Optimal Recalibration of an Online Predictor

Lunjia Hu[*] **Kevin Tian**[†] **Chutong Yang**[‡]

Abstract

We study the problem of recalibrating an online predictor [KE17, OKS24]: given an arbitrary "hint" sequence of forecasts, the learner must output new predictions that are calibrated while incurring small excess error relative to the original forecasts, under a proper loss. We give an online algorithm that achieves $(\varepsilon, \varepsilon^2)$-recalibration for Lipschitz proper losses in $T \approx \varepsilon^{-3}$ rounds, using an imbalanced extension of the recent simultaneous Blackwell approachability reduction framework of [HTY26]. We show that this tradeoff is optimal by proving a matching lower bound for recalibrating against the squared loss.

As our main application, we show how our recalibration algorithms can be combined with the online refinement method of [FH23] to obtain simultaneous ε-calibration and ε^2-calibeating for smooth proper losses at the same asymptotic rate, improving upon prior works that achieved these properties separately or with a worse ε dependence. In particular, the $\mathcal{K}_2$ variant answers a question of [CHJL26] on simultaneously achieving near-optimal calibeating and calibration rates. We also derive extensions to settings with multiple hint sequences. Finally, we empirically evaluate our algorithms on a classification dataset undergoing distribution shift.

1 Introduction

Probabilistic prediction is a central primitive in modern machine learning. In an online classification setting with binary labels, this problem is as follows: examples arrive over T timesteps, and for each $t \in [T]$ our goal is to output predictions $p_t \in [0, 1]$ (possibly depending on auxiliary example features) before seeing the label $y_t \in \{0, 1\}$. These predictions $p_{[T]} := \{p_t\}_{t\in[T]}$ are routinely interpreted as probabilities modeling the labels $y_{[T]} := \{y_t\}_{t\in[T]}$: an image classifier assigns probabilities to classes, a recommender predicts the probability of a click or purchase, and so on.

To measure the interpretability of predictions $p_{[T]}$ as a reflection of the labels $y_{[T]}$, calibration [Daw82, FV98] has arisen as a standard desideratum. For prediction-label pairs $p_{[T]}, y_{[T]}$, denote the average label and total count of a predicted value $v \in [0, 1]$ up to iteration $t \in [T]$ by

$$\bar{y}_t(p) := \frac{1}{n_t(p)} \sum_{s\in[t]} y_s \cdot \mathbb{I}_{p_s=p}, \qquad n_t(p) := \sum_{s\in[t]} \mathbb{I}_{p_s=p}. \tag{1}$$

with the convention $\bar{y}_t(v) = 0$ if $n_t(v) = 0$. Then the calibration error[4] is defined as follows.

[*]Northeastern University, lunjia@alumni.stanford.edu

[†]University of Texas at Austin, kjtian@cs.utexas.edu

[‡]University of Texas at Austin, cyang98@utexas.edu

[4]Definition 1 is sometimes called the *expected calibration error* (ECE) in the literature, to disambiguate with other calibration metrics. It is equivalent to ECE in a distributional or empirical setting. However, as it is the only metric we use in this paper, we simply call it the calibration error.

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

Definition 1 (Calibration error)**.** *For a sequence of predictor-label pairs $p_{[T]} \in [0,1]^T$, $y_{[T]} \in \{0,1\}^T$, we define the* calibration error *by*[5]

$$\mathcal{K}_1\left(p_{[T]}, y_{[T]}\right) := \sum_{v \in [0,1]} \frac{n_T(v)}{T} |v - \bar{y}_T(v)|. \tag{2}$$

Note that if $\mathcal{K}_1(p_{[T]}, y_{[T]}) = 0$, then $\bar{y}_T(v) = v$ for all predicted v; this reflects the interpretable condition that a v fraction of iterations predicting $p_t = v$ will have labels $y_t = v$. In this case, the predictions satisfy the classical definition of exact *calibration* [Daw82, FV98].

Oftentimes, calibration is only treated as a basic requirement, alongside other criteria such as *loss minimization* (e.g., as measured by the squared loss). Indeed, calibration and loss minimization are generally incomparable; a predictor can be extremely miscalibrated while achieving near-optimal loss minimization, and vice versa. In this paper, we study the problem of *recalibrating* an online predictor, which balances these two goals in a general *hinted sequential prediction* setting.

Model 1 (Hinted sequential prediction)**.** *In* hinted sequential prediction*, a learner observes a* hint *sequence $q_{[T]} \in [0,1]^T$, and outputs a* prediction *sequence $p_{[T]} \in [0,1]^T$ before observing a* label *sequence $y_{[T]} \in \{0,1\}^T$. We assume (q_t, y_t) can depend jointly on all previous $(q_{[t-1]}, p_{[t-1]}, y_{[t-1]})$, but not p_t; similarly, p_t can depend on $(q_{[t-1]}, p_{[t-1]}, y_{[t-1]})$ and q_t, but not y_t.*

This setting can also be viewed as a T-iteration game between nature and a learner. In the t^{th} iteration, nature selects a hint-label pair (q_t, y_t) (depending on the game's history) and reveals q_t to the learner. The learner then makes a prediction p_t, and then nature reveals y_t to the learner.

In Model 1, the hint $q_{[T]}$ can be viewed as an initial prediction, e.g., made by a model with good loss properties, but that does not necessarily satisfy calibration. This tension is motivated by [GPSW17, KCA+22], studies which demonstrate that modern neural networks trained with a loss minimization objective are not always well-calibrated. Thus, a natural goal under Model 1 is to observe the hint $q_{[T]}$, and output a calibrated sequence $p_{[T]}$ that achieves similar loss. This is the goal in the problem of *recalibration*, introduced by [KE17] and studied further by [OKS24].

Definition 2 (Recalibration)**.** *In the setting of Model 1, let $\alpha, \beta \geq 0$, and let $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$ be a loss. We say that (possibly randomized) $p_{[T]}$ is (α, β)-recalibrated against $(q_{[T]}, y_{[T]})$ under ℓ if*

$$\mathbb{E}\left[\mathcal{K}_1\left(p_{[T]}, y_{[T]}\right)\right] \leq \alpha, \quad \text{and} \quad \mathbb{E}\left[\ell\left(p_{[T]}, y_{[T]}\right) - \ell\left(q_{[T]}, y_{[T]}\right)\right] \leq \beta.$$

If only the first condition holds, we say that $p_{[T]}$ is α-calibrated against $y_{[T]}$.

In Definition 2, we overload notation: for a loss ℓ that takes prediction-label pairs (p, y) to a real value, we define its application to sequences by $\ell(p_{[T]}, y_{[T]}) := \frac{1}{T}\sum_{t \in [T]} \ell(p_t, y_t)$. In this introduction, we focus primarily on the *squared loss* $\ell(p, y) = \ell_{\mathsf{sq}}(p, y) := (p - y)^2$ for simplicity, although our main results apply to a much broader range of *proper losses* (cf. Section 2 for definitions).

1.1 Our results

Our main result is an optimal algorithm for recalibration with parameters $(\varepsilon, \varepsilon^2)$. In particular, we provide a matching lower bound, showing our required iteration complexity is tight.

Theorem 1 (Informal, see Theorems 4 and 8)**.** *If $T = \Omega(\varepsilon^{-3})$, there is an algorithm that achieves $(\varepsilon, \varepsilon^2)$-recalibration under ℓ_{sq} in the setting of Model 1, which is optimal up to constant factors.*

Our formal upper bound in Theorem 4 extends to all L-Lipschitz proper losses, with a polynomial dependence on L in the iteration count. In Table 1, we compare Theorem 1 with prior related results from [KE17, OKS24]. Our result improves known bounds on a question posed by [NRRX23, OKS24], on the achievable exponent pairs $(\alpha, \beta) = (T^{-a}, T^{-b})$ in recalibration. We also note that Theorem 4 extends straightforwardly to a high-probability guarantee, see Section B.3.

The horizon $T = \Omega(\varepsilon^{-3})$ in Theorem 1 is natural from the perspective of the simpler problem of *online calibration*, where the only goal is to achieve $\mathbb{E}[\mathcal{K}_1(p_{[T]}, y_{[T]})] \leq \varepsilon$. In this related setting, ε-calibration was known to be achievable in $T = \Omega(\varepsilon^{-3})$ iterations since [FV98], but improving

[5] Although the sum is over an infinite domain, only finitely many $v \in [0,1]$ have nonzero $n_T(v)$.

Work	Rate
[KE17]	$O\left(\frac{1}{\varepsilon^8}\right)$
[NRRX23, OKS24]	$O\left(\frac{1}{\varepsilon^4}\right)$
Theorems 4, 8	$\Theta\left(\frac{1}{\varepsilon^3}\right)$

Table 1: Rates for $(\varepsilon, \varepsilon^2)$-recalibration under ℓ_{sq}.

upon this tradeoff was open for nearly 30 years until a recent breakthrough of [DDF+25]. Our work shows that for the stronger goal of $(\varepsilon, \varepsilon^2)$-recalibration, taking a horizon of $T = \Omega(\varepsilon^{-3})$ is not only sufficient, it is also necessary, establishing a strict separation with online calibration.

Although our lower bound shows that no shorter horizon is possible for recalibration with parameters $(\varepsilon, \varepsilon^2)$, the excess loss requirement of ε^2 may seem artificial. Indeed, if calibration is not a consideration, achieving $\beta = 0$ in Definition 2 is trivial using $p_{[T]} = q_{[T]}$. Our next application, *calibeating*, motivates this particular $(\varepsilon, \varepsilon^2)$ tradeoff. This problem was introduced by [FH23] and expanded upon by [CHJL26], where the goal is stronger than merely not worsening the loss of a hint sequence $q_{[T]}$: instead, its loss must be improved up to the *refinement score* (3).

Definition 3 (Calibeating)**.** *In the setting of Model 1, let $\alpha \geq 0$, and let $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$ be a loss. We say that (possibly randomized) $p_{[T]}$ α-*calibeats *$(q_{[T]}, y_{[T]})$ under ℓ if*

$$\mathbb{E}\left[\ell\left(p_{[T]}, y_{[T]}\right)\right] \leq \mathbb{E}\left[\mathcal{R}_\ell\left(q_{[T]}, y_{[T]}\right)\right] + \alpha,$$

*where we define the ℓ-*refinement score *by*

$$\mathcal{R}_\ell\left(q_{[T]}, y_{[T]}\right) := \frac{1}{T}\sum_{t\in[T]} \ell\left(\bar{y}_T(q_t), y_t\right). \tag{3}$$

We note $\bar{y}_{[T]}$ is defined as Equation 1.

The refinement score (3) is the loss achieved by the *hindsight average predictor* $\bar{y}_T(q_t)$, which can observe the label sequence $y_{[T]}$ in advance and output corrected label averages. Calibeating, i.e., competing with $\bar{y}_T(q_t)$ in an online fashion, was motivated by the classical bias-variance decomposition of the Brier score (the loss ℓ_{sq}). Note that this benchmark predictor is infeasible to implement in the online setting of Model 1. Perhaps surprisingly, [FH23] showed that when $T = \Omega(\varepsilon^{-3})$ competing with the refinement score up to $\alpha = \varepsilon^2$ remains asymptotically achievable even in the *online setting* of Model 1, as long as $q_{[T]}$ takes on at most $O(\frac{1}{\varepsilon})$-distinct values.[6]

Interestingly, although the hindsight average predictor $\bar{y}_T(q_t)$ is perfectly-calibrated by definition, the online calibeating algorithm of [FH23] does not by itself guarantee calibration. Thus, the state-of-affairs prior to our work was that ε-calibration and ε^2-calibeating were achievable in isolation when $T = \Omega(\varepsilon^{-3})$, but not simultaneously. Our next result closes this gap.

Theorem 2 (Informal, see Theorem 5)**.** *If $T = \Omega(\varepsilon^{-3})$ and $q_{[T]}$ takes on $O(\frac{1}{\varepsilon})$ distinct values, there is an algorithm that achieves ε-calibration and ε^2-calibeating under ℓ_{sq} in the setting of Model 1.*

Theorem 2 is proven by combining a generalization of the main calibeating result of [FH23] (see Proposition 1) with the recalibration result in Theorem 1. It extends to a wider family of proper losses beyond ℓ_{sq}: those which are Lipschitz and smooth (Assumption 2). The formal variant states a polynomial dependence on the smoothness, as well as an explicit (linear) dependence on the sparsity of $q_{[T]}$. We also provide a comparison with prior related works in Table 2.[7]

Finally, in Appendix D, we show that both Theorems 1 and 2 generalize naturally to the setting of Model 2, a variant of Model 1 where there are m hint sequences, and the goal is to compete with the loss of the best hint sequence (in hindsight), or the appropriate refinement score extension. This problem setting is natural when a portfolio of models (e.g., with different hyperparameters) is used to

[6] We discuss this requirement further in Remark 3, which explains why nontrivial calibeating is impossible without a sparsity assumption on the hint sequence $q_{[T]}$. The regime $O(\frac{1}{\varepsilon})$ is natural for predictions on a grid of width ε.

[7] We use $\widetilde{O}$ to hide polylog$(\frac{1}{\varepsilon})$ factors.

Work	Rate	Multiple sequences?
[FH23]	$\widetilde{O}\left(\frac{1}{\varepsilon^4}\right)$	✗
[LNPR22]	$\widetilde{O}\left(\frac{1}{\varepsilon^6}\right)$	✓
[HLSS25]	$\widetilde{O}\left(\frac{1}{\varepsilon^6}\right)$	✓
[CHJL26]	$\widetilde{O}\left(\frac{1}{\varepsilon^4}\right)$	✓
Theorem 5	$\widetilde{O}\left(\frac{1}{\varepsilon^3}\right)$	✓

Table 2: Rates for ε-calibration and ε^2-calibeating under ℓ_{sq}, assuming $q_{[T]}$ is $O(\frac{1}{\varepsilon})$-sparse.

generate predictions. By using a reduction technique from [CHJL26] that we term *hint compression* (cf. Definition 8), we show in Theorems 6 and 7 that our algorithms for Theorems 4 and 5 extend to the multi-hinted setting at an additive cost of $T = \Omega(\frac{\log m}{\varepsilon^2})$. This property also holds for the prior works of [LNPR22, HLSS25, CHJL26], as indicated in Table 2.

2 Preliminaries

Notation. We denote vectors in boldface and scalars in plaintext. When the ambient dimension is clear from context, $\mathbf{e}_i$ denotes the i^{th} standard basis vector. We use $\mathbf{1}_d$ and $\mathbf{0}_d$ to denote the all-ones and all-zeroes vectors in $\mathbb{R}^d$. For $\mathbf{v} \in \mathbb{R}^d$ and $p \geq 1$, $\|\mathbf{v}\|_p$ denotes its ℓ_p norm. For $n \in \mathbb{N}$, we let $[n] := \{i \in \mathbb{N} \mid 1 \leq i \leq n\}$. We abbreviate a scalar sequence $\{p_t\}_{t\in[T]}$ by $p_{[T]}$ for shorthand; similarly, a vector sequence $\{\mathbf{v}_t\}_{t\in[T]}$ is denoted $\mathbf{v}_{[T]}$. The 0-1 indicator of an event $\mathcal{E}$ is denoted $\mathbb{I}_{\mathcal{E}}$, and $\mathsf{Bern}(p)$ is the Bernoulli distribution with mean p.

Our algorithms, particularly in Appendix A, use online regret minimization primitives as subroutines (see e.g., [Sha07] for background on this topic). In these subroutines, we use the following notation for the *Bregman divergence* in a convex function $\omega : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subseteq \mathbb{R}^d$:

$$D_\omega(\mathbf{x}'\|\mathbf{x}) := \omega(\mathbf{x}') - \omega(\mathbf{x}) - \langle \nabla\omega(\mathbf{x}), \mathbf{x}' - \mathbf{x}\rangle .$$

This is the residual of a first-order Taylor expansion from $\mathbf{x}$, and always satisfies $D_\omega(\mathbf{x}'\|\mathbf{x}) \geq 0$. For $\omega = -H$ and $\psi(p) = \ell(p, 0) - \ell(p, 1)$, propriety gives $\psi(p) \in \partial\omega(p)$.

Sequential prediction. Throughout, we study *hinted sequential prediction* tasks in binary classification, defined formally in Model 1. Informally, in applications captured by Model 1, the goal of the learner is to output predictions $p_{[T]}$ that capture the labels $y_{[T]}$ accurately, while improving the hint $q_{[T]}$ in calibration or loss minimization. We note that in Appendix D we extend Model 1 to a setting where nature provides multiple hints (Model 2), all of which must be competed against.

In our calibeating applications specifically (cf. Appendix C), we parameterize the hint sequence $q_{[T]}$ by its *sparsity*, as is consistent with the literature. More formally, we say a sequence of scalars $q_{[T]}$ is *s-sparse* if the sequence takes on at most s distinct values. We give a discussion on the necessity of this parameterization for our calibeating applications in Remark 3.

Loss minimization. Our goal will be to output a prediction sequence $p_{[T]}$ that competes with the input hint sequence $q_{[T]}$ with respect to a *loss* $\ell : [0, 1] \times \{0, 1\} \to \mathbb{R}$ that sends prediction-label pairs to scalar values. When ℓ is clear from context, we overload notation and define

$$\ell\left(p_{[T]}, y_{[T]}\right) := \frac{1}{T}\sum_{t\in[T]} \ell(p_t, y_t)$$

to be the average loss over the sequence. We focus on the following special family of losses.

