibility. ▷ Lemma 4
    Let the resulting probe scheme be $\tau^{J,\eta}_{\mathrm{probe}}(\ell, r)$.
    $\tau_{\mathrm{last}} \leftarrow \tau^{J,\eta}_{\mathrm{probe}}(\ell, r)$.
    Play $\tau^{J,\eta}_{\mathrm{probe}}(\ell, r)$ until the first informative signal i is realized. ▷ RUNUNTILINFORMATIVE
    Observe the receiver's action a_t on that signal.
    if i is LOWER **then** ▷ REFINEINTERVAL
        if $a_t = a_i$ **then**
            $\ell \leftarrow \ell + \eta$.
        else
            return $([\ell, \ell + \eta], \tau_{\mathrm{last}})$.
    else
        if $a_t = a_i$ **then**
            $r \leftarrow r - \eta$.
        else
            return $([r - \eta, r], \tau_{\mathrm{last}})$.
return $([\ell, r], \tau_{\mathrm{last}})$.

C.4 Missing Details in Section 4.2

C.4.1 Full algorithm in the safe exploration stage

See the full safe exploration algorithm in Algorithm 6, with subroutine in Algorithms 7 and 8. We also introduce several LPs below, which are used in the algorithms.

Algorithm 7 VERTEXSAFESCHEME on interval J

Require: Interval $J = [\underline{\alpha}, \overline{\alpha}]$, instance $\mathcal{I}$.

1: For every $a \in A$, construct R_a^J and test whether $\mathrm{int}_{\mathrm{IC}}(R_a^J) \neq \emptyset$.
2: Solve the safe direct-recommendation LP (equation 3) on J.
3: Obtain posterior/action pairs $\{(\lambda_a, \bar{\nu}_a, a) : \lambda_a > 0\}$.
4: **for** each a with $\lambda_a > 0$ **do**
5: Decompose $\bar{\nu}_a$ into vertices of the polytope R_a^J using Algorithm 8:

$$\bar{\nu}_a = \sum_{m=1}^{q_a} \gamma_{a,m} v_{a,m}^J, \qquad v_{a,m}^J \in \mathrm{vert}(R_a^J).$$

6: Replace $(\lambda_a, \bar{\nu}_a, a)$ by $\{(\lambda_a \gamma_{a,m}, v_{a,m}^J, a)\}_{m=1}^{q_a}$.
7: Merge all resulting atoms and return

$$\tau_{\mathrm{vtx}}^J = \{(p_i, v_i^J, a_i)\}_{i=1}^{M_J}.$$

Algorithm 8 VERTEXDECOMPOSE a posterior in a safe polytope

Require: A polytope $P = R_a^J$ and a posterior $\nu \in P$.

1: Initialize a working vertex set V by solving the initialization LP.
2: **repeat**
3: Solve the restricted membership LP.
4: **if** the membership LP is feasible **then**
5: **return** the resulting decomposition.
6: Solve the separation LP to find c with $c^\top \nu > \max_{v \in V} c^\top v$.
7: Solve the pricing LP to get a new vertex v^{new}, add it to V.
8: **until** membership is feasible

Safe direct-recommendation LP. A safe direct-recommendation scheme can be computed through the variables $x_{a,\omega} \geq 0$, interpreted as the joint probability of state ω and recommendation a:

$$\begin{aligned}
\max_{x_{a,\omega} \geq 0} \quad & \sum_{a \in A} \sum_{\omega \in \Omega} x_{a,\omega} u_S(a, \omega) \\
\text{s.t.} \quad & \sum_{a \in A} x_{a,\omega} = \mu_0(\omega), && \forall \omega \in \Omega, \\
& \sum_{\omega \in \Omega} x_{a,\omega} \Delta u_{a,a'}(\omega) \geq b_{a,a'}(J) \sum_{\omega \in \Omega} x_{a,\omega}, && \forall a \in A,\ \forall a' \in A \setminus \{a\}, \\
& x_{a,\omega} = 0, && \forall \omega \in \Omega,\ \forall a \in A \text{ such that } \mathrm{intIC}(R_a^J) = \emptyset.
\end{aligned} \tag{3}$$

The last line is the strict-feasibility filtering step.

If $\lambda_a := \sum_\omega x_{a,\omega} > 0$, the induced posterior is

$$\bar{\nu}_a(\omega) := \frac{x_{a,\omega}}{\lambda_a} \in R_a^J.$$

Thus the LP returns a Bayes-plausible safe scheme

$$\{(\lambda_a, \bar{\nu}_a, a) : \lambda_a > 0\}.$$

Initialization LP. Choose any generic vector $r \in \mathbb{Q}^{|\Omega|}$ and solve

$$v^{(1)} \in \arg\max_{x \in P} r^\top x,$$

returning any optimal basic feasible solution; this is a vertex of P.

Restricted membership LP. Given a current set $V = \{v^1, \ldots, v^m\} \subseteq \mathrm{vert}(P)$, solve

$$\text{find } \lambda \in \mathbb{R}_+^m \quad \text{s.t.} \quad \mathbf{1}^\top \lambda = 1, \qquad \sum_{j=1}^m \lambda_j v^j = \nu.$$

If this LP is feasible, then $\nu \in \mathrm{conv}(V)$ and the resulting coefficients already give a decomposition.

Separation LP. If the membership LP is infeasible, solve

$$\max_{c\in\mathbb{R}^{|\Omega|},\,\gamma\in\mathbb{R}} \; c^\top \nu - \gamma \quad \text{s.t.} \quad c^\top v^j \le \gamma \;\; (j=1,\dots,m), \qquad \|c\|_\infty \le 1.$$

Any optimal solution with positive objective value yields a separating hyperplane c such that

$$c^\top \nu > \max_{v\in V} c^\top v.$$

Pricing LP. Given such a c, solve

$$v^{\mathrm{new}} \in \arg\max_{x\in P} c^\top x,$$

and again return any optimal basic feasible solution; this is an extreme point of P.

We then describe the omitted details in Section 4.2.2.

Details of the local probe posterior construction in Section 4.2.2 Fix an informative vertex $v \in R_a^J$ and a selected movable binding constraint $\Delta u_{a,a'}^\top v = b_{a,a'}(\beta)$, where $\beta \in \{\underline{\alpha}, \overline{\alpha}\}$ is the endpoint determining the interval-safe constraint. For a probe value $m \in J$, define $v^{\mathrm{pr}}(m)$ as a minimum l_2 distance point satisfying the same simplex constraints, all non-selected IC constraints, and the modified equality

$$\Delta u_{a,a'}^\top v^{\mathrm{pr}}(m) = b_{a,a'}(m).$$

Lemma 3 shows that this perturbation has distance $O(|J|)$ from v. After perturbing all informative vertices, Lemma 4 adds at most one correction posterior to restore Bayes plausibility, losing only $O(P_{\mathrm{info}}(\tau_{\mathrm{vtx}}^J)|J|)$ utility.

C.4.2 Geometric stability and perturbation bounds

Before we show proofs of all propositions, we first list two useful lemmas here. For the definition of κ_a and κ, see Section C.1. These constants depend only on the constraint normals and not on the interval J.

Lemma 3 (Uniform distance bound under RHS perturbations)**.** *Fix $a \in A$. For any two parameter values $\beta_0, \beta_1 \in [\alpha_{\min}(I), 1]$ and any $\nu_0 \in R_a^{\beta_0}$,*

$$\mathrm{dist}(\nu_0, R_a^{\beta_1}) \le \kappa_a \left\| \left(\widetilde{b}_a(\beta_1) - \widetilde{b}_a(\beta_0) \right)_+ \right\|_2 \le \kappa_a \|\widetilde{b}_a(\beta_1) - \widetilde{b}_a(\beta_0)\|_2.$$

The same bound holds with R_a^J in place of $R_a^{\beta_1}$ by taking $\widetilde{b}_a(J)$ as the right-hand side.