Definition 4 (Proper loss)**.** *We say a loss $\ell : [0, 1] \times \{0, 1\} \to \mathbb{R}$ is* proper *if for all $p \in [0, 1]$,*

$$p \in \arg\min_{q\in[0,1]} \left\{\mathbb{E}_{y\sim\mathsf{Bern}(p)}\left[\ell(q, y)\right]\right\} .$$

Many common losses in supervised machine learning satisfy Definition 4, including the *squared loss* and *cross entropy (logistic)* loss, defined as:

$$\ell_{\mathsf{sq}}(p, y) := (p - y)^2, \quad \ell_{\mathsf{log}}(p, y) := (1 - y) \log\left(\frac{1}{1-p}\right) + y \log\left(\frac{1}{p}\right). \tag{4}$$

We also use the following assumption to parameterize our recalibration algorithms.

Assumption 1. *There exists $L \geq 1$ such that $\ell(\cdot, y)$ is L-Lipschitz for all $y \in \{0, 1\}$.*

Remark 1. *Our recalibration algorithm uses Assumption 1 in its MLOO construction (Lemma 4, following [OKS24]), and we leave it is an interesting open problem to characterize whether such an assumption is necessary for $(\varepsilon, \varepsilon^2)$-recalibration. Notably, the squared loss ℓ_{sq} is $L = 2$-Lipschitz, but the cross entropy loss ℓ_{log} is unbounded as $p \to 0$ (when $y = 1$) or $p \to 1$ (when $y = 0$). In practice, this issue could be alleviated by using a Lipschitz approximation to ℓ_{log}.*

3 Technical overview

In this section, we overview the proofs of both parts of Theorem 1. We then briefly survey its applications to calibeating (Theorem 5) and to multi-sequence generalizations (Theorems 6 and 7).

Recalibration as imbalanced simultaneous Blackwell approachability. Blackwell approachability [Bla56] is a classical strategy in online learning, and has a documented connection to regret minimization [ABH11] and online calibration [FV98, Fos99]. Our recalibration algorithm is based on encoding calibration and loss preservation as two simultaneous Blackwell approachability objectives. This general strategy of balancing approachability objectives was the explicit focus of the recent work [HTY26], and was implicit in several prior works, including [OKS24, OKK25]. Where our work differs is in introducing a variant of the [HTY26] simultaneous Blackwell approachability framework that explicitly encodes and benefits from asymmetry (Theorem 3). [OKS24] supplies the recalibration encoding and geometric oracle, while [OKK25, HTY26] contain the simultaneous approachability framework. Our contribution is recognizing that the [OKS24] geometric oracle implements the asymmetric MLOO required by our refinement of the [OKK25, HTY26] framework, and developing the imbalanced simultaneous approachability analysis that makes the $O(\varepsilon^{-4})$ to $O(\varepsilon^{-3})$ improvement possible (a nontrivial modification, even given the [OKS24] oracle). We also provide a matching lower bound and calibeating/multi-sequence consequences. Proving this imbalanced generalization of [HTY26] is the main focus of Appendix A.

Both calibration and proper loss regret can be concatenated into a single standard Blackwell target, but along this route, standard approachability controls both targets on a single accuracy scale. Setting the accuracy to ε^2 also controls the calibration coordinate at ε^2, leading to the previous $O(\varepsilon^{-4})$ horizon. Our imbalanced framework keeps separate tolerances $(\varepsilon, \varepsilon^2)$ by using a careful initialization and expert-wise guarantees. With $\eta = \varepsilon^2$, the calibration expert incurs $\eta + \frac{1}{\eta T}$ while the loss expert incurs only η; at $T = \Theta(\varepsilon^{-3})$, these are respectively $O(\varepsilon)$ and $O(\varepsilon^2)$.

We recall the encoding of recalibration as approachability, as stated in [OKS24]. The learner's action is a distribution $\mathbf{a} \in \Delta^{\mathcal{N}}$ over a fine grid $\mathcal{N} \subset [0, 1]$, which is later sampled to produce an ordinary scalar prediction. The first approachability objective is induced by the calibration vector

$$\mathbf{v}^{(1)}(\mathbf{a}, q, y) = \{\mathbf{a}_s(s - y)\}_{s \in \mathcal{N}}. \tag{5}$$

Our goal is to force this vector to approach the origin in $\mathbb{R}^{\mathcal{N}}$, by controlling its (time-averaged) ℓ_1 norm. Indeed, the ℓ_1 norm of the average $\mathbf{v}^{(1)}$ is exactly the calibration error (Definition 1).

The second approachability objective is induced by the excess loss

$$v^{(2)}(\mathbf{a}, q, y) = \sum_{s \in \mathcal{N}} \mathbf{a}_s(\ell(s, y) - \ell(q, y)), \tag{6}$$

whose average is the loss increase relative to the hint, which we want to approach 0. Thus $(\varepsilon, \varepsilon^2)$-recalibration is a two-objective approachability problem, but with different target accuracies.

Solving imbalanced simultaneous Blackwell approachability. We now explain the technical modifications to [HTY26] that our asymmetry-aware Theorem 3 makes, en route to proving the upper bound in Theorem 1. The main result of [HTY26] can be viewed as a reduction, from solving m simultaneous Blackwell approachability problems with coupled actions, to implementing m disjoint online learning algorithms and a certain oracle. This oracle, termed a *mixture linear optimization oracle* (MLOO, Definition 5), essentially asks to approximately satisfy m halfspace constraints *on average*, where each approachability instance induces one constraint.

The uniform definition of the MLOO in [HTY26] is incompatible with our asymmetric framing of recalibration. Therefore, in Definition 5, we generalize the definition to handle asymmetric tolerance parameters for the halfspace constraints, with the final tolerance being a weighted average. In Corollary 1, we show that this new oracle definition, carefully combined with an imbalanced mirror descent with two experts within the [HTY26] framework, yields a rate compatible with Theorem 1. More precisely, our mirror descent implementation with step size $\eta \approx \varepsilon^2$ has one expert pay an average regret of $\eta + \frac{1}{\eta T}$, but the other expert only pays η; this is important because only the calibration objective can afford a $\frac{1}{\eta T} = \varepsilon$ overhead when $T = \Omega(\varepsilon^{-3})$.

The remaining challenge is to implement the asymmetric MLOO for the two objectives (5), (6). Given a calibration distinguisher $\mathbf{u} \in [-1,1]^{\mathcal{N}}$, mixture weights (α, β), and the current hint q, the oracle must choose a distribution over grid points so that, simultaneously for both labels $y \in \{0,1\}$,

$$\alpha \sum_{s \in \mathcal{N}} \mathbf{a}_s u_s(s-y) + \beta \sum_{s \in \mathcal{N}} \mathbf{a}_s (\ell(s,y) - \ell(q,y)) = O(\alpha\varepsilon + \beta\varepsilon^2).$$

Fortunately, such an oracle was provided as Lemma 7 of [OKS24] for any $O(1)$-Lipschitz loss ℓ, although it was not stated that way in the original work. We recall the geometric oracle construction of [OKS24] in Lemma 4 in a way that is compatible with our approachability framework, and show it implements an $(\varepsilon, \varepsilon^2)$-MLOO in the sense of Definition 5. Finally, we put together the pieces (appropriate online learners, our MLOO, and Theorem 3) in Appendix B to prove Theorem 4.

Tightness of $(\varepsilon, \varepsilon^2)$-recalibration. Our lower bound is patterned off of a recent work by [CLNR26], but is much simpler.[8] The hard instance cycles through an ε-net inside $[\frac{1}{4}, \frac{3}{4}]$, sets q_t to the current grid value, and draws $y_t \sim \mathsf{Bern}(q_t)$. After $O(\varepsilon^{-2})$ total cycles, it is simple to see that outputting the "truthful predictor" $p_t = q_t$ at each iteration achieves ε-calibration and 0 excess loss in $T \approx \varepsilon^{-3}$ iterations. Our lower bound shows that no better strategy exists, even allowing ε^2 excess loss.

The proof first shows in Appendix E.1 that, up to constant factors in the recalibration guarantee, we may as well assume every p_t lies on the same ε-net used to define the instance. For a rounded value v, let ν_v be the number of times v is predicted, and let $\mu_v = \sum_{t:p_t=v}(q_t - y_t)$ be its stochastic fluctuation. We prove a bias-miscalibration decomposition in Lemma 13, which shows that calibration error must pay for $\sum_v |\mu_v|$, except for the amount of bias the algorithm introduces by moving away from q_t. This latter term is controlled by the excess squared loss bound implied by recalibration.

The rest of the proof lower bounds the empirical fluctuations $\sum_v |\mu_v|$. The key challenge is that the binning in these sums are defined in terms of our own rounded predictions $p_{[T]}$, rather than the original hint sequence $q_{[T]}$. We introduce a charging argument in Lemma 16 that shows the difference is negligible under an excess squared loss bound, concluding the proof.

Calibeating. In light of Theorem 4, to prove Theorem 5, our main result on simultaneously achieving calibration and calibeating, it is enough to design an online sequence that has excess loss ε^2 compared to the *refinement score* (3) of the hint $q_{[T]}$. In particular, online refinement produces $r_{[T]}$ which is $\frac{\varepsilon^2}{2}$-calibeating. We then pass $r_{[T]}$ to the recalibration algorithm, which adds at most $\frac{\varepsilon^2}{2}$ loss while guaranteeing ε-calibration. Combining therefore gives simultaneous ε-calibration and ε^2-calibeating. This is the strategy we take in Appendix C.

In the case of the squared loss ℓ_{sq}, [FH23] proves using a telescoping argument that playing the information-theoretically feasible *online prefix averages* $p_t = \bar{y}_{t-1}(q_t)$ achieves excess loss $O(\frac{s \log T}{T})$, when compared to the idealized hindsight predictor $p_t = \bar{y}_T(q_t)$ for an s-sparse sequence

[8] This lower bound construction was suggested by ChatGPT, and its proof was verified with presentation polished by the authors. The authors take full responsibility for the correctness of the proof.

$q_{[T]}$. This result is already enough to obtain our basic result in Theorem 2. For the more general variant, we show that when ℓ obeys a γ-smoothness condition (Assumption 2), we can substantially generalize the online-offline refinement gap, where we define the online refinement score

$$\widehat{\mathcal{R}}_\ell\left(q_{[T]}, y_{[T]}\right) := \frac{1}{T}\sum_{t\in[T]} \ell\left(\bar{y}_{t-1}(q_t), y_t\right). \tag{7}$$

We note $\bar{y}_{[T]}$ is defined (1). With this notation in hand, we state our generalization of Proposition 1, [FH23], which extends it to hold for any smooth proper loss. Our proof uses a decomposition of certain *Bregman divergences*, along with a similar telescoping trick as in [FH23].

Proposition 1. *In the setting of Model 1, if $q_{[T]}$ is s-sparse and Assumption 2 holds,*

$$0 \le \widehat{\mathcal{R}}_\ell\left(q_{[T]}, y_{[T]}\right) - \mathcal{R}_\ell\left(q_{[T]}, y_{[T]}\right) \le \frac{\gamma s\left(\log T + 1\right)}{2T}.$$

Combining Proposition 1 and Theorem 4 then directly yields Theorem 5.

Multiple sequences. In the case of recalibration or calibeating in the setting of Model 2, an m-sequence generalization of Model 1, our strategy is simple, and follows a similar argument as [CHJL26]. We define a problem called *hint compression* (Definition 8), which informally asks if the m hint sequences can be processed in an online fashion into a single sequence, which competes with the best of the m sequences up to average loss ε^2 in $T = O(\varepsilon^{-3})$ iterations. Fortunately, this problem is exactly equivalent to the well-studied problem of *online agnostic learning*.

Leveraging this equivalence, Appendix D.1 states some standard results on online agnostic learning as it pertains to hint compression in our setting. For the Lipschitz proper losses handled by Theorem 4, we use a classical framework called *mixability* (cf. [CBL06, RW10, vERW12]) to show that hint sequences are compressible with $O(\log m)$ excess regret, i.e., $O(\varepsilon^2)$ average regret when $T \gg \frac{\log m}{\varepsilon^2}$. This yields a tradeoff sufficient for our multi-sequence results, Theorems 6 and 7. In the more general setting (beyond Lipschitz losses), known lower bounds against hint compression rule out the stronger goal of $(\varepsilon, \varepsilon^2)$-multi-recalibration at the same $T = O(\varepsilon^{-3})$ rate, as we discuss in Appendix D.2.

3.1 Related work

To our knowledge, our main problem of recalibration (as stated in Definition 2) has only previously been formally studied in [KE17, OKS24]. Here we overview some closely-related literature.

Calibeating and calibration. Several other works [FH23, LNPR22, HLSS25, CHJL26] have considered simultaneous calibeating and calibration, as in our Theorem 2. Although we summarized their rates in Table 2, here we wish to mention for fair comparison that not all pursued our stated goal of ε-ℓ_1-calibration and ε^2-calibeating. In particular, some parameterize calibration differently [HLSS25, CHJL26], pursue a different tradeoff of $(\varepsilon, \varepsilon)$ [HLSS25], and/or are primarily motivated by the multi-sequence setting [LNPR22]. These works proceed using a diverse mix of strategies, notably including swap regret minimization primitives [HLSS25, CHJL26].

Online calibration. As discussed in Section 1.1, online calibration in its own right is a very well-studied problem in the online learning literature, inspired by practical problems like forecasting. Famously, [FV98] proved that ε-ℓ_2-calibration is achievable in $T = O(\varepsilon^{-3})$ iterations, and that $T = \Omega(\varepsilon^{-2})$ is necessary. Recent breakthroughs, first by [QV21] and subsequently by [DDF+25], have improved both bounds by constants in the exponent, although a gap persists. A related line [GJN+22, GJRR24, LSS25] pursues the stronger goal of *ε-multicalibration* [HKRR18] in an online setting, for which recent work by [CLNR26] proves a lower bound of $T = \Omega(\varepsilon^{-3})$ iterations. Their hard instance is the basis for our lower bound in Theorem 8.

Omniprediction. We briefly mention a recent line of work on *omniprediction* [GKR+22, GHK+23, HNRY23, GKR23, GJRR24, HTY25, DHI+25, OKK25, HTY26], a closely-related goal to recalibration as in Definition 2. In omniprediction, calibration is used as a means to achieve a simultaneous loss minimization guarantee, against a family of comparators. In our setting, the comparator is a single (or multiple) hint sequence(s), and our goal is to achieve a tight asymmetric tradeoff.

4 Experiments

In this section, we report on experiments comparing the loss minimization and calibration properties of four algorithms: the predictions made by a base neural network used to generate a hint sequence $q_{[T]}$ in Model 1, and three postprocessing methods applied to this hint sequence. Specifically, the four online forecasters we compare are as follows.

1. (Hint). The unmodified sigmoid output of the trained network, q_t.
2. (Calibeating). The algorithm in Section 4 of [FH23] and Appendix C of this work.
3. (Recalibration). Single hint recalibration via Blackwell approachability (Theorem 4).
4. (Calibeating + recalibration). The output of Item 2 is fed to an independent instance of Item 3, i.e., calibeating followed by recalibration.

We present their Brier score (31), ℓ_1 calibration error (2), and classification accuracy (zero-one loss). For the last metric, we threshold predictions at 0.5 to convert them to a label.

4.1 Binary classification

In this section, we consider a binary classification task. We use a binary subset of CIFAR-10 [KH+09] obtained by keeping only the cat and dog classes. This yields $11{,}000$ training images (of which $1{,}000$ are held out as a validation set) and $2{,}000$ test images, with cats relabeled 0 and dogs relabeled 1.

Hint model. As our hint model, we use a 40-layer DenseNet-BC with a single sigmoid output. The model is trained for 100 epochs with SGD with Nesterov momentum (initial learning rate 0.1, weight decay 10^{-4}, multistep schedule with drops at epochs 50 and 75) under the binary cross-entropy loss. The trained model attains 89.6% test accuracy.

Online stream and drift. The $2{,}000$ test images are reshuffled and replayed for 100 passes, producing an online stream of total length $T = 200{,}000$. At step $100{,}000$ (the midpoint, after 50 passes) the stream enters a "post-flip" phase. In this phase every label is independently flipped $y_t \leftarrow 1 - y_t$ with probability $\beta = 0.2$, otherwise kept. The model's per-image predictions are unchanged across the two phases, so the feature distribution and the forecaster's outputs are stationary while the conditional distribution $y \mid x$ changes. The flip is realized by an independent Bernoulli stream. This label drift is meant to model a distribution shift, which could simulate a predictor that is used to label both clean and noisy data. This is a reasonable setting to evaluate recalibration and calibeating, because the corrected sequence must address this miscalibration-inducing shift.

Table 3 reports cumulative metrics over all $T = 200{,}000$ stream rounds, with standard error across 20 runs with a fixed hint sequence. We report the Brier score, the ℓ_1 calibration error $\mathcal{K}_1$, and 0-1 accuracy. As theory predicts, calibeating decreases the Brier score, and empirically improves upon $\mathcal{K}_1$ as well. Interestingly, recalibration also improved the Brier score slightly. Finally, calibeating + recalibration achieved the best Brier score, and was competitive by other metrics.

Predictor	Brier	$\mathcal{K}_1$	Acc
Hint	$0.16124 \pm 7.8\text{e}{-4}$	$0.15413 \pm 8.4\text{e}{-4}$	$82.075 \pm 8.2\text{e}{-2}\%$
Calibeating	$0.13865 \pm 5.4\text{e}{-4}$	$0.05771 \pm 5.1\text{e}{-4}$	$82.428 \pm 8.4\text{e}{-2}\%$
Recalibration	$0.14708 \pm 8.3\text{e}{-4}$	$0.01188 \pm 2.5\text{e}{-4}$	$78.597 \pm 3.08\text{e}{-1}\%$
Calibeating + Recalibration	$0.13314 \pm 4.5\text{e}{-4}$	$0.01319 \pm 4.5\text{e}{-4}$	$82.408 \pm 8.8\text{e}{-2}\%$

Table 3: CIFAR-2 cats vs. dogs classification with partial concept drift (flip probability $\beta = 0.2$ after step $100{,}000$ with $T = 200{,}000$ for 20 random seeds). Cumulative metrics over the full stream. Hint and Calibeating are deterministic, and the randomness comes from random flipping and data order.

4.2 Multiclass classification

We include a multiclass experiment, for a classification task with k classes. To generate sequences $q_{[T]}, y_{[T]}$ for use in Model 1, we use two projection operations as suggested by [GPSW17]. In

particular, start from a hint vector $\mathbf{q} \in \Delta^k$ and a one-hot label $\mathbf{y} = \mathbf{e}_j \in \Delta^k$. We produce a pair $q = \mathbf{q}_i$, where $i := \arg\max_{i \in [k]} \{\mathbf{q}_i\}$ (breaking ties arbitrarily), and $y = \mathbb{I}_{i=j}$. This is the same projection as is implicitly performed in *confidence calibration*, where the goal is a top-class probability of q should mean this label is correct a q fraction of the time.

Next, suppose that our algorithm (calibeating, recalibration, or calibeating + recalibration) processes the scalar hint q and produces a label p. We lift this back up to a vector $\mathbf{p}$ as follows: we take $\mathbf{p}_i \leftarrow p$, and set the other coordinates $\mathbf{p}_{-i}$ to be the unique scaling of $\mathbf{q}_{-i}$ so that $\mathbf{p} \in \Delta^k$. When evaluating this new prediction $\mathbf{p}$ according to calibration ($\mathcal{K}_1$) or Brier score ($\mathcal{B}_2$), we freshly perform the binary projection operation, i.e., we relabel $p = \mathbf{p}_i$ where $i = \arg\max_{i \in [k]} \mathbf{p}_i$, and $y = \mathbb{I}_{i=j}$. Due to the nonlinear nature of this operation (and the inconsistent binary labels), our theory does not technically apply to this setting. Nevertheless, we include an empirical evaluation of the performance of our postprocessing operations in this multiclass setting.

Dataset. We use CIFAR-100 [KH+09], a 100-class image classification dataset on 32×32 RGB images. We use the standard splits: 45,000 training images (after holding out 5,000 for validation), and 10,000 test images. Test-set marginal class distribution is uniform. The 10,000 test images are reshuffled and replayed for 100 passes, producing an online stream of total length $T = 1{,}000{,}000$.

Hint model. As our hint model, we use a 40-layer DenseNet-BC with a 100-way softmax output. The model is trained for 300 epochs by SGD with Nesterov momentum (initial learning rate 0.1, weight decay 10^{-4}, multistep schedule with drops at 50% and 75% of training) under cross-entropy loss. The trained model attains 65.5% test accuracy.

Table 4 reports cumulative metrics over all $T = 1{,}000{,}000$ stream rounds, again with standard errors over 20 runs. We report the (projected) Brier score $\mathcal{B}_2$, the ℓ_1 calibration error $\mathcal{K}_1$, and top-1 accuracy. Despite the lack of direct theoretical guarantees, the recalibration method substantially improves the calibration, and calibeating + recalibration reduces the empirical calibration of calibeating alone. Interestingly, despite having the worst Brier score, recalibration had the best top-1 accuracy.

Predictor	Brier	$\mathcal{K}_1$	Acc
Hint	0.55862	0.23330	65.51%
Calibeating	$0.51501 \pm 1.3\text{e}{-4}$	$0.08476 \pm 2.6\text{e}{-4}$	$62.114 \pm 2.3\text{e}{-2}\%$
Recalibration	$0.55281 \pm 1.8\text{e}{-4}$	$0.03107 \pm 1.05\text{e}{-3}$	$64.379 \pm 5.1\text{e}{-2}\%$
Calibeating + Recalibration	$0.51240 \pm 1.4\text{e}{-4}$	$0.06622 \pm 7.1\text{e}{-4}$	$63.549 \pm 3.9\text{e}{-2}\%$

Table 4: CIFAR-100 multi-class with $T = 1{,}000{,}000$ for 20 random seeds. Calibeating is deterministic and the randomness comes from different orders on online data.

Numerical validation of the convergence rate. To validate the convergence rate as a function of T, the number of iterations, we ran the following convergence experiment on the pre-drift portion of the 200,000-round CIFAR-10 cat-vs-dog stream. The results below are averaged over 20 random seeds. The theoretical reference columns are anchored to the observed values at $T = 10{,}000$.

Rounds T	Mean $\mathcal{K}_2$	$T^{-1/3}$ reference	Mean calibeating gap	$T^{-2/3}$ reference
10,000	0.067580	0.067580	0.001986	0.001986
20,000	0.053549	0.053640	0.001100	0.001251
30,000	0.048584	0.046860	0.000775	0.000955
40,000	0.044750	0.042580	0.000600	0.000788
50,000	0.041521	0.039520	0.000491	0.000679
60,000	0.038812	0.037200	0.000417	0.000601
70,000	0.036361	0.035350	0.000362	0.000543
80,000	0.034299	0.033790	0.000320	0.000497
90,000	0.032405	0.032490	0.000287	0.000460
100,000	0.030767	0.031370	0.000260	0.000428

A log–log regression of mean $\mathcal{K}_2$ against T gives an empirical slope of -0.3331, essentially matching the theoretical slope $-1/3$. The mean calibeating gap has empirical slope -0.8852, which is steeper than the required $-2/3$ rate.

References

[ABH11] Jacob D. Abernethy, Peter L. Bartlett, and Elad Hazan. Blackwell approachability and no-regret learning are equivalent. In *COLT 2011 - The 24th Annual Conference on Learning Theory*, volume 19 of *JMLR Proceedings*, pages 27–46. JMLR.org, 2011.

[Bla56] David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956.

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[CHJL26] Yurong Chen, Zhiyi Huang, Michael I Jordan, and Haipeng Luo. Calibeating made simple. *arXiv preprint arXiv:2603.22167*, 2026.

[CLNR26] Natalie Collina, Jiuyao Lu, Georgy Noarov, and Aaron Roth. Optimal lower bounds for online multicalibration. *CoRR*, abs/2601.05245, 2026.

[Daw82] A Philip Dawid. The well-calibrated bayesian. *Journal of the American statistical Association*, 77(379):605–610, 1982.

[DDF+25] Yuval Dagan, Constantinos Daskalakis, Maxwell Fishelson, Noah Golowich, Robert Kleinberg, and Princewill Okoroafor. Breaking the tˆ(2/3) barrier for sequential calibration. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 2007–2018, 2025.

[DHI+25] Cynthia Dwork, Chris Hays, Nicole Immorlica, Juan C. Perdomo, and Pranay Tankala. From fairness to infinity: Outcome-indistinguishable (omni)prediction in evolving graphs. In *The Thirty Eighth Annual Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 1564–1637. PMLR, 2025.

[FH23] Dean P Foster and Sergiu Hart. "calibeating": Beating forecasters at their own game. *Theoretical Economics*, 18(4):1441–1474, 2023.

[Fos99] Dean P Foster. A proof of calibration via blackwell's approachability theorem. *Games and Economic Behavior*, 29(1-2):73–78, 1999.

[FV98] Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

[GHK+23] Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. In *14th Innovations in Theoretical Computer Science Conference, ITCS 2023*, volume 251 of *LIPIcs*, pages 60:1–60:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023.

[GJN+22] Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M. Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. In *13th Innovations in Theoretical Computer Science Conference, ITCS 2022*, LIPIcs, pages 82:1–82:24. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

[GJRR24] Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. Oracle efficient online multicalibration and omniprediction. In *Proceedings of the 2024 ACM-SIAM Symposium on Discrete Algorithms, SODA 2024*, pages 2725–2792. SIAM, 2024.

[GKR+22] Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference, ITCS 2022*, volume 215 of *LIPIcs*, pages 79:1–79:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

[GKR23] Parikshit Gopalan, Michael P. Kim, and Omer Reingold. Swap agnostic learning, or characterizing omniprediction via multicalibration. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, 2023.

[GPSW17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, Proceedings of Machine Learning Research, pages 1321–1330. PMLR, 2017.

[GR07] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[HKRR18] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, Proceedings of Machine Learning Research, pages 1944–1953. PMLR, 2018.

[HLSS25] Lunjia Hu, Haipeng Luo, Spandan Senapati, and Vatsal Sharan. Efficient swap multicalibration of elicitable properties. *arXiv preprint arXiv:2511.04907*, 2025.

[HNRY23] Lunjia Hu, Inbal Rachel Livni Navon, Omer Reingold, and Chutong Yang. Omnipredictors for constrained optimization. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 13497–13527. PMLR, 2023.

[HTY25] Lunjia Hu, Kevin Tian, and Chutong Yang. Omnipredicting single-index models with multi-index models. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 1762–1773, 2025.

[HTY26] Lunjia Hu, Kevin Tian, and Chutong Yang. Simultaneous blackwell approachability and applications to multiclass omniprediction. *arXiv preprint arXiv:2602.17577*, 2026.

[KCA+22] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[KE17] Volodymyr Kuleshov and Stefano Ermon. Estimating uncertainty online against an adversary. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 2110–2116. AAAI Press, 2017.

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[LNPR22] Daniel Lee, Georgy Noarov, Mallesh Pai, and Aaron Roth. Online minimax multiobjective optimization: Multicalibeating and other applications. *Advances in Neural Information Processing Systems*, 35:29051–29063, 2022.

[LSS25] Haipeng Luo, Spandan Senapati, and Vatsal Sharan. Improved bounds for swap multicalibration and swap omniprediction. *CoRR*, abs/2505.20885, 2025.

[Nes03] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course, volume I*. 2003.

[NRRX23] Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional prediction for sequential decision making. *arXiv preprint arXiv:2310.17651*, 2023.

[OKK25] Princewill Okoroafor, Robert Kleinberg, and Michael P. Kim. Near-optimal algorithms for omniprediction. *CoRR*, abs/2501.17205, 2025.

[OKS24] Princewill Okoroafor, Bobby Kleinberg, and Wen Sun. Faster recalibration of an online predictor via approachability. In *International Conference on Artificial Intelligence and Statistics*, pages 4690–4698. PMLR, 2024.

[QV21] Mingda Qiao and Gregory Valiant. Stronger calibration lower bounds via sidestepping. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 456–466, 2021.

[RW10] Mark D. Reid and Robert C. Williamson. Composite binary losses. *J. Mach. Learn. Res.*, 11:2387–2422, 2010.

[Sav71] Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[Sha07] Shai Shalev-Shwartz. *Online learning: theory, algorithms and applications*. PhD thesis, Hebrew University of Jerusalem, Israel, 2007.

[vERW12] Tim van Erven, Mark D. Reid, and Robert C. Williamson. Mixability is bayes risk curvature relative to log loss. *Journal of Machine Learning Research*, 13(52):1639–1663, 2012.

A Imbalanced Simultaneous Blackwell Approachability

In this section, we give a framework that is an *imbalanced* generalization of the main result of [HTY26]. That work defined a *simultaneous Blackwell approachability* problem (Problem 1), a multi-set generalization of the classical Blackwell approachibility problem and its application in online learning [Bla56, ABH11]. We recall the relevant problem statement of [HTY26] and develop some preliminaries in Appendix A.1. We then give our imbalanced generalization in Appendix A.2, and derive two basic consequences of it that see direct use in our applications.

A.1 Setup

Simultaneous Blackwell approachability. We state a variant of the main problem studied by [HTY26] in Problem 1. Here we provide a brief informal description of our variant.

In this problem, there are m vector-valued functions $\mathbf{v}^{(i)}$ each with three arguments: an action $\mathbf{a} \in \mathcal{A}$ chosen by the learner, and a context $q \in \mathcal{Q}$ and response $y \in \mathcal{Y}$ chosen by nature. There are also m sets $\mathcal{V}^{(i)}$, the i^{th} of which is implicitly defined by a "distinguisher set" $\mathcal{U}^{(i)}$ (formally, see (8)). The goal is to choose a sequence $\mathbf{a}_{[T]} \in \mathcal{A}^T$ such that for any sequences $q_{[T]} \in \mathcal{Q}^T$, $y_{[T]} \in \mathcal{Y}^T$, each average $\frac{1}{T}\sum_{t\in[T]} \mathbf{v}^{(i)}(\mathbf{a}_t, q_t, y_t)$ approximately lies in the i^{th} set $\mathcal{V}^{(i)}$. When $m = 1$, this is precisely (a contextual variant of) the classical Blackwell approachabilty as derived in [ABH11]. This special case is reviewed in Section 3.1, [HTY26], see also Problem 2 of that work.

Problem 1 (Imbalanced simultaneous Blackwell approachability)**.** *Let $a, m \in \mathbb{N}$, let $\mathcal{A} \subseteq \mathbb{R}^a$ and $\mathcal{Q}, \mathcal{Y} \subseteq \mathbb{R}$ be compact and convex, and let $\boldsymbol{\varepsilon} := \{\varepsilon^{(i)}\}_{i\in[m]} \in \mathbb{R}^m_{\geq 0}$. For all $i \in [m]$, let $\mathcal{U}^{(i)} \subseteq \mathcal{H}^{(i)}$ where $\mathcal{H}^{(i)}$ is a Hilbert space, and define $\mathcal{V}^{(i)} := \{\mathbf{v} \in \mathcal{H}^{(i)} \mid \sup_{\mathbf{u}\in\mathcal{U}^{(i)}} \langle \mathbf{u}, \mathbf{v}\rangle \leq 0\}$. Let $\mathbf{v}^{(i)} : \mathcal{A} \times \mathcal{Q} \times \mathcal{Y} \to \mathcal{H}^{(i)}$ be linear in its first argument, for all $i \in [m]$. Our goal is to observe online sequences $q_{[T]} \in \mathcal{Q}^T$, $y_{[T]} \in \mathcal{Y}^T$ and to choose $\mathbf{a}_{[T]} \in \mathcal{A}^T$ so that $\mathbf{a}_t$ depends on $q_{[t]}$ and $y_{[t-1]}$, and*

$$\sup_{\mathbf{u}^{(i)}\in\mathcal{U}^{(i)}} \left\langle \mathbf{u}^{(i)}, \frac{1}{T}\sum_{t\in[T]} \mathbf{v}^{(i)}(\mathbf{a}_t, q_t, y_t) \right\rangle \leq \varepsilon^{(i)}. \tag{8}$$

We highlight a few distinctions of Problem 1 when compared to previous versions.

First, some works require the $\mathbf{v}^{(i)}$ to bilinear functions of its inputs, to leverage certain equivalences between satisfiability notions. However, linearity in the second input is not needed in our applications. Further, Problem 1 as stated in [HTY26] had all $\varepsilon^{(i)} = \varepsilon$. Our applications use an imbalanced variant: for example, Appendix B sets $(\varepsilon^{(1)}, \varepsilon^{(2)}) = (\varepsilon, \varepsilon^2)$ for recalibration (Definition 2).

The main challenge in Problem 1 is to satisfy all m of the conditions (8) at once, by using a coupled sequence of actions $\mathbf{a}_{[T]}$. Theorem 2 of [HTY26] gives a sufficient condition for Problem 1 to be solvable: each distinguisher set $\mathcal{U}^{(i)}$ must admit a *no-regret learner*, and the m sets to be approached $\mathcal{V}^{(i)}$ must jointly satisfy a certain oracle requirement, which we now give.

Definition 5 (Mixture linear optimization oracle)**.** *In the setting of Problem 1, we call $\mathcal{O}$ an $\boldsymbol{\varepsilon}$-*mixture linear optimization oracle *(MLOO) if on inputs $\mathbf{w} \in \Delta^m$, $\{\mathbf{u}^{(i)}\}_{i\in[m]} \in \prod_{i\in[m]} \mathcal{U}^{(i)}$, and $q \in \mathcal{Q}$, the oracle outputs $\mathbf{a} \in \mathcal{A}$ satisfying*

$$\sum_{i\in[m]} \mathbf{w}_i \left(\left\langle \mathbf{u}^{(i)}, \mathbf{v}^{(i)}(\mathbf{a}, q, y) \right\rangle - \varepsilon^{(i)} \right) \le 0 \text{ for all } y \in \mathcal{Y}. \tag{9}$$

Definition 5 has an intuitive interpretation when specialized to a point mass $\mathbf{w} = \mathbf{e}_i$. In this case, the oracle requirement becomes: find $\mathbf{a}$ such that

$$\left\langle \mathbf{u}^{(i)}, \mathbf{v}^{(i)}(\mathbf{a}, q, y) \right\rangle \le \varepsilon^{(i)} \text{ for all } y \in \mathcal{B}. \tag{10}$$

As an example, this condition is attainable (including the exact case $\boldsymbol{\varepsilon} = \mathbf{0}_2$) whenever $\mathcal{V}^{(i)}$ is "approachable" in the sense of Blackwell (see Definition 2, [ABH11]). This is because [Bla56] shows approachability is equivalent to *halfspace satisfiability*, i.e., (10) with $\varepsilon^{(i)} = 0$. This relationship is discussed in detail by Theorem 3 and Lemma 1, [ABH11].