Proof. Write $\nu_0 = \bar{\nu} + Q z_0$. Then $z_0 \in \widetilde{R}_a^{\beta_0}$, and

$$\mathrm{dist}(\nu_0, R_a^{\beta_1}) = \mathrm{dist}(z_0, \widetilde{R}_a^{\beta_1}).$$

Define the componentwise violation vector

$$r := \left(\widetilde{b}_a(\beta_1) - \widetilde{A}_a z_0 \right)_+ \in \mathbb{R}_+^{m_a}.$$

Since $z_0 \in \widetilde{R}_a^{\beta_0}$, we have $\widetilde{A}_a z_0 \ge \widetilde{b}_a(\beta_0)$, and therefore

$$r \le \left(\widetilde{b}_a(\beta_1) - \widetilde{b}_a(\beta_0) \right)_+ \quad \text{componentwise.}$$

Thus

$$\|r\|_2 \le \left\| \left(\widetilde{b}_a(\beta_1) - \widetilde{b}_a(\beta_0) \right)_+ \right\|_2.$$

It remains to show $\mathrm{dist}(z_0, \widetilde{R}_a^{\beta_1}) \le \kappa_a \|r\|_2$. By definition,

$$\mathrm{dist}(z_0, \widetilde{R}_a^{\beta_1}) = \min_y \{ \|y\|_2 : \widetilde{A}_a y \ge \widetilde{b}_a(\beta_1) - \widetilde{A}_a z_0 \}.$$

Since $r \geq \widetilde{b}_a(\beta_1) - \widetilde{A}_a z_0$ componentwise, it suffices to consider

$$\min_y \{\|y\|_2 : \widetilde{A}_a y \geq r\}.$$

Let $y^\star$ be a minimum-norm solution. Let B be a full-row-rank subset of binding constraints at $y^\star$. Then $B \in \mathcal{B}_a$, and $y^\star$ is the minimum-norm solution of

$$\widetilde{A}_{a,B} y = r_B.$$

Hence

$$y^\star = \widetilde{A}_{a,B}^\dagger r_B, \qquad \|y^\star\|_2 \leq \|\widetilde{A}_{a,B}^\dagger\|_2 \|r_B\|_2 \leq \kappa_a \|r\|_2.$$

Mapping back from z-coordinates to posteriors preserves Euclidean distance. This proves the first inequality. The second follows from $\| (x)_+ \|_2 \leq \|x\|_2$. □

Lemma 4 (Bayes-plausibility repair under small perturbations)**.** *Let $\tau = \{(\lambda_i, \nu_i, a_i)\}_{i=1}^k$ be Bayes-plausible: $\lambda_i \geq 0, \sum_i \lambda_i = 1, \sum_i \lambda_i \nu_i = \mu_0$. Suppose we perturb a subset of posteriors and obtain $\nu_i' \in \Delta(\Omega)$ such that*

$$\|\nu_i' - \nu_i\|_2 \leq \epsilon \quad \textit{whenever } \nu_i' \neq \nu_i, \qquad \sum_{\nu_i' \neq \nu_i} \lambda_i \leq p.$$

Then there exists an augmented Bayes-plausible scheme $\widehat{\tau} = \{(\widehat{\lambda}_i, \nu_i', a_i)\}_{i=1}^k \cup \{(\widehat{\lambda}_+, \nu_+, a_0)\}$, where $\nu_+ \in R_{a_0}^J$, such that

$$|U_S(\tau) - U_S(\widehat{\tau})| \leq C_{\mathrm{rep}}\, p\epsilon, \qquad C_{\mathrm{rep}} := U_{\max}\left(\sqrt{|\Omega|} + \frac{4}{\delta_{\mu_0}}\right).$$

If no repair is needed, then $\widehat{\lambda}_+ = 0$.

Proof. Let

$$\mu_0' := \sum_i \lambda_i \nu_i', \qquad r := \mu_0 - \mu_0'.$$

Since both μ_0 and μ_0' lie in the affine hull of the simplex, $\mathbf{1}^\top r = 0$. Moreover,

$$\|r\|_2 = \left\| \sum_i \lambda_i (\nu_i - \nu_i') \right\|_2 \leq \sum_{\nu_i' \neq \nu_i} \lambda_i \|\nu_i - \nu_i'\|_2 \leq p\epsilon.$$

If $r = 0$, set $\widehat{\lambda}_+ = 0$, $\widehat{\lambda}_i = \lambda_i$, and we are done.

Assume $r \neq 0$. Define

$$\widehat{\lambda}_+ := \frac{2\|r\|_2}{\delta_{\mu_0} + 2\|r\|_2}, \qquad \widehat{\lambda}_i := (1 - \widehat{\lambda}_+)\lambda_i,$$

and

$$\nu_+ := \mu_0 + \frac{1 - \widehat{\lambda}_+}{\widehat{\lambda}_+} r.$$

By construction,

$$\|\nu_+ - \mu_0\|_2 = \frac{\delta_{\mu_0}}{2}.$$

Because $\mathbf{1}^\top r = 0$, we have $\mathbf{1}^\top \nu_+ = 1$. By the standing margin convention, $\nu_+ \in \Delta(\Omega) \cap R_{a_0}^J$.

Bayes plausibility holds exactly:

$$\begin{aligned} \sum_i \widehat{\lambda}_i \nu_i' + \widehat{\lambda}_+ \nu_+ &= (1 - \widehat{\lambda}_+)\mu_0' + \widehat{\lambda}_+ \left(\mu_0 + \frac{1 - \widehat{\lambda}_+}{\widehat{\lambda}_+} r\right) \\ &= (1 - \widehat{\lambda}_+)\mu_0' + \widehat{\lambda}_+ \mu_0 + (1 - \widehat{\lambda}_+)(\mu_0 - \mu_0') = \mu_0. \end{aligned}$$

Also,

$$\widehat{\lambda}_+ \le \frac{2\|r\|_2}{\delta_{\mu_0}} \le \frac{2p\epsilon}{\delta_{\mu_0}}.$$

For utility, write

$$v_a(\nu) := \sum_{\omega\in\Omega} \nu(\omega) u_S(a,\omega).$$

Then

$$|v_a(\nu) - v_a(\nu')| \le U_{\max}\sqrt{|\Omega|}\,\|\nu - \nu'\|_2.$$

Therefore,

$$\begin{aligned}
|U_S(\tau) - U_S(\widehat{\tau})| &\le U_{\max}\sqrt{|\Omega|} \sum_{\nu_i'\neq\nu_i} \lambda_i \|\nu_i - \nu_i'\|_2 + U_{\max}\sum_i |\lambda_i - \widehat{\lambda}_i| + U_{\max}\widehat{\lambda}_+ \\
&\le U_{\max}\sqrt{|\Omega|}\,p\epsilon + 2U_{\max}\widehat{\lambda}_+ \\
&\le U_{\max}\sqrt{|\Omega|}\,p\epsilon + \frac{4U_{\max}}{\delta_{\mu_0}} p\epsilon.
\end{aligned}$$

This proves the claim. □

C.4.3 Proof of Proposition 3

Since $\tau_{\text{opt}} = \{(\lambda_i, \nu_i, a_i)\}_{i=1}^k$ is a full-information optimal posterior-form scheme at the true bias $\alpha^\star$, where $a_i = a^\star(\nu_i; \alpha^\star)$. Thus $\nu_i \in R_{a_i}^{\alpha^\star}$. We project each ν_i into the interval-safe region for the same action:

$$\nu_i' \in \arg\min_{\nu\in R_{a_i}^J} \|\nu - \nu_i\|_2.$$

The localization stage ensures that $R_{a_i}^J \neq \emptyset$ whenever $R_{a_i}^{\alpha^\star} \neq \emptyset$, so this projection is well-defined.

For every $a' \neq a_i$ and every $\beta \in J$,

$$|b_{a_i,a'}(\beta) - b_{a_i,a'}(\alpha^\star)| \le \sup_{\alpha\in[\alpha_{\min}(I),1]} \left|\frac{d}{d\alpha} b_{a_i,a'}(\alpha)\right| |J|.$$

Since

$$\frac{d}{d\alpha} b_{a,a'}(\alpha) = \frac{C_{a,a'}}{\alpha^2},$$

and Proposition 7 gives $\alpha_{\min}(I) \ge \delta_{\mu_0}/\sqrt{2}$, we have

$$|b_{a_i,a'}(\beta) - b_{a_i,a'}(\alpha^\star)| \le \frac{2G_{\max}}{\delta_{\mu_0}^2}|J|.$$

Applying Lemma 3 yields an instance-dependent constant $K_{\text{dist}} = \frac{2G_{\max}\kappa}{\delta_{\mu_0}^2}$ such that

$$\|\nu_i' - \nu_i\|_2 \le K_{\text{dist}}|J| \qquad \forall i.$$

Now apply Lemma 4 to the perturbation $\nu_i \mapsto \nu_i'$, with $p = 1$ and $\epsilon = K_{\text{dist}}|J|$. We obtain a Bayes-plausible scheme $\widehat{\tau}$ supported on R_{safe}^J such that

$$U_S(\tau_{\text{opt}}) - U_S(\widehat{\tau}) \le C_{\text{rep}} K_{\text{dist}}|J|.$$

Since τ_{vtx}^J is optimal among interval-safe schemes,

$$U_S(\tau_{\text{vtx}}^J) \ge U_S(\widehat{\tau}).$$

Therefore

$$0 \le U_S(\tau_{\text{opt}}) - U_S(\tau_{\text{vtx}}^J) \le C_{\text{rep}} K_{\text{dist}}|J| = O(|J|).$$

C.4.4 Proof of Proposition 4

Fix scan endpoints $(\ell, r) \subseteq J$. Let

$$p_u := P_{\text{info}}(\tau^J_{\text{vtx}}).$$

Only informative posteriors are perturbed, and their total probability is p_u. From Lemma 3, for every informative index i,

$$\|v_i^{\text{pr}}(m_i) - v_i^J\|_2 \le K_{\text{pr}}|J|,$$

where $K_{\text{pr}} = \frac{2G_{\max}\kappa}{\delta^2_{\mu_0}}$, similar to the proof of Proposition 3.