Definition 5 asks that the inequalities (10) can be satisfied *on average* as specified by any mixture weights $\mathbf{w} \in \Delta^m$, allowing for slacks given by $\boldsymbol{\varepsilon}$. In general, this is a strictly stronger requirement than asking each $\mathcal{V}^{(i)}$ to be approachable in isolation (see a counterexample in Lemma 6, [HTY26]). Consequently, solving (8) can be strictly harder than running m independent approachability procedures, due to the requirement of coupled actions. In Theorem 3, we derive a new sufficient condition for solving Problem 1, specialized to the imbalanced setting.

Online learning. We summarize a standard result from online learning and stochastic convex optimization, for use in proving Theorem 3 and our later applications.

Lemma 1 (Theorem 4.2, [Bub15])**.** *Fix $T \in \mathbb{N}$, $\delta \in (0, 1)$, and a set $\mathcal{X} \subseteq \mathbb{R}^d$ of diameter at most R under a norm $\|\cdot\|$. Let $r : \mathcal{X} \to \mathbb{R}$ be 1-strongly convex in $\|\cdot\|$,*[9] *initialize at $\mathbf{x}_1 \in \mathcal{X}$, and for each round $t \in [T]$, let $\mathbf{g}_t$ be a vector with $\|\mathbf{g}_t\|_* \le L$, and define a sequence of iterates by*

$$\mathbf{x}_{t+1} \leftarrow \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \eta \mathbf{g}_t - \nabla r(\mathbf{x}_t), \mathbf{x} \rangle + r(\mathbf{x}) \right\}, \text{ for all } t \in [T].$$

Then there exists a choice of η such that, for all $\mathbf{x} \in \mathcal{X}$,

$$\sum_{t\in[T]} \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x} \rangle \le \frac{\eta L^2}{2} + \frac{D_r(\mathbf{x}\|\mathbf{x}_1)}{\eta T}.$$

In what follows, we only apply Lemma 1 in the standard regime of quadratic regularization over a Euclidean ball of radius R (projected gradient descent). These applications use

$$r(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2, \quad D_r(\mathbf{x}\|\mathbf{x}') = \frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|_2^2, \quad \|\cdot\| = \|\cdot\|_* = \|\cdot\|_2. \tag{11}$$

A.2 Main framework

We are now ready to state our main framework in Algorithm 1, and analyze it in Theorem 3.

Theorem 3 (Imbalanced simultaneous Blackwell approachability)**.** *In the setting of Problem 1, let $\mathcal{O}$ be an $\boldsymbol{\varepsilon}$-MLOO. Also assume that for every $i \in [m]$ and horizon $T \in \mathbb{N}$, there is an online learner $\mathsf{alg}^{(i)}$ which, on inputs $(\mathbf{a}_{[T]}, q_{[T]}, y_{[T]}) \in \mathcal{A}^T \times \mathcal{Q}^T \times \mathcal{Y}^T$, outputs $\mathbf{u}^{(i)}_{[T]}$ with each $\mathbf{u}^{(i)}_t \in \mathcal{U}^{(i)}$ depending only on $\mathbf{a}_{[t-1]}$, $q_{[t-1]}$ and $y_{[t-1]}$, and satisfies*

$$\sup_{\mathbf{u}^{(i)}\in\mathcal{U}^{(i)}} \sum_{t\in[T]} \left\langle \mathbf{v}^{(i)}(\mathbf{a}_t, q_t, y_t), \mathbf{u}^{(i)} - \mathbf{u}^{(i)}_t \right\rangle \le \mathrm{reg}^{(i)}(T), \tag{12}$$

for some $\mathrm{reg}^{(i)} : \mathbb{N} \to \mathbb{R}_{\ge 0}$. Finally let $\eta > 0$ and suppose

$$\left| \left\langle \mathbf{v}^{(i)}(\mathbf{a}, q, y), \mathbf{u}^{(i)} \right\rangle \right| \le L, \quad \varepsilon^{(i)} \le L, \tag{13}$$

[9] This means $D_r(\mathbf{x}'\|\mathbf{x}) \ge \frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|^2$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$.

Algorithm 1: ImbalancedSimultaneousApproach$(q_{[T]}, y_{[T]}, \{\mathsf{alg}^{(i)}\}_{i\in[m]}, \mathcal{O}, \mathbf{w}_1, r, \eta)$

1 **Input:** Sequences $q_{[T]} \in \mathcal{Q}^T$, $y_{[T]} \in \mathcal{Y}^T$, online learners $\{\mathsf{alg}^{(i)}\}_{i\in[m]}$ satisfying (12), ε-MLOO $\mathcal{O}$ (following notation in Problem 1, Definition 5), initial $\mathbf{w}_1 \in \Delta^m$, convex $r : \Delta^m \to \mathbb{R}$, $\eta > 0$

2 **Output:** $\mathbf{p}_{[T]} \in \mathcal{A}^T$ such that each $\mathbf{p}_t$ is output after observing q_t and before observing y_t

3 $\mathbf{u}_1^{(i)} \leftarrow \mathsf{alg}^{(i)}(\{\})$ for all $i \in [m]$

`// Initialize each` $\mathbf{u}_1^{(i)}$ `as` $\mathsf{alg}^{(i)}$ `does before observing any examples.`

4 $\mathbf{a}_1 \leftarrow \mathcal{O}(\mathbf{w}_1, \{\mathbf{u}_1^{(i)}\}_{i\in[m]}, q_1)$

5 **for** $2 \le t \le T$ **do**

6 $\quad \mathbf{p}_{t-1} \leftarrow$ any random element of $\mathcal{A}$ such that $\mathbb{E}[\mathbf{p}_{t-1} \mid \mathbf{a}_{t-1}] = \mathbf{a}_{t-1}$

7 $\quad \mathbf{v}_{t-1}^{(i)} \leftarrow \mathbf{v}^{(i)}(\mathbf{p}_{t-1}, q_{t-1}, y_{t-1})$ for all $i \in [m]$

8 $\quad \tilde{\mathbf{g}}_{t-1} \leftarrow$ vector in $\mathbb{R}^m$ such that $[\tilde{\mathbf{g}}_{t-1}]_i = \langle \mathbf{u}_{t-1}^{(i)}, \mathbf{v}_{t-1}^{(i)} \rangle - \varepsilon^{(i)}$ for all $i \in [m]$

9 $\quad \mathbf{u}_t^{(i)} \leftarrow \mathsf{alg}^{(i)}(\mathbf{p}_{[t-1]}, q_{[t-1]}, y_{[t-1]})$ for all $i \in [m]$

10 $\quad \mathbf{w}_t \leftarrow \arg\min_{\mathbf{w}\in\Delta^m} \{-\langle \eta \mathbf{g}_{t-1} + \nabla r(\mathbf{w}_{t-1}), \mathbf{w}\rangle + r(\mathbf{w})\}$

11 $\quad \mathbf{a}_t \leftarrow \mathcal{O}(\mathbf{w}_t, \{\mathbf{u}_t^{(i)}\}_{i\in[m]}, q_t)$

12 **end**

13 **Return:** $\mathbf{a}_{[T]}$

for all $i \in [m]$, $(\mathbf{a}, q, y) \in \mathcal{A} \times \mathcal{Q} \times \mathcal{Y}$, and $\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}$. Then, for any sequences $q_{[T]} \in \mathcal{Q}^T$, $y_{[T]} \in \mathcal{Y}^T$, Algorithm 1 produces $\mathbf{p}_{[T]} \in \mathcal{A}^T$ such that $\mathbf{p}_t$ depends only on $q_{[t]}$, $y_{[t-1]}$, and

$$\mathbb{E}\left[\sup_{\mathbf{u}^{(i)}\in\mathcal{U}^{(i)}} \left\langle \mathbf{u}^{(i)}, \frac{1}{T}\sum_{t\in[T]} \mathbf{v}^{(i)}(\mathbf{p}_t, q_t, y_t)\right\rangle\right] \le \varepsilon^{(i)} + \frac{\mathrm{reg}^{(i)}(T)}{T} + 2\eta L^2 + \frac{D_r(\mathbf{e}_i\|\mathbf{w}_1)}{\eta T}, \quad (14)$$

for all $i \in [m]$, where $r : \Delta^m \to \mathbb{R}$ is the convex function used to implement Line 10.

Proof. We adopt the notation of Algorithm 1 in this proof. Note that $\mathbf{a}_t$ is chosen on Line 11 before y_t is revealed, and $\mathbf{p}_t$ only depends on $\mathbf{a}_t$. We also let $\mathbf{g}_t \in \mathbb{R}^m$ satisfy

$$[\mathbf{g}_t]_i = \left\langle \mathbf{u}_t^{(i)}, \mathbf{v}^{(i)}(\mathbf{a}_t, q_t, y_t)\right\rangle - \varepsilon^{(i)} \text{ for all } i \in [m],$$

so that $\mathbb{E}[\tilde{\mathbf{g}}_t \mid \tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_{t-1}] = \mathbf{g}_t$ by bilinearity of each $\mathbf{v}^{(i)}$ in its first argument.

The first step is to prove that for all $i \in [m]$,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} \left(\left\langle \mathbf{u}_t^{(i)}, \mathbf{v}_t^{(i)}\right\rangle - \varepsilon^{(i)}\right)\right] \le 2\eta L^2 + \frac{D_r(\mathbf{e}_i\|\mathbf{w}_1)}{\eta T}. \quad (15)$$

Indeed, this follows from plugging $\mathbf{w} \leftarrow \mathbf{e}_i$ into

$$\mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} \langle \tilde{\mathbf{g}}_t, \mathbf{w}\rangle\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} \langle \tilde{\mathbf{g}}_t, \mathbf{w}_t\rangle\right] + \mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} \langle \tilde{\mathbf{g}}_t, \mathbf{w} - \mathbf{w}_t\rangle\right] \le 2\eta L^2 + \frac{D_r(\mathbf{w}\|\mathbf{w}_1)}{\eta T}. \quad (16)$$

In the only inequality, we bounded each $\mathbb{E}[\langle \mathbf{w}_t, \tilde{\mathbf{g}}_t\rangle] = \langle \mathbf{w}_t, \mathbf{g}_t\rangle \le 0$ using the oracle condition (9), which is valid for every q_t, y_t defining $\mathbf{g}_t$. Here, the claimed equality holds because $\mathbb{E}[\tilde{\mathbf{g}}_t \mid \mathbf{w}_t] = \mathbf{g}_t$. We also applied the regret bound from Lemma 1 to the sequence of $\tilde{\mathbf{g}}_{[T]}$, with the adjustment $L \leftarrow 2L$ (due to triangle inequality over the two assumptions in (13)), which holds for any deterministic sequence of $\tilde{\mathbf{g}}_t$. In particular, the first bound in (13) holds for any $\mathbf{a} \leftarrow \mathbf{p}_t \in \mathcal{A}$.

To conclude, we combine (15) with (12) using $\mathbf{a}_{[T]} \leftarrow \mathbf{p}_{[T]}$: for all $i \in [m]$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}} \left\langle \mathbf{u}^{(i)}, \frac{1}{T}\sum_{t\in[T]} \mathbf{v}_t^{(i)} \right\rangle\right] \\
&\le \mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]} \left\langle \mathbf{u}_t^{(i)}, \mathbf{v}_t^{(i)} \right\rangle\right] + \mathbb{E}\left[\sup_{\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}} \frac{1}{T}\sum_{t\in[T]} \left\langle \mathbf{u}^{(i)} - \mathbf{u}_t^{(i)}, \mathbf{v}_t^{(i)} \right\rangle\right] \\
&\le \varepsilon^{(i)} + \frac{\mathrm{reg}^{(i)}(T)}{T} + 2\eta L^2 + \frac{D_r(\mathbf{e}_i \| \mathbf{w}_1)}{\eta T}.
\end{aligned}$$

□

We give a convenient corollaries of Theorem 3, which applies the quadratic regularizer (11) to a setting with $m = 2$ approachability sets, and enables tight control for our recalibration applications.

Corollary 1. *In the setting of Theorem 3, choosing* $\mathbf{w}_1 \leftarrow \mathbf{e}_2$ *and* $r = \frac{1}{2}\|\cdot\|_2^2$ *in Algorithm 1 yields*

$$\begin{aligned}
\mathbb{E}\left[\sup_{\mathbf{u}^{(1)} \in \mathcal{U}^{(1)}} \left\langle \mathbf{u}^{(1)}, \frac{1}{T}\sum_{t\in[T]} \mathbf{v}^{(1)}(\mathbf{p}_t, q_t, y_t) \right\rangle\right] &\le \varepsilon^{(1)} + \frac{\mathrm{reg}^{(1)}(T)}{T} + 2\eta L^2 + \frac{2}{\eta T}, \\
\mathbb{E}\left[\sup_{\mathbf{u}^{(2)} \in \mathcal{U}^{(2)}} \left\langle \mathbf{u}^{(2)}, \frac{1}{T}\sum_{t\in[T]} \mathbf{v}^{(2)}(\mathbf{p}_t, q_t, y_t) \right\rangle\right] &\le \varepsilon^{(2)} + \frac{\mathrm{reg}^{(2)}(T)}{T} + 2\eta L^2.
\end{aligned}$$

Proof. This follows from Theorem 3, using $D_r(\mathbf{e}_1\|\mathbf{e}_2) = 2$ and $D_r(\mathbf{e}_2\|\mathbf{e}_2) = 0$. □

B Recalibration

In this section, we apply Theorem 3 to derive a new algorithm for recalibration of an online predictor $q_{[T]} \in [0,1]^T$ (Definition 2) under a proper loss ℓ. We instate our framework with sets

$$\mathcal{A} := \Delta^{\mathcal{N}}, \qquad \mathcal{Q} = \mathcal{Y} := [0,1], \tag{17}$$

where

$$\mathcal{N} := \left\{\frac{i}{n}\right\}_{0\le i\le n}, \quad n \in \mathbb{N}$$

is an ε-net of $[0,1]$, for some fixed $\varepsilon \in (0,1)$. This requires taking $n = \Theta(\frac{1}{\varepsilon})$, which we set in the proof of Theorem 4. We view $\mathcal{A}$ as actions corresponding to distributions over predictions in $\mathcal{N}$, and our random sampling in Line 6 will simply sample a random strategy $p \in \mathcal{N}$ according to $\mathbf{a} \in \mathcal{A}$.

Moreover, we define approachability sets

$$\mathcal{U}^{(1)} := [-1,1]^{\mathcal{N}}, \qquad \mathbf{v}^{(1)}(\mathbf{a}, q, y) := \{\mathbf{a}_s(s-y)\}_{s\in\mathcal{N}}, \tag{18}$$

living in $\mathcal{H}^{(1)} := \mathbb{R}^{\mathcal{N}}$, and

$$\mathcal{U}^{(2)} := \{1\}, \qquad v^{(2)}(\mathbf{a}, q, y) := \sum_{s\in\mathcal{N}} \mathbf{a}_s(\ell(s,y) - \ell(q,y)), \tag{19}$$

living in $\mathcal{H}^{(2)} := \mathbb{R}$. We pause to make a few observations about this setup.

First, the role of $\mathbf{v}^{(1)}$ is to enforce calibration, for which the hint sequence in Definition 2 is irrelevant, so it is independent of q. Moreover, $\|\mathbf{v}\|_1 = \sup_{\mathbf{u}\in\mathcal{U}^{(1)}} \langle \mathbf{u}, \mathbf{v}\rangle$, so the left-hand side of (14) with $i \leftarrow 1$ is $\mathbb{E}[\mathcal{K}_1(p_{[T]}, y_{[T]})]$, where $p_t \in \mathcal{N}$ is the prediction associated with a deterministic action (point mass) $\mathbf{p}_t \in \mathcal{A}$. Thus, the first approachability set (18) exactly captures calibration error.

Analogously, when all actions $\mathbf{a}$ taken by the algorithm are point masses, it is straightforward to see that the second approachability set (19) captures the excess loss aspect of Definition 2.

To instantiate Theorem 3 in the setting of (17), (18), (19), we require three ingredients: no-regret learners over $\mathcal{U}^{(1)}$ and $\mathcal{U}^{(2)}$ in the sense of (12), and an MLOO for the instance. In Appendix B.1, we repurpose a result from [OKS24] for use as an MLOO with appropriate parameters. In Appendix B.2, we put together the components to prove Theorem 4, our main recalibration result.

B.1 Imbalanced MLOO implementation

In this section, we give a short rederivation of Lemma 7, [OKS24]. The subroutine in that result is used within Theorem 3 as our MLOO with parameters $(\varepsilon^{(1)}, \varepsilon^{(2)}) \approx (\varepsilon, \varepsilon^2)$, in the setting of (17), (18), (19). Recall from Definition 5 that the input to the MLOO (when $m = 2$) is mixture weights $\mathbf{w} = (\alpha, \beta)$ satisfying $\alpha + \beta = 1$, $\mathbf{u} \in \mathcal{U}^{(1)}$, and $q \in \mathcal{Q}$, and the output is $\mathbf{a} \in \mathcal{A}$ satisfying

$$\alpha \left\langle \mathbf{u}, \mathbf{v}^{(1)}(\mathbf{a}, q, y) \right\rangle + \beta v^{(2)}(\mathbf{a}, q, y) \le \alpha\varepsilon^{(1)} + \beta\varepsilon^{(2)}, \text{ for all } y \in \{0, 1\}.$$

We begin by recalling some basic facts about proper losses.

Lemma 2. *Let be $\ell : [0, 1] \times \{0, 1\}$ a proper loss. Then, $\ell(\cdot, 0)$ is monotone nondecreasing and $\ell(\cdot, 1)$ is monotone nonincreasing (both in their first argument).*

Proof. Let $0 \le p < q \le 1$, and define $\Delta_0 := \ell(q, 0) - \ell(p, 0)$ and $\Delta_1 := \ell(q, 1) - \ell(p, 1)$. Properness of ℓ implies $(1 - p)\ell(p, 0) + p\ell(p, 1) \le (1 - p)\ell(q, 0) + p\ell(q, 1)$ (and likewise with p, q swapped), so

$$(1 - p)\Delta_0 + p\Delta_1 \ge 0 \ge (1 - q)\Delta_0 + q\Delta_1.$$

Rearranging yields

$$(q - p)(\Delta_0 - \Delta_1) \ge 0 \implies \Delta_1 \le \Delta_0.$$

Finally, we have the desired

$$\Delta_0 \ge (1 - p)\Delta_0 + p\Delta_1 \ge 0, \quad \Delta_1 \le (1 - q)\Delta_0 + q\Delta_1 \le 0.$$

□

We now prove our main helper result in the MLOO implementation.

Lemma 3 (Lemma 7, [OKS24])**.** *Let $\mathbf{u} \in \mathbb{R}^{\{0\} \cup [n]}$ satisfy $\|\mathbf{u}\|_\infty \le 1$. Let $c, d : [0, 1] \times \{0, 1\} \to \mathbb{R}$, and define a function $f : (\{0\} \cup [n]) \times \{0, 1\} \to \mathbb{R}$ by*

$$f(i, y) := \mathbf{u}_i c\left(\frac{i}{n}, y\right) + d\left(\frac{i}{n}, y\right).$$

If the following assumptions hold:

1. *$c(p, y) = p - y$,*
2. *d is proper, L-Lipschitz, $d(\cdot, 0)$ is monotone nondecreasing and $d(\cdot, 1)$ is monotone nonincreasing (both in the first argument), and there exists a $p \in [0, 1]$ such that $d(p, 0) = d(p, 1) = 0$,*
3. *$f(0, 0) \le 0$ and $f(n, 1) \le 0$,*

then there exists a distribution $\mathcal{D}$ over $\{0\} \cup [n]$ such that

$$\mathbb{E}_{i \sim \mathcal{D}}\left[\begin{pmatrix} f(i, 0) \\ f(i, 1) \end{pmatrix}\right] \in \left\{\left(-\infty, \frac{1}{n} + \frac{4L}{n^2}\right] \times \left(-\infty, \frac{1}{n} + \frac{4L}{n^2}\right]\right\}. \tag{20}$$

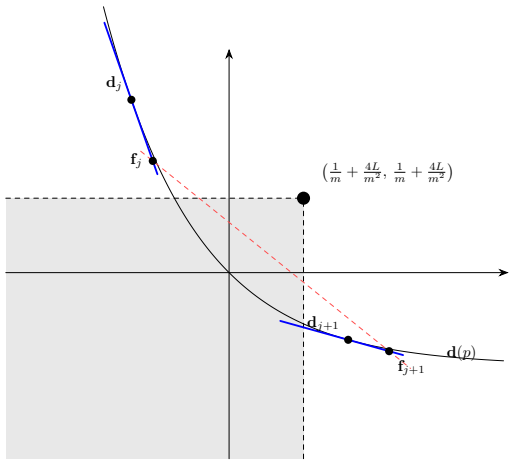
Moreover, $\mathcal{D}$ is either a point mass, or supported on two neighboring indices $\{j, j + 1\}$.

Proof. For convenience in this proof, for $p \in [0, 1]$, we define vector-valued functions

$$\mathbf{c}(p) := \begin{pmatrix} c(p, 0) \\ c(p, 1) \end{pmatrix}, \quad \mathbf{d}(p) := \begin{pmatrix} d(p, 0) \\ d(p, 1) \end{pmatrix},$$

and denote for shorthand $\mathbf{c}_i := \mathbf{c}(\frac{i}{n})$, $\mathbf{d}_i := \mathbf{d}(\frac{i}{n})$, and $\mathbf{f}_i := \mathbf{u}_i\mathbf{c}_i + \mathbf{d}_i$ for all $0 \le i \le n$, so that

$$\mathbf{f}_i = \begin{pmatrix} f(i, 0) \\ f(i, 1) \end{pmatrix}.$$



Our goal is to show that either some $\mathbf{f}_i$ lies in the set on the right-hand side of (20), or the line segment between two adjacent $\mathbf{f}_j, \mathbf{f}_{j+1}$ lies in that set. We also define the orthogonal direction

$$\mathbf{c}_i^\perp := \begin{pmatrix} 1 - \frac{i}{n} \\ \frac{i}{n} \end{pmatrix} \text{ for all } i \in \{0\} \cup [n].$$

Next, for all $i \in \{0\} \cup [n]$, let L_i be the line through $\mathbf{d}_i$ and $\mathbf{f}_i$: since $\left\langle \mathbf{c}_i, \mathbf{c}_i^\perp \right\rangle = 0$,

$$\mathsf{L}_i := \{\mathbf{d}_i + \lambda \mathbf{c}_i \mid \lambda \in \mathbb{R}\} = \left\{\mathbf{v} \in \mathbb{R}^2 \mid \left\langle \mathbf{c}_i^\perp, \mathbf{v} - \mathbf{d}_i \right\rangle = 0\right\}.$$

We first show L_i is a supporting line of $\mathbf{d}(p)$. Since d is proper, for any point $\mathbf{x} = \mathbf{d}(p)$,

$$\left(1 - \frac{i}{n}\right) d(p, 0) + \frac{i}{n} d(p, 1) = \left\langle \mathbf{c}_i^\perp, \mathbf{x} \right\rangle \geq \left\langle \mathbf{c}_i^\perp, \mathbf{d}_i \right\rangle = \left(1 - \frac{i}{n}\right) d\left(\frac{i}{n}, 0\right) + \frac{i}{n} d\left(\frac{i}{n}, 1\right). \quad (21)$$

Indeed, a similar argument shows that there is a supporting line of $\mathbf{d}(p)$ at any $p \in [0, 1]$ with slope $\frac{c(p,1)}{c(p,0)} = 1 - \frac{1}{p}$, which is increasing in p. Thus, the region of $\mathbb{R}^2$ lying above $\mathbf{d}(p)$ is convex.

Now, if any $\mathbf{f}_i$ lies in the third quadrant $\mathbb{R}^2_{\leq 0}$, (20) holds for a point mass $\mathcal{D}$. Assume henceforth there is no such $\mathbf{f}_i$. We claim that no $\mathbf{f}_i$ is strictly in the first quadrant either; if there were such a $\mathbf{f}_i$, then it cannot be $i = 0$ or $i = n$ by assumption, so $\left\langle \mathbf{c}_i^\perp, \mathbf{f}_i \right\rangle > 0$. However,

$$\left\langle \mathbf{c}_i^\perp, \mathbf{f}_i \right\rangle = \left\langle \mathbf{c}_i^\perp, \mathbf{d}_i \right\rangle \leq 0, \quad (22)$$

where we used the definition of L_i and the inequality applied (21) with the p such that $\mathbf{d}(p) = \mathbf{0}_2$.

Thus have that every $\mathbf{f}_i$ lies either in the second or fourth quadrants. By assumption, $\mathbf{f}_0$ must lie in the second quadrant and $\mathbf{f}_n$ must lie in the fourth quadrant. There then exists a $j \in \{0\} \cup [n]$ such that $\mathbf{f}_j$ is in the second quadrant and $\mathbf{f}_{j+1}$ is in the fourth quadrant. Without loss of generality, assume $j \geq \frac{n}{2}$; the case $j \leq \frac{n}{2}$ is symmetric after swapping labels $y = 0$ and $y = 1$.

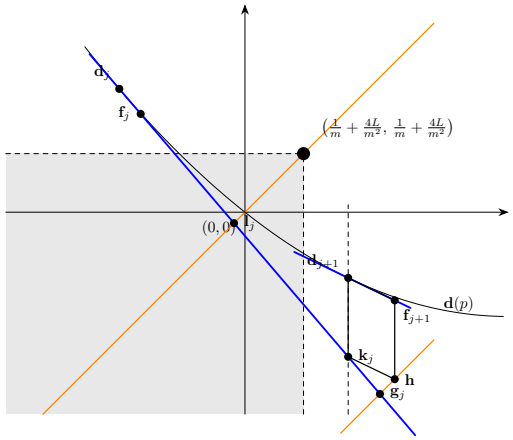
Let $\mathbf{k}_j$ be the intersection of the vertical line through $\mathbf{d}_{j+1}$ with L_j. Since the slope of L_j is $1 - \frac{n}{j}$ and the slope of L_{j+1} is $1 - \frac{n}{j+1}$, the point $\mathbf{k}_j$ lies below $\mathbf{d}_{j+1}$. Because $\mathbf{d}_j$ lies above L_{j+1}, a direct calculation shows that the vertical gap between $\mathbf{d}_{j+1}$ and $\mathbf{k}_j$ is at most the product of the horizontal gap between $\mathbf{d}_j, \mathbf{d}_{j+1}$, and the difference in slopes between $\mathsf{L}_j, \mathsf{L}_{j+1}$. This is

$$\frac{n}{j(j+1)} \cdot \left(d\left(\frac{j+1}{n}, 0\right) - d\left(\frac{j}{n}, 0\right)\right) \leq \frac{n}{(n/2)(n/2+1)} \cdot \frac{L}{n} \leq \frac{4L}{n^2}, \quad (23)$$

where we applied Lipschitzness of d.

Finally, we are ready to define $\mathcal{D}$. Let $\mathbf{i}_j$ be the intersection between L_j and the line through the origin with slope 1. We claim that $\mathbf{i}_j$ lies between $\mathbf{f}_j$ and the following point on L_j:

$$\mathbf{g}_j := \mathbf{k}_j + \mathbf{u}_{j+1}\mathbf{c}_j = \mathbf{k}_j + (\mathbf{f}_{j+1} - \mathbf{d}_{j+1}) + \mathbf{u}_{j+1}(\mathbf{c}_{j+1} - \mathbf{c}_j).$$



This is because $\mathbf{f}_j$'s second coordinate is larger than its first. On the other hand,

$$\mathbf{g}_j = \mathbf{f}_{j+1} + (\mathbf{k}_j - \mathbf{d}_{j+1}) + \frac{\mathbf{u}_{j+1}}{n}\mathbf{1}_2,$$

so its first coordinate is larger than its second, because $\mathbf{f}_{j+1}$ has this property and $\mathbf{k}_j - \mathbf{d}_{j+1}$ is vertically downwards. Thus, $\mathbf{i}_j = (1-\lambda)\mathbf{f}_j + \lambda\mathbf{g}_j$ for some $\lambda \in [0,1]$.

We output $\mathbf{f}_j$ with probability $1-\lambda$, and $\mathbf{f}_{j+1}$ with probability λ. We claim that $(1-\lambda)\mathbf{f}_j + \lambda\mathbf{f}_{j+1}$ lies in the set (20), which would conclude the proof. To see this, note that $\mathbf{i}_j$ is in the third quadrant, because its two coordinates are equal (say, to t), so combining (22) with $\mathbf{i}_j \in \mathsf{L}_j$,

$$t = \left(1 - \frac{i}{n}\right) t + \frac{i}{n} \cdot t = \left\langle \mathbf{c}_j^{\perp}, \mathbf{d}_j \right\rangle \leq 0.$$

Finally, we can bound the coordinates of $(1-\lambda)\mathbf{f}_j + \lambda\mathbf{f}_{j+1}$:

$$(1-\lambda)\mathbf{f}_j + \lambda\mathbf{f}_{j+1} = \mathbf{i}_j + \lambda\left((\mathbf{d}_{j+1} - \mathbf{k}_j) - \frac{\mathbf{u}_{j+1}}{n}\mathbf{1}_2\right).$$

Indeed, both coordinates of $\lambda(\mathbf{d}_{j+1} - \mathbf{k}_j)$ are at most $\frac{4L}{n^2}$ using (23), and both coordinates of $\frac{\lambda \mathbf{u}_{j+1}}{n}\mathbf{1}_2$ are at most $\frac{1}{n}$ using the assumption on $\mathbf{u}$. Since $\mathbf{i}_j \in \mathbb{R}^2_{\leq 0}$, we are done. □

Lemma 4. *Under Assumption 1, there is a ε-MLOO for (17), (18), (19), with*

$$\varepsilon^{(1)} = \frac{1}{n}, \quad \varepsilon^{(2)} = \frac{4L}{n^2}.$$

Moreover, the output is a 2-sparse element of $\Delta^{\mathcal{N}}$ supported on neighboring predictions $\{\frac{j}{n}, \frac{j+1}{n}\}$.

Proof. Consider inputs to the MLOO, $\mathbf{w} = (\alpha, \beta) \in \Delta^2$, $\mathbf{u} \in \mathcal{U}^{(1)}$, and $q \in \mathcal{Q}$. Define

$$d(p, y) := \frac{\beta}{\alpha}\left(\ell(p, y) - \ell(q, y)\right).$$

We check that this function satisfies the assumptions of Lemma 3. Indeed, it is proper and $\frac{\beta L}{\alpha}$-Lipschitz by assumption, satisfies the monotonicity conditions by Lemma 2, and satisfies $d(q,0) = d(q,1) = 0$. Moreover, propriety implies $f(0,0) = d(0,0) \leq 0$ and $f(n,1) = d(1,1) \leq 0$.

Now let $\mathbf{a}$ be the natural identification of the output distribution from Lemma 3 with an element of $\Delta^{\mathcal{N}}$. By the conclusion of Lemma 3, we have for either $y \in \mathcal{Y}$,

$$\left\langle \mathbf{u}, \mathbf{v}^{(1)}(\mathbf{a}, q, y) \right\rangle + \frac{\beta}{\alpha} v^{(2)}(\mathbf{a}, q, y) \leq \frac{1}{n} + \frac{4\beta L}{\alpha n^2}.$$

Rescaling this guarantee by α and comparing with the MLOO definition gives the result. □

B.2 Main result

We can now combine our MLOO implementation in Lemma 4 with standard regret minimization bounds within Theorem 3 to obtain our main result on recalibration. We briefly discuss the required online learners $\mathsf{alg}^{(1)}$ and $\mathsf{alg}^{(2)}$ satisfying (12). First, because $\mathcal{U}^{(2)}$ is a singleton, $\mathsf{alg}^{(2)}$ can trivially achieve $\mathrm{reg}^{(2)}(T) = 0$ by returning 1 every iteration. We provide $\mathsf{alg}^{(1)}$ in the next result.

Lemma 5. *There is an online learning algorithm* $\mathsf{alg}^{(1)}$ *that satisfies* (12) *with*

$$\mathrm{reg}^{(1)}(T) := \frac{\varepsilon T}{8} + \frac{16n}{\varepsilon}.$$

Proof. This follows from projected gradient descent (Lemma 1 with the settings in (11)), $\mathbf{x}_1 \leftarrow \mathbf{0}_{\mathcal{N}}$, and $\eta \leftarrow \frac{\varepsilon}{16}$. More precisely, we use that each $\mathbf{u} \in \mathcal{U}^{(1)}$ has

$$\|\mathbf{u}\|_2 \le \sqrt{n+1}\,\|\mathbf{u}\|_\infty \le \sqrt{2n} =: R,$$

and each feedback vector $\mathbf{v} := \mathbf{v}^{(1)}(\mathbf{a}, q, y)$ for $(\mathbf{a}, q, y) \in \mathcal{A} \times \mathcal{Q} \times \mathcal{Y}$ has $\|\mathbf{v}\|_2 \le \|\mathbf{v}\|_1 \le 2 =: L$. □

Theorem 4. *Let* $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$ *be proper such that Assumption 1 holds, and let* $\varepsilon \in (0, \frac{1}{2})$. *In the setting of Model 1, there is an algorithm that outputs* $p_{[T]} \in [0,1]^T$ *that is* $(\varepsilon, \varepsilon^2)$*-recalibrated against* $(q_{[T]}, y_{[T]})$ *under* ℓ*, if for an appropriate constant,*

$$T = \Omega\left(\frac{L^2}{\varepsilon^3}\right).$$

Proof. We apply Corollary 1 with the setup in (17), (18), (19), where we use Lemma 4 with $n = \lceil \frac{2}{\varepsilon} \cdot \sqrt{L} \rceil$ as the MLOO, Lemma 5 as $\mathsf{alg}^{(1)}$, and the trivial learner over $\mathcal{U}^{(2)}$ as $\mathsf{alg}^{(2)}$. We now bound each of the quantities in Theorem 3 and explain its implications to recalibration.