Applying Lemma 4 with

$$p = p_u, \qquad \epsilon = K_{\text{pr}}|J|$$

gives a Bayes-plausible probe scheme $\tau^{J,\eta}_{\text{probe}}(\ell, r)$ such that

$$\left|U_S(\tau^J_{\text{vtx}}) - U_S(\tau^{J,\eta}_{\text{probe}}(\ell, r))\right| \le C_{\text{rep}} K_{\text{pr}}\, p_u |J|.$$

Thus

$$U_S(\tau^J_{\text{vtx}}) - U_S(\tau^{J,\eta}_{\text{probe}}(\ell, r)) = O\big(P_{\text{info}}(\tau^J_{\text{vtx}})|J|\big)\,.$$

Moreover, the repair step scales the original atoms by a factor $1 - O(p_u|J|)$ and adds only a default correction posterior. Therefore, for all sufficiently small intervals J,

$$P_{\text{info}}(\tau^{J,\eta}_{\text{probe}}(\ell, r)) \ge \frac{1}{2}\, P_{\text{info}}(\tau^J_{\text{vtx}}).$$

C.4.5 Proof of Proposition 5

Finite vertex-index representation. For the proof of Proposition 5, we use a fixed finite index set for all possible vertices of the interval-safe regions. Each interval-safe action region R^J_a is a polytope, and its constraints consist of the simplex equality $\mathbf{1}^\top \nu = 1$, the IC inequalities $\Delta u_{a,a'}^\top \nu \ge b^J_{a,a'}$, and the non-negativity inequalities $\nu(\omega) \ge 0$.

To solve a possible candidate of vertex, we first fix one action a and select $n-1$ candidate binding constraints chosen from the IC and non-negativity constraints of R^J_a, then turn these constraints into equalities and add the simplex equality. This gives a linear system. If the linear system has a unque solution, it then corresponds to one possible vertex. We use *vertex index* $i = (B_i, a(i))$ to do such process for every possible action and set of constraints, where B_i is a collection of $n-1$ candidate binding constraints and $a(i)$ is the corresponding recommending action. Then the linear system is denoted by $\boldsymbol{A}_i \nu = \boldsymbol{b}_i^J$, where $\boldsymbol{A}_i$ depends only on the chosen constraint normals and not on J, while $\boldsymbol{b}_i^J$ depends on J through the safe right-hand sides $b^J_{a(i),a'}$.

We keep only indices for which $\boldsymbol{A}_i$ is invertible, namely there exists a unique solution of the linear system. Use $\mathcal{I}^J_{\text{vtx}}$ to denote the set of such indexes. Note that firstly, since $\boldsymbol{A}_i$ doesn't depend on J, then $\mathcal{I}^J_{\text{vtx}}$ doesn't depend on J, namely each index i corresponds to the same action and the same set of constraints. By saying the same constraint, we mean it is the IC constraint on the same action pair or the same non-negative constraints.

Also, the actual set of vertices is actually a subset of $\mathcal{I}^J_{\text{vtx}}$. Besides, $\mathcal{I}^J_{\text{vtx}}$ may include infeasible points which violate other constraints, or duplicate points like degenerate cases or overlapping of different action regions. But since our goal is to have an uniform representation to denote the vertex-supported scheme, $\mathcal{I}^J_{\text{vtx}}$ is enough to use. Infeasible candidate vertices are assigned probability zero.

For notational simplicity, enumerate $\mathcal{I}^J_{\text{vtx}} = \{1, \dots, M\}$. Each $i \in \mathcal{I}^J_{\text{vtx}}$ therefore labels a possible vertex $v_i^J \in \text{vert}\left(R^J_{a(i)}\right)$.

Thus any interval-safe vertex-supported scheme can be represented as

$$\tau^J = \{(p_i, v_i^J, a_i)\}_{i=1}^M, \qquad p \in \Delta(\{1, \dots, M\}), \qquad \sum_{i=1}^M p_i v_i^J = \mu_0,$$

with $p_i = 0$ whenever v_i^J is infeasible.

Formal proof We prove the dichotomy by contradiction. Suppose the first condition fails. Then for every $n \geq 1$, there exists an interval $J_n = [\underline{\alpha}_n, \overline{\alpha}_n]$ with

$$\alpha^\star \in J_n, \qquad |J_n| \leq \frac{1}{n},$$

and an interval-safe vertex-supported optimizer

$$\tau^n = \{(p_i^n, v_i^{J_n}, a_i)\}_{i=1}^M$$

such that

$$P_{\text{info}}(\tau^n) \leq \frac{1}{n}.$$

Because the probability simplex in $\mathbb{R}^M$ is compact, there is a subsequence, still denoted by n, such that

$$p^n \to p^\star.$$

Since $\underline{\alpha}_n, \overline{\alpha}_n \in [\alpha_{\min}(I), 1]$, compactness also gives a subsequence such that

$$\underline{\alpha}_n \to \alpha^-, \qquad \overline{\alpha}_n \to \alpha^+.$$

Because $\alpha^\star \in J_n$ and $|J_n| \to 0$, we have

$$\alpha^- = \alpha^+ = \alpha^\star.$$

By continuity of each candidate vertex,

$$v_i^{J_n} \to v_i^{\alpha^\star} \qquad \forall i.$$

Define the limiting scheme

$$\tau^\star := \{(p_i^\star, v_i^{\alpha^\star}, a_i)\}_{i=1}^M.$$

We first show feasibility at $\alpha^\star$. If $p_i^\star > 0$, then for all sufficiently large n, $p_i^n > 0$. Since τ^n is feasible for J_n, the candidate posterior $v_i^{J_n}$ satisfies all IC and nonnegativity constraints of action a_i under J_n. Taking limits gives

$$\Delta u_{a_i,a'}^\top v_i^{\alpha^\star} \geq b_{a_i,a'}(\alpha^\star), \qquad \forall a' \neq a_i,$$

and also $v_i^{\alpha^\star} \in \Delta(\Omega)$. Bayes plausibility follows from

$$\sum_i p_i^n v_i^{J_n} = \mu_0$$

by taking limits. Therefore $\tau^\star$ is feasible at the true bias $\alpha^\star$.

Next we show optimality. Since τ^n is safe-optimal on J_n,

$$U_S(\tau^n) = \text{OPT}^{\text{safe}}(J_n).$$

Because J_n tightens the IC constraints relative to the true-bias problem,

$$\text{OPT}^{\text{safe}}(J_n) \leq \text{OPT}_I(\alpha^\star)\,.$$

On the other hand, Proposition 3 gives

$$\text{OPT}^{\text{safe}}(J_n) \geq \text{OPT}_I(\alpha^\star) - C_{\text{rep}} K_{\text{dist}} |J_n|$$

where constants are only related to instance $\mathcal{I}$, but not J. Hence

$$\text{OPT}^{\text{safe}}(J_n) \to \text{OPT}_I(\alpha^\star)\,.$$

By continuity of the finite representation,

$$U_S(\tau^n) \to U_S(\tau^\star).$$

Therefore

$$U_S(\tau^\star) = \text{OPT}_I(\alpha^\star)\,,$$

so $\tau^\star$ is optimal at the true bias.

Finally, since $P_{\text{info}}(\tau^n) \le 1/n$, we must have

$$P_{\text{info}}(\tau^\star) = 0.$$

Indeed, if some movable informative index had $p_i^\star > 0$, then $p_i^n \ge p_i^\star/2$ for all large n, contradicting $P_{\text{info}}(\tau^n) \le 1/n$. Thus the second condition holds: there exists a true-bias optimal scheme with zero informative probability.

The remaining statement we will show is that if such a zero-informative optimal scheme exists, then there exists $\delta_{\text{stab}} > 0$, for every interval $J \ni \alpha^\star$ with $|J| \le \delta_{\text{stab}}$, the same scheme $\tau^\star$ is feasible and optimal for the interval-safe problem on J. Consequently, $\mathrm{OPT}^{\text{safe}}(J) = \mathrm{OPT}_I(\alpha^\star)$.

Write the finite-support optimal scheme as

$$\tau^\star = \{(p_i, \nu_i, a_i)\}_{i=1}^k, \qquad p_i > 0, \qquad \sum_i p_i \nu_i = \mu_0,$$

where $a_i = a^\star(\nu_i; \alpha^\star)$.

For every action pair (a, a'), define the IC slack

$$\text{slack}_{a,a'}(\nu, \alpha) := \Delta u_{a,a'}^\top \nu - b_{a,a'}(\alpha).$$

A movable constraint is one with $C_{a,a'} \neq 0$. The condition $P_{\text{info}}(\tau^\star) = 0$ means that no support posterior of $\tau^\star$ lies on a movable binding IC constraint. Hence, for every support atom i and every movable pair (a_i, a'),

$$\text{slack}_{a_i,a'}(\nu_i, \alpha^\star) > 0.$$

Because the support is finite, the minimum positive movable slack

$$s^\star := \min_i \min_{\substack{a' \neq a_i \\ C_{a_i,a'} \neq 0}} \text{slack}_{a_i,a'}(\nu_i, \alpha^\star)$$

is strictly positive, ignoring pairs over an empty set.

By uniform continuity of $b_{a,a'}(\cdot)$ on $[\alpha_{\min}(I), 1]$, there exists $\delta_{\text{stab}} > 0$ such that whenever $J \ni \alpha^\star$ and $|J| \le \delta_{\text{stab}}$,

$$b_{a,a'}^J \le b_{a,a'}(\alpha^\star) + \frac{s^\star}{2}$$

for every movable pair that appears in the support of $\tau^\star$. Therefore all movable IC constraints remain satisfied on J. Non-movable constraints have $C_{a,a'} = 0$, so $b_{a,a'}(\alpha) \equiv 0$ and are unchanged. Thus every support posterior of $\tau^\star$ lies in its corresponding interval-safe region, and $\tau^\star$ is feasible for the safe problem on J.

Since the safe feasible set is a subset of the feasible set at the true bias $\alpha^\star$,

$$\mathrm{OPT}^{\text{safe}}(J) \le \mathrm{OPT}_I(\alpha^\star).$$

Feasibility of $\tau^\star$ gives the reverse inequality. Hence

$$\mathrm{OPT}^{\text{safe}}(J) = \mathrm{OPT}_I(\alpha^\star),$$

and $\tau^\star$ is safe-optimal on J.

C.4.6 Proof of Proposition 6

Consider a safe-exploration phase starting from interval $J_r = [\underline{\alpha}_r, \overline{\alpha}_r]$ of length L_r. The algorithm uses step size

$$\eta_r := L_r^2.$$

Each safe informative realization moves either the left endpoint or the right endpoint inward by η_r. A unsafe informative realization localizes $\alpha^\star$ to an adjacent interval of length η_r. Therefore the phase contains at most

$$\left\lceil \frac{L_r}{\eta_r} \right\rceil = O\left(\frac{1}{L_r}\right)$$

safe probes and at most one unsafe probe. In all cases,

$$L_{r+1} \le \eta_r = L_r^2.$$

We now bound the regret of one phase. Let

$$p_u := P_{\mathrm{info}}(\tau_{\mathrm{vtx}}^{J_r})$$

for the safe vertex-supported optimizer used at the start of the phase. By the proof of Proposition 4, the corresponding probe scheme has informative-signal probability at least $p_u/2$ on sufficiently small intervals.

There are two cases from Proposition 5.

Case 1: informative probability is bounded below. There exist constants $c > 0$ and $\epsilon > 0$ such that for every interval $J \ni \alpha^\star$ with $|J| \le \epsilon$, every interval-safe vertex-supported optimizer satisfies

$$P_{\mathrm{info}}(\tau_{\mathrm{vtx}}^{J}) \ge c.$$

Then every probe has informative-signal probability at least $c/2$, so the expected waiting time for an informative realization is $O(1/c)$.

By Proposition 3, the safe scheme is $O(L_r)$-optimal relative to the full-information benchmark. By Proposition 4, the probe perturbation changes utility by at most $O(p_u|L_r|) \le O(|L_r|)$. Multiplying by the expected waiting time, one safe probe contributes $O(|L_r|/c)$ expected regret. Since there are $O(|L_r|/|L_r|^2) = O(1/|L_r|)$ safe probes, their total contribution in this phase is $O(1/c)$.

There is at most one unsafe probe. Its per-round regret is at most $\Delta U_{\max}$, and its expected waiting time is $O(1/c)$, so it contributes $O(1/c)$. Thus the whole phase contributes $O(1)$ expected regret, with constants depending only on the instance.

Case 2: a zero-informative optimal scheme exists. For all sufficiently small intervals $J_r \ni \alpha^\star$,

$$\mathrm{OPT}^{\mathrm{safe}}(J_r) = \mathrm{OPT}_I(\alpha^\star)\,.$$

Hence the baseline safe scheme already achieves the full-information benchmark. The only loss comes from the probe perturbation.

For an safe probe, Proposition 4 gives per-round regret $O(p_u|L_r|)$, while the expected waiting time is $O(1/p_u)$. Thus one safe probe contributes $O(|L_r|)$. Summing over $O(|L_r|/|L_r|^2) = O(1/|L_r|)$ safe probes gives

$$O\left(\frac{1}{|L_r|} \cdot |L_r|\right) = O(1).$$

For the single unsafe probe, only informative posteriors can induce the wrong action. Their total probability is $O(p_u)$, and the expected waiting time is $O(1/p_u)$, so the unsafe probe contributes $O(1)$. Therefore this phase also has $O(1)$ expected regret.

Number of phases. The interval lengths satisfy

$$L_{r+1} \le L_r^2, \qquad L_0 \le \frac{1}{\log T}.$$

Thus

$$L_r \le \left(\frac{1}{\log T}\right)^{2^r}.$$

The smallest r such that $L_r \le 1/T$ is $O(\log\log T)$. Therefore the total regret over all safe-exploration phases is $O(\log\log T)$.

C.5 Proof of Theorem 2

C.5.1 Runtime of the general algorithm

The localization stage solves the threshold-test LP, which has $|A||\Omega|$ variables and polynomially many constraints. The safe direct-recommendation LP used in VERTEXSAFESCHEME has the same polynomial scale.

For vertex decomposition, the polytope R_a^J is given explicitly by polynomially many linear inequalities. Each call to VERTEXDECOMPOSE uses only polynomial-size LPs: a generic linear optimization

LP over R_a^J, a restricted membership LP over the currently discovered vertices, a separation LP, and a pricing LP over R_a^J. Equivalently, one may use the standard separation–optimization implementation of convex decomposition to obtain a formal polynomial-time routine. The final Carathéodory compression is a linear-algebraic elimination step.

The probe construction only solves constant-size linear systems associated with active vertex bases and then applies the Bayes-plausibility repair step. Therefore each update of the online algorithm is polynomial in the finite instance description, and the total runtime up to horizon T is polynomial in T, $|A|$, and $|\Omega|$.

C.5.2 Regret of the general algorithm

We decompose the regret into the localization stage, the safe-exploration stage, and the final commitment stage.

Localization stage. The localization stage starts from the interval $[\alpha_{\min}(I), 1]$ and performs binary search using THRESHOLDTEST. It stops once the interval length is at most $1/\log T$. Therefore it requires $O(\log\log T)$ informative threshold-test realizations. By Proposition 8, the expected localization regret is $O(\log\log T)$.

Let J_0 denote the interval returned by localization, with length $L_0 \le 1/\log T$.

Safe exploration stage. From Proposition 6, the expected total regret in this stage is $O(\log\log T)$.