First, the MLOO parameters yielded by Lemma 4 with the stated value of n are

$$\varepsilon^{(1)} = \frac{1}{n} \le \frac{\varepsilon}{4}, \quad \varepsilon^{(2)} = \frac{L}{n^2} \le \frac{\varepsilon^2}{4}. \tag{24}$$

This implies we can take L as given in the theorem statement in (13), because $|\langle \mathbf{u}, \mathbf{v} \rangle| \le 1$ for $\mathbf{u} \in \mathcal{U}^{(1)}$ and $\mathbf{v} = \mathbf{v}^{(1)}(\mathbf{a}, q, y)$ by the ℓ_1-ℓ_∞ Hölder's inequality, and $|\mathbf{v}^{(2)}(\mathbf{a}, q, y)| \le L$ by Lipschitzness of ℓ. Next, Lemma 5 gives

$$\frac{\mathrm{reg}^{(1)}(T)}{T} \le \frac{\varepsilon}{2} + \frac{4n}{\varepsilon T}$$

and we clearly have $\mathrm{reg}^{(2)}(T) = 0$. Therefore, if we take

$$\eta \leftarrow \frac{\varepsilon^2}{8L^2}, \quad T \leftarrow \frac{128L^2}{\varepsilon^3},$$

then the two bounds in Corollary 1 are at most $(\varepsilon, \varepsilon^2)$ respectively. This holds for any randomized estimator $\mathbf{p}_t$ such that $\mathbb{E}[\mathbf{p}_t] = \mathbf{a}_t$ in each iteration, so we let $\mathbf{p}_t$ be a randomly sampled point mass according to $\mathbf{a}_t$, and output the associated prediction p_t.

By the definitions in (18), (19), when every prediction output is a point mass, Corollary 1 implies $(\varepsilon, \varepsilon^2)$-recalibration: the first inequality is exactly

$$\mathbb{E}\left[\frac{1}{T}\sum_{s \in \mathcal{N}} |n_T(s)(s - \bar{y}_T(s))|\right] = \mathbb{E}\left[\mathcal{K}_1\left(p_{[T]}, y_{[T]}\right)\right] \le \varepsilon,$$

and the second inequality is exactly

$$\mathbb{E}\left[\frac{1}{T}\sum_{t \in [T]} \ell(p_t, y) - \ell(q_t, y)\right] = \mathbb{E}\left[\ell\left(p_{[T]}, y_{[T]}\right) - \ell\left(q_{[T]}, y_{[T]}\right)\right] \le \varepsilon^2.$$

□

As shown in Appendix E (precisely, Theorem 8), Theorem 4 is minimax optimal up to constant factors.

B.3 High-probability recalibration

Theorem 4 proves that taking $T \approx \varepsilon^{-3}$ suffices for the expected recalibration guarantee in Definition 2 for Lipschitz proper losses. Here we improve Theorem 4 to give a high-probability recalibration bound. Our applications in Sections C and D are fully deterministic except in their reliance on Theorem 4, so they also straightforwardly extend to hold with high probability.

To obtain our high-probability result, we separate the prediction output to nature from the feedback used by the approachability procedure. On round t, after nature has fixed (q_t, y_t) and revealed q_t, the procedure chooses a mixed action $\mathbf{a}_t$ and samples and outputs $p_t \sim \mathbf{a}_t$ before y_t is revealed. Thus, future hint-label pairs may depend on the realized predictions $p_{[t]}$, as allowed by Model 1. After y_t is revealed, however, the internal approachability update uses the deterministic mixed action $\mathbf{a}_t$ rather than its sampled point mass. Using the same action, context, and approachability sets (17), (18), (19) as before gives the following pathwise guarantee for the mixed actions.

Lemma 6. *Let $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$ be proper such that Assumption 1 holds, and let $\varepsilon \in (0, \frac{1}{2})$. In the setting of Model 1, there is an even $N = \Theta(\sqrt{L}/\varepsilon)$ and an algorithm that outputs $\mathbf{a}_{[T]} \in (\Delta^{\mathcal{N}})^T$ (where each $\mathbf{a}_t$ has the same measurability assumptions as each p_t in Model 1), such that the following conditions simultaneously hold deterministically, if for an appropriate constant, $T = \Omega(\frac{L^2}{\varepsilon^3})$:*

$$\left\|\mathbf{c}\left(\mathbf{a}_{[T]}, y_{[T]}\right)\right\|_1 \le \frac{\varepsilon}{2}, \quad \frac{1}{T}\sum_{t\in[T]}\left(\ell\left(\mathbf{a}_t, y_t\right) - \ell(q_t, y_t)\right) \le \frac{\varepsilon^2}{2}, \tag{25}$$

where we overload the definition $\ell(\mathbf{a}, y) := \sum_{s\in\mathcal{N}} \mathbf{a}_s \ell(s, y)$ for any $\mathbf{a} \in \Delta^{\mathcal{N}}$, and we define the miscalibration vector $\mathbf{c}(\mathbf{a}_{[T]}, y_{[T]}) \in \mathbb{R}^{\mathcal{N}}$ by

$$\left[\mathbf{c}(\mathbf{a}_{[T]}, y_{[T]})\right]_s = \frac{1}{T}\sum_{t\in[T]}\left[\mathbf{a}_t\right]_s (s - y_t), \text{ for all } s \in \mathcal{N}. \tag{26}$$

Moreover, each $\mathbf{a}_t$ is deterministically supported on at most two strategies s, s', with $|s - s'| \le \frac{\varepsilon}{2}$.

Proof. Run the proof of Corollary 1 with $\mathbf{p}_t = \mathbf{a}_t$ in every internal update. The conditional-unbiasedness step in that proof then becomes a pointwise equality, so its guarantees hold pathwise for every realized adaptive sequence. Combining this observation with our approachability sets (18), (19), using the same parameters as in Theorem 4 up to constant factors, gives (25). The condition on each $\mathbf{a}_t$ results from our MLOO construction in Lemma 4, using that $\frac{1}{N} \le \frac{\varepsilon}{2}$. □

We next bound the error from sampling each mixed action $\mathbf{a}_t$ into the online prediction $p_t \in \mathcal{N}$, on both terms in (25). The bounds are stated conditionally so that the mixed actions, labels, and future play may all depend on previous sampled predictions.

Lemma 7. *Let $\delta \in (0, \frac{1}{2})$, and let $\{\mathcal{G}_t\}_{t\in[T]}$ describe the history immediately before p_t is sampled, after the current pair (q_t, y_t) has been fixed. Suppose $\mathbf{a}_t \in \Delta^{\mathcal{N}}$ and $y_t \in \{0,1\}$ are $\mathcal{G}_t$-measurable, and conditionally sample $p_t \sim \mathbf{a}_t$, so that $\mathbb{E}[\mathbf{p}_t \mid \mathcal{G}_t] = \mathbf{a}_t$ for $\mathbf{p}_t := \mathbf{e}_{p_t}$. Then if for an appropriate constant,*

$$T = \Omega\left(\frac{|\mathcal{N}|}{\varepsilon^2}\log\left(\frac{4|\mathcal{N}|}{\delta}\right)\right),$$

we have with probability $\ge 1 - \frac{\delta}{2}$,

$$\left\|\mathbf{c}\left(\mathbf{a}_{[T]}, y_{[T]}\right) - \mathbf{c}\left(\mathbf{p}_{[T]}, y_{[T]}\right)\right\|_1 \le \frac{\varepsilon}{2}.$$

Proof. For each $t \in [T]$, define a vector $\mathbf{x}_t \in \mathbb{R}^{\mathcal{N}}$ by

$$[\mathbf{x}_t]_s := \left(\mathbb{I}\left(p_t = s\right) - [\mathbf{a}_t]_s\right)(s - y_t), \qquad s \in \mathcal{N}.$$

For every sign vector $\boldsymbol{\xi} \in \{-1, 1\}^{\mathcal{N}}$, the scalar random variables $Z_t(\boldsymbol{\xi}) := \langle \boldsymbol{\xi}, \mathbf{x}_t \rangle$ form a martingale difference sequence with respect to the sampling history. Indeed, conditional on $\mathcal{G}_t$,

$$Z_t(\boldsymbol{\xi}) = \xi_{p_t}(p_t - y_t) - \sum_{s\in\mathcal{N}}[\mathbf{a}_t]_s \xi_s (s - y_t),$$

which has conditional mean zero and absolute value at most 2. Therefore, Azuma–Hoeffding's inequality and a union bound over the $2^{|\mathcal{N}|}$ sign vectors imply that, with probability at least $1-\delta/2$,

$$\left\|\sum_{t\in[T]}\mathbf{x}_t\right\|_1 = \max_{\boldsymbol{\xi}\in\{-1,1\}^{\mathcal{N}}}\sum_{t\in[T]}Z_t(\boldsymbol{\xi}) \le \sqrt{8T\left(|\mathcal{N}|\log 2+\log\left(\frac{2}{\delta}\right)\right)}.$$

Since

$$\left\|\mathbf{c}\left(\mathbf{a}_{[T]},y_{[T]}\right)-\mathbf{c}\left(\mathbf{p}_{[T]},y_{[T]}\right)\right\|_1 = \frac{1}{T}\left\|\sum_{t\in[T]}\mathbf{x}_t\right\|_1,$$

the conclusion follows from the stated lower bound on T. □

Lemma 8. *Let $\delta\in(0,\frac{1}{2})$, and suppose the conditional sampling setup of Lemma 7 holds. Also, suppose that Assumption 1 holds, and that each $\mathbf{a}_t$ is supported on at most two strategies s,s', with $|s-s'|\le\frac{\varepsilon}{2}$. Then if for an appropriate constant,*

$$T=\Omega\left(\frac{L^2\log\left(\frac{1}{\delta\varepsilon}\right)}{\varepsilon^2}\right),$$

we have with probability $\ge 1-\frac{\delta}{2}$,

$$\frac{1}{T}\sum_{t\in[T]}\left(\ell\left(p_t,y_t\right)-\ell\left(\mathbf{a}_t,y_t\right)\right)\le\frac{\varepsilon^2}{2},$$

where $\ell(\mathbf{a}_t,y_t):=\sum_{s\in\mathcal{N}}[\mathbf{a}_t]_s\ell(s,y_t)$.

Proof. For all $t\in[T]$, define a random variable

$$Y_t:=\ell(p_t,y_t)-\ell(\mathbf{a}_t,y_t).$$

Then $\mathbb{E}[Y_t\mid\mathcal{G}_t]=0$, so Y_t is a martingale difference sequence satisfying

$$|Y_t|\le\frac{L\varepsilon}{2}\text{ with probability }1.$$

Here, we used Lipschitzness of ℓ, and the support assumption on $\mathbf{a}_t$. Thus, for T taken as stated, the Azuma-Hoeffding inequality shows the desired claim holds with probability $\ge 1-\frac{\delta}{2}$:

$$\frac{1}{T}\sum_{t\in[T]}Y_t\le\frac{\varepsilon^2}{2}.$$

□

We conclude with our high-probability recalibration result, by combining Lemmas 6, 7, and 8.

Corollary 2. *Let $\ell:[0,1]\times\{0,1\}\to\mathbb{R}$ be proper such that Assumption 1 holds, and let $(\varepsilon,\delta)\in(0,\frac{1}{2})^2$. In the setting of Model 1, there is an algorithm that outputs $p_{[T]}\in[0,1]^T$ such that the following conditions hold with probability $\ge 1-\delta$:*

$$\mathcal{K}_1\left(p_{[T]},y_{[T]}\right)\le\varepsilon,\quad \ell\left(p_{[T]},y_{[T]}\right)-\ell\left(q_{[T]},y_{[T]}\right)\le\varepsilon^2,$$

if for an appropriate constant,

$$T=\Omega\left(\frac{L^2}{\varepsilon^3}+\frac{\sqrt{L}\log\left(\frac{\sqrt{L}}{\delta\varepsilon}\right)}{\varepsilon^3}+\frac{L^2\log\left(\frac{1}{\delta\varepsilon}\right)}{\varepsilon^2}\right).$$

Proof. Run the mixed-action procedure from Lemma 6 with the even choice $N=\Theta(\sqrt{L}/\varepsilon)$ specified there. On each round, conditionally sample and output $p_t\sim\mathbf{a}_t$ before y_t is revealed, while using $\mathbf{a}_t$ itself in the internal approachability update after the reveal. The pathwise guarantees (25) therefore hold for the realized adaptive sequence. Since $|\mathcal{N}|=N+1=O(\sqrt{L}/\varepsilon)$, the three terms in the stated horizon respectively imply the conditions of Lemmas 6, 7, and 8. The two sampling conclusions hold simultaneously with probability at least $1-\delta$ by a union bound, and combining them with (25) gives the claim. □

C Calibeating and Calibration

In this section, we apply Theorem 4 to improve existing calibeating algorithms, to simultaneously ensure ε-calibration (Definition 1) and ε^2-calibeating (Definition 3) in $T \approx \varepsilon^{-3}$ iterations. Previously, both results were known in isolation [FV98, FH23], but not simultaneously.

While natural in its own right, this problem is well-motivated from the perspective of the *refinement score* (3), the loss benchmark used in the literature on calibeating [LNPR22, FH23, CHJL26]. Recall that the refinement score is the loss achieved by using the hindsight average-label predictor $\bar{y}_T(v)$, in all iterations where $q_t = v$. This hindsight predictor is infeasible to implement without knowledge of $y_{[T]}$, so calibeating aims to compete with its loss in an online setting.

Beyond achieving a strong loss guarantee, however, the hindsight predictor $v \to \bar{y}_T(v)$ also is clearly calibrated. On the other hand, existing calibeating algorithms do not ensure calibration at the same rate as [FV98]; indeed, requiring this additional property has thus far resulted in worse asymptotics. As a result, it is natural to ask if one can simultaneously achieve calibration in an online setting.

We begin by providing a calibeating result (Proposition 1) that extends special cases in [FH23] to all smooth proper losses. Particularly, [FH23] proved variants of Proposition 1 for the squared loss and (truncated) logistic loss (4), which we extend to hold for a more general family.

To state our result, we require a standard reformulation of proper losses as a representative univariate function, often called the *Savage representation* [Sav71, GR07].

Lemma 9 (Theorem 1, [GR07])**.** *Let $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$ be proper, and define*

$$H(p) := (1-p)\ell(p,0) + p\ell(p,1) = \mathbb{E}_{y\sim\mathsf{Bern}(p)}[\ell(p,y)] \text{ for all } p \in [0,1]. \tag{27}$$

Then H is concave. Moreover, if H is differentiable, we have the equivalence

$$\ell(p,y) = H(p) + H'(p)(y-p), \text{ for all } y \in \{0,1\}.$$

For the rest of the section, we fix a proper loss ℓ and identify it with its univariate Savage representation H. We make the following assumption on H to parameterize our results.

Assumption 2 (γ-smoothness)**.** *H is differentiable, and there exists $\gamma \geq 1$ such that for $\omega := -H$,*

$$D_\omega(p\|q) \leq \frac{\gamma}{2}(p-q)^2 \text{ for all } (p,q) \in [0,1]^2. \tag{28}$$

Remark 2. *Assumption 2 is only used in Proposition 1, and not in the recalibration step afterwards. We believe that Proposition 1 likely extends to a broader range of losses than those satisfying Assumption 2: for example, [FH23] give a somewhat more ad hoc argument that handles the cross entropy loss ℓ_{log}, for which Assumption 2 does not hold with $\gamma < \infty$. We give Proposition 1 under a smoothness bound for simplicity, as this covers the squared loss ℓ_{sq} as a special case.*

It is helpful to note that Assumption 2 is implied by Assumption 1 with a slight blowup in parameters. However, we choose to isolate Assumption 2 because it lets us state Proposition 1 in a convenient way, which may be of independent future interest.

Lemma 10. *Assumption 1 implies Assumption 2 with $G \leftarrow 2L$.*

Proof. Let $\omega := -H$ and $\psi(p) := \ell(p,0) - \ell(p,1)$. For any $p, r \in [0,1]$, properness gives

$$H(r) \leq (1-r)\ell(p,0) + r\ell(p,1) = H(p) + (r-p)(\ell(p,1) - \ell(p,0)).$$

Equivalently,

$$\omega(r) \geq \omega(p) + \psi(p)(r-p),$$

so $\psi(p) \in \partial\omega(p)$ for every $p \in [0,1]$. Assumption 1 and the triangle inequality imply

$$|\psi(p) - \psi(q)| \leq |\ell(p,0) - \ell(q,0)| + |\ell(p,1) - \ell(q,1)| \leq 2L|p-q|.$$

Thus ψ is a continuous subgradient selection of the one-dimensional convex function ω, and hence ω is differentiable with $\omega' = \psi$ (using one-sided derivatives at the endpoints). By the standard equivalence between Lipschitz gradients and quadratic upper bounds for Bregman divergence (Theorem 2.1.5, [Nes03]), ω is $2L$-smooth on $[0,1]$, which is Assumption 2 with $G \leftarrow 2L$. □

Next, to state our main helper result, analogously to the (offline) refinement score (3), we additionally define the *online refinement score* (7), where we use the available prefix average $\bar{y}_{t-1}(q_t)$ at every iteration, rather than the hindsight average $\bar{y}_T(q_t)$. We now bound the gap between refinement scores.