Commitment stage. If safe exploration reaches an interval J_R with $|J_R| \le 1/T$, the algorithm commits to an interval-safe vertex-supported optimizer on J_R. Proposition 3 implies that its per-round gap relative to the full-information benchmark is $O(1/T)$, so the total commitment regret is $O(1)$. If instead SAFEEXPLORE stops early because $P_{\text{info}}(\tau_{\text{vtx}}^{J_R}) = 0$, then Case 1 of Proposition 5 cannot hold on sufficiently small intervals. Hence Case 2 holds, and this implies that the current safe scheme is already optimal at the true bias. Committing to it incurs no additional asymptotic regret.

Combining the localization stage, the $O(\log\log T)$ safe-exploration phases, and the final commitment stage gives

$$\mathrm{Reg}_T^{\mathcal{I}}\big(\Pi^{\mathrm{GSE}}; \alpha^\star\big) = O(\log\log T).$$

D Receiver Welfare under Biased Persuasion

This appendix compares the receiver's ex ante utility under persuasion with her no-persuasion utility. All notation follows the model section. In the classical setting [27], the receiver's utility cannot decrease under persuasion: additional information cannot reduce the value of an optimal decision, because the receiver can always ignore the signal and play the default action.

In our setting, the receiver is non-Bayesian. Even if the sender provides more information, it is not clear whether the receiver's true expected utility increases, because the receiver chooses actions based on a biased belief. We show that under linear bias model, the non-decreasing guarantee still holds.

Proposition 9 (Preservation of consistency)**.** *If ρ is Bayes-plausible, then the induced distribution of distorted posteriors also has mean μ_0: $\mathbb{E}_{\nu\sim\rho}\big[\alpha^\star\nu + (1-\alpha^\star)\mu_0\big] = \mu_0$.*

Proposition 10 (Persuasion does not hurt the biased receiver)**.** *For any bias level $\alpha^\star$ and any feasible signal scheme $\tau = \{(p_i, \nu_i)\}_{i=1}^k$, the receiver's expected utility under τ is at least her no-persuasion utility.*

Proof. Consider any feasible signal scheme $\tau = \{p_i, \nu_i\}_{i=1}^k$. Under the signal scheme, the receiver's utility can be rewritten as below.

$$\sum_i p_i \sum_\omega u_R(a^*(\nu_i;\alpha^\star),\omega)\nu_i(\omega) = \sum_i p_i \sum_\omega u_R(a^*(\nu_i^{\alpha^\star}),\omega)\left(\frac{1}{\alpha^\star}\nu_i^{\alpha^\star}(\omega) - \frac{1-\alpha^\star}{\alpha^\star}\mu_0(\omega)\right)$$

$$(\nu_i^{\alpha^\star} = \alpha^\star\nu_i + (1-\alpha^\star)\mu_0)$$

$$\geq \sum_i p_i \sum_\omega u_R(a_0,\omega)\frac{1}{\alpha^\star}\nu_i^{\alpha^\star}(\omega) - \frac{1-\alpha^\star}{\alpha^\star}\sum_\omega u_R(a_0,\omega)\mu_0(\omega)$$

($a^*(\nu_i^{\alpha^\star})$ is the optimal action facing belief $\nu_i^{\alpha^\star}$, a_0 is the optimal action facing belief μ_0)

$$= \sum_\omega u_R(a_0,\omega)\mu_0(\omega)$$

$$(\textstyle\sum_i p_i\nu_i^{\alpha^\star} = \mu_0 \text{ from Proposition 9, } \frac{1}{\alpha^\star} - \frac{1-\alpha^\star}{\alpha^\star} = 1)$$

Since $\sum_\omega u_R(a_0,\omega)\mu_0(\omega)$ is the receiver's utility without persuasion, we conclude the proof. □

Another question is: for a fixed bias level $\alpha^\star$, how does the receiver's utility compare between (i) a sender who correctly accounts for the receiver's bias when optimizing the signal scheme, and (ii) a sender who incorrectly treats the receiver as Bayesian when optimizing the signal scheme? The comparison can go in either direction.

The main intuition is that when the sender does not know the receiver's true bias level, the sender cannot precisely target the receiver's decision thresholds. This imprecision can cut in either direction: the sender may fail to extract as much utility as under correct knowledge, which can leave the receiver better off, or the sender may inadvertently push the receiver into suboptimal actions under the true bias, reducing the receiver's utility and potentially lowering the sender's utility as well.

Example 1. *Let $\Omega = \{0,1\}$, $A = \{a_1, a_2, a_3\}$, and $\mu_0(1) = 0.1$. The receiver's utilities are*

$$u_R(a_1,0) = u_R(a_1,1) = 0, \qquad u_R(a_2,0) = -1, \quad u_R(a_2,1) = 4,$$
$$u_R(a_3,0) = -51, \qquad u_R(a_3,1) = 54.$$

For a binary belief $q = \Pr(\omega = 1)$, the receiver obtains

$$\mathbb{E}[u_R(a_2,\omega) \mid q] = -1 + 5q, \qquad \mathbb{E}[u_R(a_3,\omega) \mid q] = -51 + 105q.$$

Thus she chooses a_1 for $q \leq 0.2$, a_2 for $0.2 \leq q \leq 0.5$, and a_3 for $q \geq 0.5$. Let the sender's utility depend only on the induced action:

$$u_S(a_1) = 0, \qquad u_S(a_2) = 2, \qquad u_S(a_3) = 10.$$

First take $\alpha^\star = 0.2$. A sender who mistakenly treats the receiver as Bayesian chooses the two-posterior scheme $q_L = 0$, $q_H = 0.5$, with probabilities 0.8 and 0.2. Under the true bias,

$$\widehat{q}_L = 0.2 \cdot 0 + 0.8 \cdot 0.1 = 0.08, \qquad \widehat{q}_H = 0.2 \cdot 0.5 + 0.8 \cdot 0.1 = 0.18,$$

so the receiver chooses a_1 after both signals and obtains utility 0. By contrast, a sender who accounts for $\alpha^\star = 0.2$ cannot induce a_3, and induces a_2 using $q_L = 0$, $q_H = 0.6$, with probabilities $5/6$ and $1/6$. The receiver's utility is

$$\frac{1}{6}(-1 + 5 \cdot 0.6) = \frac{1}{3}.$$

Hence the receiver is better off when the sender knows $\alpha^\star$.

Now take $\alpha^\star = 1/3$. The same Bayesian-designed scheme $(q_L, q_H) = (0, 0.5)$, with probabilities $(0.8, 0.2)$, induces distorted high belief

$$\widehat{q}_H = \frac{1}{3} \cdot 0.5 + \frac{2}{3} \cdot 0.1 = \frac{7}{30},$$

so the receiver chooses a_2 after the high signal and obtains

$$0.2(-1 + 5 \cdot 0.5) = 0.3.$$

A sender who correctly accounts for $\alpha^\star = 1/3$ again cannot induce a_3. To induce a_2, the minimal Bayesian posterior is $q_H = 0.4$, so the optimal scheme uses $q_L = 0$, $q_H = 0.4$, with probabilities 0.75 and 0.25. The receiver obtains

$$0.25(-1 + 5 \cdot 0.4) = 0.25.$$

Hence the receiver is worse off when the sender knows $\alpha^\star$.

E Extension: Changing Priors and Utility Functions

We now allow the prior and utility functions to vary across rounds while the receiver's bias $\alpha^\star$ remains fixed. The state and action spaces Ω and A are fixed and finite. At round t, a context

$$x_t = (\mu_{x_t}, u_{S,x_t}, u_{R,x_t})$$

is realized and observed by the sender before she chooses a signaling scheme. Conditional on $x_t = x$, the state is drawn from μ_x. We do *not* require the sender to know the full family of possible contexts in advance, nor do we impose any stochastic assumption on the sequence $(x_t)_{t=1}^T$. Instead, we only assume that there exists a context set $\mathcal{X}$ such that $x_t \in \mathcal{X}$ for every t.

For each context x, write $I_x = (\Omega, A, \mu_x, u_{S,x}, u_{R,x})$. Define

$$\begin{aligned}
a_x^\star(\nu;\alpha) &\in \arg\max_{a\in A} \sum_{\omega\in\Omega} \big((1-\alpha)\mu_x(\omega) + \alpha\nu(\omega)\big) u_{R,x}(a,\omega),\\
R_{x,a}^\alpha &:= \{\nu \in \Delta(\Omega) : a_x^\star(\nu;\alpha) = a\},\\
V_x^\alpha(\nu) &:= \sum_{\omega\in\Omega} \nu(\omega) u_{S,x}\big(a_x^\star(\nu;\alpha),\omega\big),\\
\mathrm{OPT}_x(\alpha) &:= \sup_{\rho\in\Delta(\Delta(\Omega)):\mathbb{E}_{\nu\sim\rho}[\nu]=\mu_x} \mathbb{E}_{\nu\sim\rho}\big[V_x^\alpha(\nu)\big].
\end{aligned}$$

The dynamic oracle knows $\alpha^\star$ and, in every round t, plays an optimal Bayes-plausible scheme for the realized context x_t. For a fixed context sequence $\mathbf{x} = (x_1, \ldots, x_T)$, define the expected dynamic regret by

$$\mathrm{Reg}_T^{\mathbf{x}}(\Pi;\alpha^\star) := \sum_{t=1}^T \Big(\mathrm{OPT}_{x_t}(\alpha^\star) - \mathbb{E}_\Pi\big[V_{x_t}^{\alpha^\star}(\nu_t)\big]\Big).$$

Assumption 2 (Uniform contextual regularity)**.** *For every $x \in \mathcal{X}$, the instance I_x satisfies the Assumption 1, together with the tie-breaking rule of the relevant-action restriction. In addition:*

1. *The true bias is detectable in every context: $\alpha^\star \geq \bar{\alpha}_{\min} := \max_{x\in\mathcal{X}} \alpha_{\min}(I_x)$.*
2. *The sender utilities are uniformly bounded: $U_{\max}^{\mathcal{X}} := \max_{x\in\mathcal{X}} \max_{a,\omega} |u_{S,x}(a,\omega)| < \infty$.*
3. *For every context x, the analogue of Proposition 5 holds, and there exists a positive constant $c = \min_x c_x > 0, \epsilon = \min_x \epsilon_x > 0$ for the first branch in Proposition 5, and a positive constant $\delta = \min_x \delta_x > 0$ for the second branch.*
4. *The constants in Theorem 2 can be chosen uniformly over $x \in \mathcal{X}$, see more details in Section C.1. Equivalently, there exist positive constant $\delta_{\mu_0} = \min_{x\in\mathcal{X}} \delta_{\mu_0,x}$ and finite constants $\kappa, G_{\max}, L_b$ satisfying*

$$\kappa = \max_{x\in\mathcal{X}} \kappa_x, \quad G_{\max} = \max_{x\in\mathcal{X}} G_{\max,x}, \quad L_b = \max_{x\in\mathcal{X}} L_{b,x}.$$

Note that the above assumption naturally holds if the context set $\mathcal{X}$ is finite.

Contextual algorithm. The contextual algorithm maintains one global confidence interval J for the common parameter $\alpha^\star$. In each round, every LP, threshold test, safe scheme, and probe is instantiated with the realized context x_t. Thus the algorithm is the contextual analogue of Algorithm 2.

Theorem 3. *Under Assumption 2, for every fixed context sequence $\mathbf{x} \in \mathcal{X}^T$, the contextual version of Algorithm 2 Π_{CGSE} satisfies that*

$$\mathrm{Reg}_T^{\mathbf{x}}(\Pi_{\mathrm{CGSE}};\alpha^\star) = O(\log\log T).$$

Proof. The following proof is a contextual analogue of proof of Theorem 2.

Localization. The localization stage runs threshold tests at midpoints of the current global interval until its length is at most $1/\log T$. Since the interval is contained in $[\bar{\alpha}_{\min}, 1]$, the threshold test is feasible in every realized context. By the uniform version of Proposition 7, in every round of a threshold test the conditional probability of an informative non-default recommendation is at least $p_{\min} = \delta_{\mu_0}/\sqrt{2} > 0$. Therefore the waiting time for one informative comparison is stochastically dominated by a geometric random variable with mean $1/p_{\min}$, even though contexts may vary over time. The number of informative comparisons needed to reduce the interval from constant length to $1/\log T$ is $O(\log\log T)$. Since one-round regret is bounded by $2U^{\mathcal{X}}_{\max}$, localization contributes $O(\log\log T)$ expected regret.

One safe-exploration phase. Consider a safe-exploration phase that starts with interval J and step size $\eta = L^2$. For each realized context x, let $\tau^J_{x,\mathrm{vtx}}$ be the contextual interval-safe vertex-supported optimizer, and let

$$p^J_x = P_{\mathrm{info}}(\tau^J_{x,\mathrm{vtx}}).$$

If $p^J_x = 0$, the algorithm plays $\tau^J_{x,\mathrm{vtx}}$ and does not attempt to update the interval in that round. If $p^J_x > 0$, it plays the contextual probe constructed using the binding-movable-constraint construction from Section 4.2.2. By the uniform version of Proposition 4, the probability of an informative realization under this probe is at least $p^J_x/2$ for all sufficiently small intervals.

We first bound the regret accumulated until the next informative realization that corresponds to a safe probe, i.e., a realization that moves one endpoint inward by η. There are two cases, matching the dichotomy in Proposition 5. If context x is in the first branch of the dichotomy, then $p^J_x \geq c$, so the informative probability in that round is at least $c/2$. The safe-optimal gap is at most $O(L)$, and the probe perturbation adds at most $O(L)$. Thus the one-round regret on such a safe probe is at most $O(L)$, while its informative probability is at least $c/2$.

If context x is in the second branch, then the contextual safe optimum already attains $\mathrm{OPT}_x(\alpha^\star)$ on sufficiently small intervals. The only loss is the probe perturbation, at most $O(p^J_x L)$, while the informative probability is at least $p^J_x/2$. Hence the one-round regret is at most $O(L)$.

Combining the two cases, there is a uniform constant C such that, on every safe-probe round before the next safe informative realization,

$$\text{one-round regret} \leq CL \cdot \Pr(\text{informative realization in that round} \mid \mathcal{F}_{t-1}).$$

Let σ be the first informative realization time for the current probe. The standard hazard-rate identity gives

$$\mathbb{E}\left[\sum_{t \leq \sigma} \Pr(\sigma = t \mid \mathcal{F}_{t-1})\right] \leq 1.$$

Therefore the expected regret before one safe endpoint update is at most CL. During a phase, each safe informative realization moves one endpoint by $\eta = L^2$, so there are at most $\lceil L/\eta \rceil = O(1/L)$ such updates. The total expected regret from safe updates in the phase is therefore $O(1)$.

There can be at most one unsafe informative realization in the phase, because once such a realization occurs the algorithm localizes $\alpha^\star$ to an adjacent interval of length η and ends the phase. In the first branch of the dichotomy, the informative probability is uniformly bounded below, and one-round regret is uniformly bounded by $2U^{\mathcal{X}}_{\max}$, so the unsafe realization contributes $O(1)$ expected regret. In the second branch, the baseline safe scheme is optimal and only the perturbed informative mass can induce the wrong action; its total probability is $O(p^J_x)$, while the informative probability is at least $p^J_x/2$. Thus the unsafe realization also contributes $O(1)$ expected regret. Hence every completed safe-exploration phase contributes $O(1)$ expected regret. A phase truncated by the horizon is bounded by the same argument.

Number of phases and commitment. At the end of any completed safe-exploration phase, the new interval length satisfies

$$L_{r+1} \leq \eta_r = L_r^2.$$

After localization, $L_0 \leq 1/\log T$. Therefore after $O(\log\log T)$ phases the interval length is at most $1/T$. When this happens, the algorithm commits, in each realized context x_t, to the contextual

interval-safe optimizer for the final interval. The uniform safe-optimal gap gives per-round regret $O(1/T)$, and hence the total commitment regret is $O(1)$.

Adding the localization regret, the $O(\log\log T)$ safe-exploration phases, and the final commitment cost yields the claimed bound.

□

F Extension: Jointly Unknown Prior and Bias

F.1 Binary case with jointly unknown prior and bias.

We now consider the binary case in which both the receiver's bias and the prior are fixed across rounds and unknown to the sender. As in Section 3, let $\Omega = \{0,1\}$, $A = \{0,1\}$, $\mu^* := \Pr(\omega = 1) \in (0,1)$, and $\alpha^* \in [0,1]$. The sender utility is $u_S(a,\omega) = \mathbf{1}\{a = 1\}$, so the sender only cares about inducing action 1. For a realized signal s, let $\nu = \Pr(\omega = 1 \mid s)$ be the Bayesian posterior. Under the linear-bias model, the receiver evaluates this signal using the distorted posterior $(1-\alpha^*)\mu^* + \alpha^*\nu$.