Proposition 1. *In the setting of Model 1, if $q_{[T]}$ is s-sparse and Assumption 2 holds,*

$$0 \le \widehat{\mathcal{R}}_\ell\left(q_{[T]}, y_{[T]}\right) - \mathcal{R}_\ell\left(q_{[T]}, y_{[T]}\right) \le \frac{\gamma s\left(\log T + 1\right)}{2T}.$$

Proof. Throughout denote $\omega := -H$. We first handle the special case where $q_t = v$ for all $t \in [T]$ (i.e., $s = 1$). In this case, denoting $\bar{y}_t := \bar{y}_t(v)$ for all $t \in [T]$ for simplicity,

$$\mathcal{R}_\ell(q_{[T]}, y_{[T]}) = \frac{1}{T}\sum_{t\in[T]} \ell(\bar{y}_T, y_t) = (1-\bar{y}_T)\ell(\bar{y}_T, 0) + \bar{y}_T\ell(\bar{y}_T, 1) = H(\bar{y}_T).$$

Hence,

$$\widehat{\mathcal{R}}_\ell(q_{[T]}, y_{[T]}) - \mathcal{R}_\ell(q_{[T]}, y_{[T]}) = \frac{1}{T}\sum_{t\in[T]} \ell(\bar{y}_{t-1}, y_t) - H(\bar{y}_T). \tag{29}$$

Moreover, for every $t \in [T]$, $\bar{y}_t = \bar{y}_{t-1} + \frac{1}{t}(y_t - \bar{y}_{t-1})$, so Lemma 9 gives

$$\ell(\bar{y}_{t-1}, y_t) = H(\bar{y}_{t-1}) + tH'(\bar{y}_{t-1})(\bar{y}_t - \bar{y}_{t-1}) = tH(\bar{y}_t) - (t-1)H(\bar{y}_{t-1}) + tD_\omega\left(\bar{y}_t \| \bar{y}_{t-1}\right).$$

Telescoping the above identity over all $t \in [T]$, and plugging this into (29), gives

$$\widehat{\mathcal{R}}_\ell(q_{[T]}, y_{[T]}) - \mathcal{R}_\ell(q_{[T]}, y_{[T]}) = \frac{1}{T}\sum_{t\in[T]} tD_\omega(\bar{y}_t \| \bar{y}_{t-1}).$$

Next, applying Assumption 2 gives

$$tD_\omega(\bar{y}_t \| \bar{y}_{t-1}) \le \frac{\gamma t}{2}(\bar{y}_t - \bar{y}_{t-1})^2 = \frac{\gamma}{2t}(y_t - \bar{y}_{t-1})^2 \le \frac{\gamma}{2t},$$

because $y_t, \bar{y}_{t-1} \in [0,1]$. Therefore,

$$0 \le \widehat{\mathcal{R}}_\ell(q_{[T]}, y_{[T]}) - \mathcal{R}_\ell(q_{[T]}, y_{[T]}) \le \frac{\gamma}{2T}\sum_{t\in[T]} \frac{1}{t} \le \frac{\gamma}{2}\cdot\frac{\log T + 1}{T}.$$

For the general case, we apply the preceding argument separately to each "bin" $v \in [0,1]$ such that $q_t = v$ at least once. Let $T_v := |\{t \in [T] \mid q_t = v\}|$. Both $\mathcal{R}_\ell$ and $\widehat{\mathcal{R}}_\ell$ decompose over bins, so

$$0 \le \widehat{\mathcal{R}}_\ell(q_{[T]}, y_{[T]}) - \mathcal{R}_\ell(q_{[T]}, y_{[T]}) \le \sum_v \frac{T_v}{T}\cdot\frac{\gamma}{2}\cdot\frac{\log T_v + 1}{T_v}.$$

Since there are at most s nonempty bins and each $T_v \le T$, the claim follows. □

We are now ready to prove our main result on simultaneous calibeating and calibration.

Theorem 5. *Let $\ell : [0,1]\times\{0,1\} \to \mathbb{R}$ be proper such that Assumptions 1, 2 hold, and let $\varepsilon \in (0, \frac{1}{2})$. In the setting of Model 1, if $q_{[T]}$ is s-sparse, there is an algorithm that outputs $p_{[T]} \in [0,1]^T$ that is ε-calibrated against $y_{[T]}$ and ε^2-calibeats $(q_{[T]}, y_{[T]})$ under ℓ, if for an appropriate constant,*

$$T = \Omega\left(\frac{L^2}{\varepsilon^3} + \frac{\gamma s\log\left(\frac{\gamma s}{\varepsilon}\right)}{\varepsilon^2}\right).$$

Proof. Define a sequence $r_t := \bar{y}_{t-1}(q_t)$, which can be computed online under Model 1. We first observe that if T is taken as stated, r_t $\frac{\varepsilon^2}{2}$-calibeats $(q_{[T]}, y_{[T]})$ under ℓ, because Proposition 1 shows

$$\frac{1}{T}\sum_{t\in[T]} \ell(r_t, y_t) \le \mathcal{R}_\ell\left(q_{[T]}, y_{[T]}\right) + \frac{\gamma s(\log T + 1)}{T} \le \mathcal{R}_\ell\left(q_{[T]}, y_{[T]}\right) + \frac{\varepsilon^2}{2}.$$

Now to achieve the result, it is enough to ensure that $p_{[T]}$ is $(\varepsilon, \frac{\varepsilon^2}{2})$-recalibrated against $(r_{[T]}, y_{[T]})$, for which Theorem 4 suffices when T is taken as stated. □

Remark 3 (Sparsity in calibeating). *As in prior works on calibeating [FH23, LNPR22, CHJL26], we assume that the hint sequence $q_{[T]}$ is s-sparse for some s independent of T (we focus on the regime $s \approx \frac{1}{\varepsilon}$). We remark that some such sparsity assumption is essentially necessary. More precisely, if each q_t is distinct, then each $\overline{y}_T(q_t) = y_t$ and thus the refinement score $\mathcal{R}_\ell(q_{[T]}, y_{[T]}) = 0$ for any ℓ with $\ell(y, y) = 0$ (e.g., those in (4)). On the other hand, in many simple settings (e.g., $\ell = \ell_{\mathsf{sq}}$ and each $y_t \sim \mathsf{Bern}(\frac{1}{2})$), it is impossible to obtain $\ell(p_{[T]}, y_{[T]}) = o(1)$, and thus α-calibeating cannot occur as we take $\alpha \to 0$. On the other hand, $\alpha \to 0$ is the interesting regime to study asymptotic (re)calibration, because if $\alpha = \Omega(1)$, then writing T asymptotically in α is ill-posed.*

D Extensions to Multiple Sequences

In this section, we generalize our results to the more challenging setting with multiple hint sequences, where multi-recalibration and multi-calibeating are required against *every* hint sequence. We formally define the problems below. Following ideas from [CHJL26], we solve them using a reduction to the single-hint sequence problems we addressed in previous sections. We describe the reduction in Appendix D.1 and present our multi-sequence results in Appendix D.2, D.3.

Model 2 (Multi-hinted sequential prediction). *In the* multi-hinted sequential prediction *setting, for some $m \in \mathbb{N}$, a learner observes m different* hint *sequences $q^i_{[T]} \in [0,1]^T$ for all $i \in [m]$, and outputs a* prediction *sequence $p_{[T]} \in [0,1]^T$ before observing a* label *sequence $y_{[T]} \in \{0,1\}^T$. We assume $(q_t^{[m]} := \{q_t^i\}_{i \in [m]}, y_t)$, can depend jointly on all previous $(q^{[m]}_{[t-1]}, p_{[t-1]}, y_{[t-1]})$, but not p_t; similarly, p_t can depend on $(q^{[m]}_{[t-1]}, p_{[t-1]}, y_{[t-1]})$ and $q_t^{[m]}$, but not y_t.*

Definition 6 (Multi-recalibration). *In the setting of Model 2, let $\alpha, \beta \geq 0$, and let $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$ be a loss. We say $p_{[T]}$ is (α, β)-recalibrated against $(q^{[m]}_{[T]}, y_{[T]})$ under ℓ if for all $i \in [m]$, $p_{[T]}$ is (α, β)-recalibrated (Definition 2) against $q^i_{[T]}$ under ℓ.*

Definition 7 (Multi-calibeating). *In the setting of Model 2, let $\alpha \geq 0$, and let $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$ be a loss. We say that $p_{[T]}$ α-*multi-calibeats *$(q^{[m]}_{[T]}, y_{[T]})$ under ℓ if for all $i \in [m]$, $p_{[T]}$ α-calibeats (Definition 3) $(q^i_{[T]}, y_{[T]})$ under ℓ.*

D.1 Hint compression

Following [CHJL26], we handle multiple hint sequences by a reduction to the single-hint sequence problems studied in previous sections. The reduction is achieved via standard results for online agnostic learning: indeed, the key component of the reduction, which we term *hint compression* and formally describe in Definition 8 below, can be viewed as equivalent to online agnostic learning.

Definition 8 (Hint compression). *In the setting of Model 2, let $\varepsilon \geq 0$ and let $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$ be a loss. We say $p_{[T]}$ achieves ε-hint compression against $(q^{[m]}_{[T]}, y_{[T]})$ under ℓ if for all $i \in [m]$,*

$$\mathbb{E}\left[\ell\left(p_{[T]}, y_{[T]}\right) - \ell\left(q^i_{[T]}, y_{[T]}\right)\right] \leq \varepsilon.$$

The following results for hint compression are standard in the online optimization literature.

Proposition 2 (Proposition 3.2, [CBL06]). *In the setting of Model 2, let $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$ be a proper loss satisfying Assumption 1. There is a deterministic algorithm that outputs $p_{[T]}$ achieving $O(\frac{L \log m}{T})$-hint compression.*

More precisely, Proposition 3.2 of [CBL06] shows that every η-mixable loss achieves $\frac{\log m}{\eta T}$-hint compression, and it is known (e.g., by combining Theorem 1, [RW10] with Section 3.1, [vERW12]) that every L-Lipschitz proper loss is $\eta = \frac{1}{L}$-mixable. We also note that Proposition 2 is minimax optimal up to constant factors assuming $T = \Omega(\log m)$, by Theorem 3.6 of [CBL06].

D.2 Multi-recalibration

In this section, we state our generalization of Theorem 4 to multiple hint sequences.

Theorem 6. *Let $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$ be proper such that Assumption 1 holds, and let $\varepsilon \in (0, \frac{1}{2})$. In the setting of Model 2, there is an algorithm that outputs $p_{[T]} \in [0,1]^T$ that is $(\varepsilon, \varepsilon^2)$-recalibrated against $(q^{[m]}_{[T]}, y_{[T]})$ under ℓ, if for an appropriate constant,*

$$T = \Omega\left(\frac{L^2}{\varepsilon^3} + \frac{L \log m}{\varepsilon^2}\right).$$

Proof. The theorem follows from combining Proposition 2 and Theorem 4. Concretely, we apply Proposition 2 to compress the m hint sequences into a single sequence, incurring $\frac{\varepsilon^2}{2}$ additional loss when $T = \Omega(L \log(m) \cdot \varepsilon^{-2})$. We then apply Theorem 4 to achieve recalibration w.r.t. the single sequence, incurring additional loss $\frac{\varepsilon^2}{2}$ when $T = \Omega(\frac{L^2}{\varepsilon^3})$. □

We make some remarks regarding the optimality of Theorem 6. For the family of $O(1)$-Lipschitz proper losses, both terms in the iteration complexity of Theorem 6 are necessary up to constant factors. Indeed, $(\varepsilon, \varepsilon^2)$-multi-recalibration implies $(\varepsilon, \varepsilon^2)$-recalibration for a single hint sequence (by taking $m = 1$), as well as ε^2-hint compression (by using the output solely for its excess loss guarantee). Thus, the minimax optimality of Theorem 6 for this family follows from our recalibration lower bound in Theorem 8 and Theorem 3.6 of [CBL06].

Interestingly, a Lipschitz assumption as in Assumption 1, or another mixability criterion, is necessary for $(\varepsilon, \varepsilon^2)$-multi-recalibration in $\approx \varepsilon^{-3}$ iterations. This is because even ε^2-hint compression against arbitrary bounded proper losses requires $T = \Omega(\log m \cdot \varepsilon^{-4})$ iterations (Theorem 3.7, [CBL06]).

D.3 Multi-calibeating and calibration

We conclude by stating an analog of Theorem 5 for multiple hint sequences.

Theorem 7. *Let $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$ be proper such that Assumptions 1, 2 hold, and let $\varepsilon \in (0, \frac{1}{2})$. In the setting of Model 2, if $q^i_{[T]}$ is s-sparse for every i, there is an algorithm that outputs $p_{[T]} \in [0,1]^T$ that is ε-calibrated against $y_{[T]}$ and ε^2-calibeats $(q^{[m]}_{[T]}, y_{[T]})$ under ℓ, if for an appropriate constant,*

$$T = \Omega\left(\frac{L^2}{\varepsilon^3} + \frac{\gamma s \log\left(\frac{\gamma s}{\varepsilon}\right)}{\varepsilon^2} + \frac{L \log m}{\varepsilon^2}\right).$$

Proof. This theorem follows from combining Theorem 5, Proposition 2 and Theorem 4. Concretely, we apply Theorem 5 to each hint sequence separately to get m refined sequences that each calibeats the corresponding original sequence. We then apply Proposition 2 to combine the m refined sequences into a single sequence that multi-calibeats the m original hint sequences. Finally, we apply Theorem 4 to obtain calibration while maintaining the multi-calibeating property. □

E Lower Bound

In this section, we give a hard instance that demonstrates the recalibration rates achieved by our algorithm in Appendix B are optimal up to constant factors. Throughout, we fix a target $\varepsilon \in (0, \frac{1}{2})$ and lower bound the number of iterations required for an online algorithm to achieve $(\varepsilon, \varepsilon^2)$-recalibration (against the squared loss). For notational convenience, in this section we define

$$\mathcal{N} := \{i\varepsilon\}_{i \in [\lfloor \varepsilon^{-1} \rfloor]} \tag{30}$$

be an ε-net of $[0,1]$. We also define the *Brier score* (average squared loss)

$$\mathcal{B}_2\left(p_{[T]}, y_{[T]}\right) := \frac{1}{T} \sum_{t \in [T]} (p_t - y_t)^2 \tag{31}$$

for any T-length sequences $p_{[T]} \in [0,1]^T$, $y_{[T]} \in \{0,1\}^T$. We now outline our hard instance.

Construction 1. *Let $\varepsilon \in (0, \frac{1}{2})$ and let $T \in \mathbb{N}$. Following notation (30), let $\mathcal{M} := \mathcal{N} \cap [\frac{1}{4}, \frac{3}{4}] = \{v_i\}_{i \in [m]}$ where $m := |\mathcal{M}|$. Our hard instance is as follows: for $t \in [T]$, we let*

$$q_t = v_i \text{ where } i \equiv t \pmod{m}, \quad y_t \sim \mathsf{Bern}(q_t).$$

In other words, for $L := \lceil \frac{T}{m} \rceil$, we loop through all elements of $\mathcal{M}$ each for $L-1$ or L times in total, set $q_t = v_i$ for an appropriate index $i \in [m]$, and sample $y_t \sim \mathsf{Bern}(q_t)$. This is the same hard instance as is described in Section 3.1.1, [CLNR26].

Observe that the initial predictions $\{q_t\}_{t\in[T]}$ truthfully capture the generation process for the labels $\{y_t\}_{t\in[T]}$. Moreover, it is straightforward to see that the initial calibration error satisfies

$$\mathcal{K}_1(q_{[T]}, y_{[T]}) = O\left(\frac{1}{\sqrt{L}}\right)$$

because the variance of each $y_t \sim \mathsf{Bern}(v_i)$ is $\Theta(1)$, and each v_i is set to q_t for either L or $L+1$ times. Thus, it is enough to take $T = O(L\varepsilon^{-1}) = O(\varepsilon^{-3})$ for the initial predictions q_t to already achieve ε-calibration, i.e., setting $p_{[T]} = q_{[T]}$ trivially is a $(\varepsilon, 0)$-recalibration of $q_{[T]}$.

We show that there is (asymptotically) no better strategy. In Appendix E.1 we first show that up to negligible loss, we may assume that all predictions are in $\mathcal{N}$. We then lower bound the rounded miscalibration error in Appendix E.2, with our main result stated as Theorem 8.

E.1 Rounding

Throughout this section, let $\mathcal{F}_t$ be the filtration generated by $\{q_{[t]}, p_{[t-1]}, y_{[t-1]}\}$, so that our t^{th} prediction p_t is $\mathcal{F}_t$-measurable and $p_t \perp y_t \mid \mathcal{F}_t$. We similarly let $\mathcal{F}'_t$ be the filtration generated by $\{q_{[t]}, p_{[t]}, y_{[t]}\}$, i.e., including the t^{th} prediction and label. We also define

$$r_t := \arg\min_{v\in\mathcal{N}} |v - p_t| \tag{32}$$

to be the deterministic rounding of p_t to our net $\mathcal{N}$, breaking ties arbitrarily but consistently.

We next relate the calibration error and squared loss regret of the rounded predictions $r_{[T]}$ to those achieved by the predictions $p_{[T]}$, allowing us to focus on the former sequence.