Let $\widehat{q} \in (0,1)$ be the receiver's cutoff in distorted-belief space: the receiver takes action 1 iff $(1-\alpha^*)\mu^* + \alpha^*\nu \geq \widehat{q}$. Equivalently, there is a Bayesian-posterior threshold ν^* such that the receiver takes action 1 iff $\nu \geq \nu^*$, where ν^* is determined by $(1-\alpha^*)\mu^* + \alpha^*\nu^* = \widehat{q}$, i.e.,

$$\nu^* = \frac{\widehat{q} - (1-\alpha^*)\mu^*}{\alpha^*} = \mu^* + \frac{\widehat{q} - \mu^*}{\alpha^*}.$$

Thus ν^* is the critical threshold in posterior space.

Compared with the known-prior case, the difficulty is that when μ^* is known, one can guess α, compute the target posterior $\nu(\alpha) = \frac{\widehat{q}-(1-\alpha)\mu^*}{\alpha}$, and implement the corresponding binary scheme supported on $\{0, \nu(\alpha)\}$. When μ^* is unknown, this direct reduction is unavailable because implementing a target posterior itself depends on the prior. Nevertheless, under a natural two-signal family, the sender still faces a one-dimensional threshold-learning problem, much as in the binary-action unknown-prior model of [30].

Specifically, consider the implementable family $\{\pi_m\}_{m\in[0,1]}$ given by $\pi_m(\text{High} \mid \omega = 1) = 1$ and $\pi_m(\text{High} \mid \omega = 0) = m$. Thus the sender always sends High in state 1 and sends High with probability m in state 0. Under π_m, Bayes' rule gives $\nu(m;\mu^*) = \Pr(\omega = 1 \mid \text{High}) = \frac{\mu^*}{\mu^* + (1-\mu^*)m}$. Hence the receiver accepts High iff $(1-\alpha^*)\mu^* + \alpha^* \frac{\mu^*}{\mu^*+(1-\mu^*)m} \geq \widehat{q}$. This motivates the critical implementable threshold m^*, defined as the unique solution of

$$(1-\alpha^*)\mu^* + \alpha^* \frac{\mu^*}{\mu^* + (1-\mu^*)m^*} = \widehat{q},$$

namely

$$m^* = \frac{\mu^*(\mu^*(1-\alpha^\star) + \alpha^\star - \widehat{q})}{(1-\mu^*)\big(\widehat{q} - (1-\alpha^*)\mu^*\big)}.$$

Since $\nu(m;\mu^*)$ is decreasing in m, the scheme π_m is persuasive for High iff $m \leq m^*$. Therefore, although the primitive unknown is the pair (α^*, μ^*), within the family $\{\pi_m\}_{m\in[0,1]}$ the sender's problem reduces to identifying the single boundary m^*. The two thresholds are linked by $\nu^* = \frac{\mu^*}{\mu^*+(1-\mu^*)m^*}$, or equivalently $m^* = \frac{\mu^*(1-\nu^*)}{(1-\mu^*)\nu^*}$. Thus ν^* is the critical threshold in posterior space, while m^* is the corresponding threshold in the implementable signaling family.

Proposition 11 (Binary-case regret upper bound for jointly unknown prior and bias)**.** *Consider the binary model described above. Assume:*

1. $\Omega = A = \{0,1\}$*;*
2. $u_S(a,\omega) = \mathbf{1}\{a = 1\}$*;*
3. *the receiver takes action* 1 *if and only if the distorted posterior is at least* $\widehat{q}$*, where* $0 < \mu^* < \widehat{q} < 1$*;*

Algorithm 9 Safe Exploration for Unknown Bias and Unknown Prior

Input: horizon T, threshold $\widehat{q}$.
Initialize: interval $[a, b] \leftarrow [0, 1]$, step size $\varepsilon \leftarrow \frac{1}{2}$, time $t \leftarrow 1$.
Stage 1: Safe exploration
while $b - a > T^{-1}$ **and** $t \leq T$ **do**
  $m \leftarrow a, \quad m_{\text{prev}} \leftarrow a$
  while $m \leq b$ **and** $t \leq T$ **do**
    Commit to the signaling scheme π_m defined by

$$\pi_m(\text{High} \mid \omega = 1) = 1, \qquad \pi_m(\text{High} \mid \omega = 0) = m$$

    Realize signal $s_t \in \{\text{Low}, \text{High}\}$
    if $s_t = \text{Low}$ **then**
      $t \leftarrow t + 1$
      continue
    else
      Observe receiver action $a_t \in \{0, 1\}$
      if $a_t = 1$ **then**
        $m_{\text{prev}} \leftarrow m, \quad m \leftarrow m + \varepsilon, \quad t \leftarrow t + 1$
      else
        $[a, b] \leftarrow [m_{\text{prev}}, m], \quad \varepsilon \leftarrow \varepsilon^2, \quad t \leftarrow t + 1$
        break
  if $m > b$ **then**
    $[a, b] \leftarrow [m_{\text{prev}}, b], \quad \varepsilon \leftarrow \varepsilon^2$
Stage 2: Commit
$\widehat{m} \leftarrow a$
Commit to the signaling scheme $\pi_{\widehat{m}}$ for all remaining rounds

4. *persuasion is feasible, namely*

$$\alpha^* \geq \alpha_{min} = \frac{\widehat{q} - \mu^*}{1 - \mu^*}$$

   so that $m^* \in [0, 1]$ *is well-defined;*

5. *ties are broken in favor of action* 1.

Let Π^{SEJ} *denote Algorithm 9. Then, for every* $T \geq 4$*,*

$$\text{Reg}_T\big(\Pi^{\text{SEJ}}; \alpha^*, \mu^*\big) = O(\log \log T).$$

In particular,

$$\text{Reg}_T\big(\Pi^{\text{SEJ}}; \alpha^*, \mu^*\big) \leq \left(\frac{1 - \mu^*}{\mu^*} + 1\right)(2 + \lceil \log_2 \log_2 T \rceil) + 1.$$

Proof. Let V^* denote the full-information benchmark value. As noted above, when $(\alpha^\star, \mu^*)$ is known, an optimal scheme is supported on $\{0, \nu^*\}$, so $V^* = \frac{\mu^*}{\nu^*} = \mu^* + (1 - \mu^*)m^*$.

Fix $m \in [0, 1]$ and consider π_m. The probability of the High signal is $p_H(m) = \mu^* + (1 - \mu^*)m$. By definition of m^*, the High recommendation is persuasive if and only if $m \leq m^*$. Hence the sender's one-round utility is $V(m) = \mu^* + (1 - \mu^*)m$ if $m \leq m^*$, and $V(m) = 0$ otherwise. Therefore the one-round regret is $r(m) = (1 - \mu^*)(m^* - m)$ if $m \leq m^*$, and $r(m) = V^*$ if $m > m^*$.

Viewing one test point m as an epoch that lasts until the first High realization, the epoch length is geometric with mean $\mathbb{E}[N(m)] = 1/p_H(m)$. Thus the expected regret of that epoch is $R_{\text{epoch}}(m) = r(m)/p_H(m)$. If $m \leq m^*$, then

$$R_{\text{epoch}}(m) = \frac{(1 - \mu^*)(m^* - m)}{\mu^* + (1 - \mu^*)m} \leq \frac{1 - \mu^*}{\mu^*}(m^* - m).$$

If $m > m^*$, then $p_H(m) \geq p_H(m^*) = V^*$, so

$$R_{\text{epoch}}(m) = \frac{V^*}{p_H(m)} \leq 1.$$

Now consider the exploration phase. The maintained interval $[a, b]$ always satisfies $a \leq m^* \leq b$: this is true initially since $[a, b] = [0, 1]$, and each update preserves it because acceptance is equivalent to $m \leq m^*$. Let L_k be the interval length at the start of phase k. Since $L_0 = 1$ and $\varepsilon_0 = 1/2$, after phase 0 we have $L_1 = 1/2$. For every $k \geq 1$, the new interval length equals the previous step size and the algorithm updates $\varepsilon \leftarrow \varepsilon^2$, so $\varepsilon_k = L_k^2$ and $L_{k+1} = L_k^2$. Hence $L_k = (1/2)^{2^{k-1}}$ for $k \geq 1$. Exploration stops once $L_k \leq T^{-1}$, so the number of phases is at most

$$P \leq 2 + \lceil \log_2 \log_2 T \rceil .$$

Fix one phase with interval $[a, b]$, length $L = b - a$, and step size ε. Let the accepted test points be $m_j = a + j\varepsilon$ for $j = 0, \ldots, J - 1$, where $J \leq L/\varepsilon$, and possibly there is one rejected point m_J. Then

$$\sum_{j=0}^{J-1} R_{\text{epoch}}(m_j) \leq \frac{1 - \mu^*}{\mu^*} \sum_{j=0}^{J-1} (m^* - m_j).$$

Since rejection first occurs at m_J, we have $m^* < a + J\varepsilon$, so $m^* - m_j \leq (J - j)\varepsilon$. Therefore

$$\sum_{j=0}^{J-1} (m^* - m_j) \leq \varepsilon \sum_{r=1}^{J} r = \varepsilon \frac{J(J+1)}{2} \leq \frac{L^2}{2\varepsilon} + \frac{L}{2}.$$

Hence the regret from accepted points in this phase is at most $\frac{1-\mu^*}{\mu^*} \left(\frac{L^2}{2\varepsilon} + \frac{L}{2} \right)$. For phase 0, this is at most $\frac{3}{2} \cdot \frac{1-\mu^*}{\mu^*}$. For every later phase, $\varepsilon = L^2$, so it is at most $\frac{1-\mu^*}{\mu^*} \left(\frac{1}{2} + \frac{L}{2} \right) \leq \frac{1-\mu^*}{\mu^*}$. Each phase has at most one rejected point, and its epoch regret is at most 1. Therefore

$$\begin{aligned} R_{\text{explore}} &\leq \frac{3}{2} \cdot \frac{1 - \mu^*}{\mu^*} + 1 + (P - 1) \left(\frac{1 - \mu^*}{\mu^*} + 1 \right) \\ &\leq \left(\frac{1 - \mu^*}{\mu^*} + 1 \right) P. \end{aligned}$$

In the commitment phase, the algorithm commits to $\widehat{m} = a$. By the interval invariant, $\widehat{m} \leq m^*$ and $m^* - \widehat{m} \leq b - a \leq 1/T$. Thus $\pi_{\widehat{m}}$ is persuasive, and its per-round regret is $(1 - \mu^*)(m^* - \widehat{m}) \leq (1 - \mu^*)/T \leq 1/T$. So the total commitment regret is at most 1.

Combining the two parts,

$$\begin{aligned} \text{Reg}_T\big(\Pi^{\text{SEJ}}; \alpha^*, \mu^*\big) &\leq \left(\frac{1 - \mu^*}{\mu^*} + 1 \right) P + 1 \\ &\leq \left(\frac{1 - \mu^*}{\mu^*} + 1 \right) (2 + \lceil \log_2 \log_2 T \rceil) + 1. \end{aligned}$$

This proves the proposition. □

F.2 General case with jointly unknown prior and bias

Proposition 12 (General-case regret lower bound for jointly unknown prior and bias)**.** *In the jointly unknown prior and bias model with finite state and action spaces, for any learning algorithm* Π*, there exists an instance such that*

$$\text{Reg}_T(\Pi; \alpha^*, \mu^*) = \Omega(\log T).$$

Proof. The jointly unknown model strictly contains the unknown-prior model studied in [30] as the special case $\alpha^* = 1$. Indeed, when $\alpha^* = 1$, the distorted posterior reduces to the Bayesian posterior.

Therefore any learning algorithm for the jointly unknown model induces a learning algorithm for the unknown-prior model. By the $\Omega(\log T)$ lower bound for the unknown-prior problem in [30], the same lower bound must hold for the jointly unknown model. □

Why the general jointly unknown case is difficult. Although the lower bound $\Omega(\log T)$ follows immediately from the unknown-prior model by restricting to the special case $\alpha^* = 1$, obtaining a matching $O(\log T)$ upper bound for the general jointly unknown case appears substantially harder. The main difficulties are as follows.

- **The identification step from unknown-prior model no longer decouples into pairwise prior-ratio estimation.**
  In the unknown-prior model of [30], the sender learns μ^* by constructing experiments in which the receiver's action reveals the sign of a linear expression involving only a pair of states. This makes it possible to estimate ratios such as $\mu^*(\omega_i)/\mu^*(\omega_j)$ and then reconstruct the prior. In the jointly unknown model, however, after a signal inducing Bayesian posterior ν, the receiver evaluates actions using the distorted belief $(1-\alpha^*)\mu^* + \alpha^*\nu$. Hence, for any two actions a, a', the receiver compares
  $$\sum_{\omega\in\Omega}\Big((1-\alpha^*)\mu^*(\omega) + \alpha^*\nu(\omega)\Big)\big(u_R(a,\omega) - u_R(a',\omega)\big).$$
  Since ν itself depends on μ^* through Bayes' rule, the sign of this expression depends jointly on the entire prior vector μ^* and on α^*. Therefore a single experiment no longer isolates a pairwise ratio of prior masses. In particular, the ratio-estimation argument from [30] cannot be applied directly.
- **The binary-action reduction from [30] does not extend directly once bias is also unknown.**
  In the binary-action setting of [30], the sender's optimization problem can be written as a fractional knapsack problem whose value-to-weight ordering is independent of the prior, because the common factor $\mu^*(\omega)$ cancels in the ratio. This allows [30] to define a universal one-parameter family $\{\pi_M\}$ and to characterize persuasion by a single threshold M^*. In the jointly unknown biased model, however, the persuasive constraint for recommending action 1 takes the form
  $$\sum_{\omega\in\Omega}\mu^*(\omega)x_\omega\Big(\alpha^* g_\omega + (1-\alpha^*)B(\mu^*)\Big) \le 0,$$
  where $x_\omega = \pi(1 \mid \omega)$, $g_\omega = u_R(0,\omega) - u_R(1,\omega)$, and $B(\mu^*) = \sum_{\omega\in\Omega}\mu^*(\omega)g_\omega$. Thus the effective "weight" of state ω is
  $$w_\omega(\alpha^*, \mu^*) = \alpha^* g_\omega + (1-\alpha^*)B(\mu^*),$$
  which depends jointly on α^* and on the full prior vector μ^*. Consequently, the state ordering relevant for the greedy argument is no longer fixed ex ante, and there is no direct analogue of the universal family $\{\pi_M\}$ used in [30]. Therefore the $O(\log\log T)$ binary-action argument of [30] does not extend mechanically to the jointly unknown setting.
- **The robustification step from the unknown-prior paper is not directly available under joint misspecification of prior and bias.**
  The $O(\log T)$ algorithm in [30] has an exploitation phase based on robustification: once the sender obtains an estimate $\widehat{\mu}$ of the true prior μ^*, she computes a signaling scheme that is optimal for $\widehat{\mu}$ and then modifies it so that it remains persuasive and near-optimal for all priors in a neighborhood of $\widehat{\mu}$. This argument is tailored to perturbations in the prior alone. In the jointly unknown model, however, persuasion is determined by distorted beliefs of the form $(1-\alpha)\mu + \alpha\nu$.
  Therefore, if the sender only knows an approximate pair $(\widehat{\mu}, \widehat{\alpha})$, then to carry out an exploitation phase one would need a *joint robustification* result: namely, a theorem showing that a scheme designed for $(\widehat{\mu}, \widehat{\alpha})$ can be modified so as to remain persuasive and near-optimal uniformly over all nearby pairs (μ, α).

For these reasons, the techniques of the two papers do not currently yield a general $O(\log T)$ algorithm. The lower bound $\Omega(\log T)$ is inherited immediately from the unknown-prior model, but the corresponding upper-bound techniques rely on tools that are specific either to prior uncertainty alone or to a one-dimensional threshold structure. In the general jointly unknown case, neither of these simplifications is currently available. Therefore, while an $O(\log T)$ algorithm may still exist, establishing such a result would require new ideas beyond the current identification and robustification arguments. For this reason, we view the general jointly unknown prior and bias problem as an open question.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction summarize the paper's scope: learning to persuade a receiver with an unknown linear belief-distortion parameter, via safe-exploration algorithms, regret upper bounds, a matching lower bound, and extensions. These claims are supported by the binary-case results in Section 3, the general-case results in Section 4, the extensions in Sections D to F, and the appendix proofs.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses important scope limitations and open directions in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper is primarily theoretical and states all relevant assumptions in Sections 2 and 4. Formal proofs and supporting lemmas for the main propositions and theorems are provided in Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer:[N/A]

Justification: The paper does not include experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We read and follow it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper motivates positive applications such as better calibrated warnings in Section 1. A potential negative impact is that the same techniques could be misused to manipulate receivers, so deployment should be accompanied by transparency, oversight, and safeguards against exploitative targeting.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.