Lemma 11. *For any $p_{[T]} \in [0,1]^T$, $y_{[T]} \in \{0,1\}^T$, following notation (32),*

$$\mathcal{K}_1\left(r_{[T]}, y_{[T]}\right) \le \mathcal{K}_1\left(p_{[T]}, y_{[T]}\right) + \varepsilon.$$

Proof. Throughout this proof, for each $v \in \mathcal{N}$, let $I_v \subseteq [0,1]$ denote the set of possible p such that $r_t = v$ if $p_t = p$ (i.e., the set of p that would be rounded to v using (32)). We have

$$\begin{aligned}
\mathcal{K}_1\left(r_{[T]}, y_{[T]}\right) &= \frac{1}{T}\sum_{v\in\mathcal{N}}\left|\sum_{t\in[T]:r_t=v}(v - y_t)\right| = \frac{1}{T}\sum_{v\in\mathcal{N}}\left|\sum_{t\in[T]:p_t\in I_v}(v - y_t)\right| \\
&\le \frac{1}{T}\sum_{v\in\mathcal{N}}\left(\left|\sum_{t\in[T]:p_t\in I_v}(p_t - y_t)\right| + \sum_{t\in[T]:p_t\in I_v}|p_t - v|\right) \\
&\le \frac{1}{T}\sum_{v\in\mathcal{N}}\left|\sum_{t\in[T]:p_t\in I_v}(p_t - y_t)\right| + \varepsilon.
\end{aligned}$$

Here, the last inequality used that every $p_t \in I_v$ has $|p_t - v| \le \varepsilon$. Next, for each $v \in \mathcal{N}$,

$$\left|\sum_{t\in[T]:p_t\in I_v}(p_t - y_t)\right| \le \sum_{p\in I_v}\left|\sum_{t\in[T]:p_t=p}(p - y_t)\right|$$

by the triangle inequality. Finally, combining the above two displays gives the claim. □

In the remainder of the section, any unconditional expectations are always with respect to $\mathcal{F}'_T$.

Lemma 12. *For any $p_{[T]} \in [0,1]^T$, following Construction 1 and notation (32),*

$$\mathbb{E}\left[\mathcal{B}_2\left(r_{[T]}, y_{[T]}\right) - \mathcal{B}_2\left(q_{[T]}, y_{[T]}\right)\right] \le 2\mathbb{E}\left[\mathcal{B}_2\left(p_{[T]}, y_{[T]}\right) - \mathcal{B}_2\left(q_{[T]}, y_{[T]}\right)\right] + 2\varepsilon^2.$$

Proof. Because of the bias-variance decomposition

$$\mathbb{E}\left[(p_t - y_t)^2 \mid \mathcal{F}_t\right] = \mathbb{E}\left[(q_t - y_t)^2 \mid \mathcal{F}_t\right] + \mathbb{E}\left[(p_t - q_t)^2 \mid \mathcal{F}_t\right], \tag{33}$$

and a similar decomposition for $\mathbb{E}[(r_t - y_t)^2 \mid \mathcal{F}_t]$, it is enough to show that for every $t \in [T]$,

$$\mathbb{E}\left[(r_t - q_t)^2 \mid \mathcal{F}_t\right] \le 2\mathbb{E}\left[(p_t - q_t)^2 \mid \mathcal{F}_t\right] + 2\varepsilon^2.$$

This follows deterministically from $(a+b)^2 \le 2a^2 + 2b^2$, and $|p_t - r_t| \le \varepsilon$. □

Combining Lemmas 11 and 12, we have shown that any $(\varepsilon, \varepsilon^2)$-recalibration algorithm playing predictions $p_{[T]} \in [0,1]^T$ implies a $(2\varepsilon, 4\varepsilon^2)$-recalibration algorithm playing predictions $r_{[T]} \in \mathcal{N}^T$. Henceforth we focus on arguing the latter goal is impossible if $T = o(\varepsilon^{-3})$.

E.2 Lower bounding miscalibration

In this section, let $r_t \in \mathcal{N}$ be a $\mathcal{F}_t$-measurable prediction restricted to being on the net $\mathcal{N}$, for all $t \in [T]$. Our goal is to lower bound $\mathcal{K}_1(r_{[T]}, y_{[T]})$ assuming that $r_{[T]}$ achieves good Brier score. Our starting point is a bias-miscalibration decomposition (Lemma 13), where we define

$$\nu_v := \sum_{t \in [T]} \mathbb{I}_{r_t = v}, \quad \mu_v := \sum_{t \in [T]} \mathbb{I}_{r_t = v}(q_t - y_t), \tag{34}$$

for all $v \in \mathcal{N}$. Note that ν_v and μ_v are $\mathcal{F}'_T$-measurable random variables, that respectively correspond to the times $r_t = v$ was chosen, and the total miscalibration of $q_{[T]}, y_{[T]}$ on those rounds.

Lemma 13. *Following Construction 1 and notation (32), (34),*

$$\mathbb{E}\left[\mathcal{K}_1\left(r_{[T]}, y_{[T]}\right)\right] \ge \frac{1}{T}\sum_{v \in \mathcal{N}} \mathbb{E}\left[|\mu_v|\right] - \sqrt{\mathbb{E}[\mathcal{B}_2(r_{[T]}, y_{[T]}) - \mathcal{B}_2(q_{[T]}, y_{[T]})]}.$$

Proof. In addition to (34) we also denote the per-prediction bias by

$$\beta_v := \sum_{t \in [T]} \mathbb{I}_{r_t = v}(r_t - q_t),$$

for all $v \in \mathcal{N}$. Then,

$$\mathcal{K}_1\left(r_{[T]}, y_{[T]}\right) = \frac{1}{T}\sum_{v \in \mathcal{N}} |\beta_v + \mu_v| \ge \frac{1}{T}\sum_{v \in \mathcal{N}} |\mu_v| - \frac{1}{T}\sum_{v \in \mathcal{N}} |\beta_v|.$$

By taking expectations, and bounding

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{v \in \mathcal{N}} |\beta_v|\right] &\le \frac{1}{T}\mathbb{E}\left[\sum_{v \in \mathcal{N}} \sqrt{\nu_v \cdot \sum_{t \in [T]} \mathbb{I}_{r_t = v}(r_t - q_t)^2}\right] \\
&\le \frac{1}{T}\mathbb{E}\left[\sqrt{\left(\sum_{v \in \mathcal{N}} \nu_v\right)\left(\sum_{t \in [T]} (r_t - q_t)^2\right)}\right] \\
&= \mathbb{E}\left[\sqrt{\frac{1}{T}\sum_{t \in [T]} (r_t - q_t)^2}\right] \le \sqrt{\mathbb{E}\left[\frac{1}{T}\sum_{t \in [T]} (r_t - q_t)^2\right]},
\end{aligned}$$

we have the claim upon applying the decomposition (33). The first and second inequalities above were the Cauchy-Schwarz inequality, and the last applied Jensen's inequality. □

Lemma 13 shows that to rule out a rounded predictor sequence $r_{[T]} \in \mathcal{N}^T$ achieving $(\varepsilon, \varepsilon^2)$-recalibration (in expectation) where $T = o(\varepsilon^{-3})$, it is enough to show

$$\sum_{v \in \mathcal{N}} \mathbb{E}[|\mu_v|] = \Omega(\varepsilon^{-2}).$$

We next develop tools for lower bounding $\sum_{v \in \mathcal{N}} \mathbb{E}[|\mu_v|]$. We begin with a technical helper result.

Lemma 14. *Following Construction 1 and notation (32), (34), for all $v \in \mathcal{N}$,*

$$\mathbb{E}\left[\mu_v^4\right] \le 66(\mathbb{E}\left[\nu_v^2\right] + 1).$$

Proof. To ease notation, for all $t \in [T]$ let

$$I_t := \mathbb{I}_{r_t = v}, \quad \Delta_t := I_t(q_t - y_t),$$

so that the following bounds hold for all $t \in [T]$: $\mathbb{E}[\Delta_t \mid \mathcal{F}_t] = 0$, and as $q_t \in [\frac{1}{4}, \frac{3}{4}]$ always,

$$\mathbb{E}[\Delta_t^2 \mid \mathcal{F}_t] = q_t(1-q_t)I_t \le \frac{1}{4}I_t, \quad \mathbb{E}[\Delta_t^3 \mid \mathcal{F}_t] = q_t(1-q_t)(2q_t-1)I_t \le \frac{1}{4}I_t. \tag{35}$$

Also, we define $\sigma_t := \sum_{s\in[t]} \Delta_s$, so that $\mu_v = \sigma_T$. We bound the conditional increment in $\mathbb{E}[\sigma_t^4]$:

$$\begin{aligned}
\mathbb{E}\left[\sigma_t^4 - \sigma_{t-1}^4 \mid \mathcal{F}_t\right] &= 6\sigma_{t-1}^2\mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_t\right] + 4\sigma_{t-1}\mathbb{E}\left[\Delta_t^3 \mid \mathcal{F}_t\right] + \mathbb{E}\left[\Delta_t^4 \mid \mathcal{F}_t\right] \\
&\le 6\sigma_{t-1}^2\mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_t\right] + 2\sigma_{t-1}^2\mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_t\right] + 2\mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_t\right] + I_t \\
&= 8\sigma_{t-1}^2\mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_t\right] + 2\mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_t\right] + I_t,
\end{aligned}$$

where the second line above uses that the second bound in (35) is dominated in magnitude by the first. Summing across all iterations we have

$$\mathbb{E}\left[\sigma_T^4\right] \le 8\sum_{t\in[T]} \mathbb{E}\left[\sigma_{t-1}^2\Delta_t^2\right] + 2\sum_{t\in[T]} \mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_t\right] + \mathbb{E}\left[\nu_v\right] \le 8\sum_{t\in[T]} \mathbb{E}\left[\sigma_{t-1}^2\Delta_t^2\right] + 2\mathbb{E}\left[\nu_v\right],$$

again using (35). Finally, Jensen's inequality implies $\mathbb{E}[\sigma_T^2 \mid \mathcal{F}_t'] \ge \sigma_{t-1}^2$ for all $t \in [T]$, so

$$\begin{aligned}
\sum_{t\in[T]} \mathbb{E}\left[\sigma_{t-1}^2\Delta_t^2\right] \le \mathbb{E}\left[\sigma_T^2 \sum_{t\in[T]} \Delta_t^2\right] &\le \sqrt{\mathbb{E}\left[\sigma_T^4\right]\mathbb{E}\left[\left(\sum_{t\in[T]} \Delta_t^2\right)^2\right]} \\
&\le \sqrt{\mathbb{E}\left[\sigma_T^4\right]\mathbb{E}\left[\nu_v^2\right]} \le \frac{1}{16}\mathbb{E}\left[\sigma_T^4\right] + 4\mathbb{E}\left[\nu_v^2\right].
\end{aligned}$$

Combining the above two displays gives the conclusion, as $\sigma_T^4 = \mu_v^4$ and $2\mathbb{E}[\nu_v] \le \mathbb{E}[\nu_v^2] + 1$:

$$\mathbb{E}\left[\mu_v^4\right] \le \frac{1}{2}\mathbb{E}\left[\mu_v^4\right] + 33\mathbb{E}\left[\nu_v^2\right] + 1.$$

□

We next give our main miscalibration bound, inspired by the proof of Proposition 1, [CLNR26].

Lemma 15. *Following Construction 1 and notation (32), (34),*

$$\sum_{v\in\mathcal{N}} \mathbb{E}\left[|\mu_v|\right] \ge \frac{T^{\frac{3}{2}}}{48\sqrt{\sum_{v\in\mathcal{N}}(\mathbb{E}[\nu_v^2] + 1)}}.$$

Proof. By Hölder's inequality with exponents $\frac{1}{3}, \frac{2}{3}$,

$$\mathbb{E}\left[\mu_v^2\right] = \mathbb{E}\left[|\mu_v|^{\frac{2}{3}} \cdot |\mu_v|^{\frac{4}{3}}\right] \le \mathbb{E}\left[|\mu_v|\right]^{\frac{2}{3}}\mathbb{E}\left[\mu_v^4\right]^{\frac{1}{3}} \implies \mathbb{E}\left[|\mu_v|\right] \ge \frac{\mathbb{E}[\mu_v^2]^{\frac{3}{2}}}{\mathbb{E}[\mu_v^4]^{\frac{1}{2}}} \ge \frac{\mathbb{E}[\mu_v^2]^{\frac{3}{2}}}{9\sqrt{\mathbb{E}[\nu_v^2] + 1}},$$

where we used Lemma 14 in the last step. Further, note that

$$\mathbb{E}[\mu_v^2] = \mathbb{E}\left[\left(\sum_{t\in[T]} \mathbb{I}_{r_t=v}(q_t - y_t)\right)^2\right] = \mathbb{E}\left[\sum_{t\in[T]} \mathbb{I}_{r_t=v}(q_t - y_t)^2\right] \ge \frac{3}{16}\mathbb{E}\left[\nu_v\right],$$

where the last inequality is because the variance $\mathbb{E}[(q_t - y_t)^2] = q_t(1-q_t) \ge \frac{3}{16}$ under Construction 1, and the second equality is because all cross-terms vanish: for $s < t$,

$$\mathbb{E}\left[\mathbb{I}_{r_t=v}\mathbb{I}_{r_s=v}(q_t - y_t)(q_s - y_s)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_{r_t=v}(q_t - y_t) \mid \mathcal{F}_s'\right]\mathbb{I}_{r_s=v}(q_s - y_s)\right] = 0.$$

Combining the above displays, we have shown that for all $v \in \mathcal{N}$,

$$\mathbb{E}\left[|\mu_v|\right] \geq \frac{\mathbb{E}[\nu_v]^{\frac{3}{2}}}{48\sqrt{\mathbb{E}[\nu_v^2]+1}}. \tag{36}$$

Finally, letting $\alpha_v := \mathbb{E}[\nu_v]$ and $\beta_v := \mathbb{E}[\nu_v^2]$, Hölder's inequality gives

$$T = \sum_{v\in\mathcal{N}} \alpha_v = \sum_{v\in\mathcal{N}} \left(\frac{\alpha_v^{\frac{3}{2}}}{\beta_v^{\frac{1}{2}}}\right)^{\frac{2}{3}} \beta_v^{\frac{1}{3}} \leq \left(\sum_{v\in\mathcal{N}} \frac{\alpha_v^{\frac{3}{2}}}{\beta_v^{\frac{1}{2}}}\right)^{\frac{2}{3}} \left(\sum_{v\in\mathcal{N}} \beta_v\right)^{\frac{1}{3}}$$

so that

$$\sum_{v\in\mathcal{N}} \frac{\mathbb{E}[\nu_v]^{\frac{3}{2}}}{\sqrt{\mathbb{E}[\nu_v^2]+1}} \geq \frac{T^{\frac{3}{2}}}{\sqrt{\sum_{v\in\mathcal{N}}(\mathbb{E}[\nu_v^2]+1)}}.$$

Summing (36) across all $v \in \mathcal{N}$, and using the above bound, gives the desired claim. □

To make Lemma 15 more usable, we next upper bound $\sum_{v\in\mathcal{N}} \mathbb{E}[\nu_v^2]$.

Lemma 16. *Following Construction 1 and notation (32), (34),*

$$\sum_{v\in\mathcal{N}} \mathbb{E}[\nu_v^2] \leq 128\left(\left(\varepsilon T^2 + \frac{1}{\varepsilon}\right) + \left(\frac{T}{\varepsilon} + \frac{1}{\varepsilon^2}\right)\mathbb{E}\left[\sum_{t\in[T]}(q_t - r_t)^2\right]\right).$$

Proof. We will prove the stronger statement

$$\sum_{v\in\mathcal{N}} \nu_v^2 \leq \varepsilon T^2 + \frac{T}{\varepsilon}\sum_{t\in[T]}(q_t - r_t)^2 \tag{37}$$

for any realization of the $\{q_t, r_t\}_{t\in[T]}$. Taking expectations then gives the claim.

To prove (37), we set up some helpful notation. For all $v \in \mathcal{N}$, let $c_v \in \{0, L-1, L\}$ be the number of times that $q_t = v$ in Construction 1, where $L := \lceil \frac{T}{m} \rceil$ is the number of loops. Let

$$a_v := \max\left\{\nu_v - c_v, 0\right\}, \quad b_v := \sum_{t\in[T]}(q_t - r_t)^2 \mathbb{I}_{r_t = v}$$

respectively denote the "overflow" of v (the number of excess times it was selected as r_t compared to q_t), and the unscaled contribution of v to the second term in (37).

The key observation is that of the a_v excess times $r_t = v$ was chosen, at most $2L$ of the corresponding q_t can satisfy $|q_t - r_t| \leq \varepsilon$, i.e., the ones from neighboring buckets $q_t \in \{r_t - \varepsilon, r_t + \varepsilon\}$. Similarly, at most $2L$ more can satisfy $|q_t - r_t| \leq 2\varepsilon$, and so on. Thus,

$$b_v \geq \sum_{i\in[a_v]} \left(\varepsilon\left\lceil\frac{i}{2L}\right\rceil\right)^2 \geq \sum_{i\in[a_v]} \frac{\varepsilon^2 i}{2L} \geq \frac{\varepsilon^2 a_v^2}{4L}.$$

Thus we expand

$$\nu_v^2 \leq (a_v + c_v)^2 \leq 2c_v^2 + 2a_v^2 \leq 2L^2 + \frac{8Lb_v}{\varepsilon^2}.$$

To conclude the proof we sum over all $v \in \mathcal{N}$. If $L = 1$ then (37) clearly holds. Otherwise, observe that $m \in [\frac{1}{4\varepsilon}, \frac{1}{\varepsilon}]$ for $\varepsilon \in (0, \frac{1}{2})$. If $L \geq 2$, we have $L \leq \frac{2T}{m} \leq 8\varepsilon T$, so

$$\sum_{v\in\mathcal{N}} \nu_v^2 \leq \frac{2L^2}{\varepsilon} + \frac{8L}{\varepsilon^2}\sum_{t\in[T]}(q_t - r_t)^2 \leq 128\varepsilon T^2 + \frac{64T}{\varepsilon}\sum_{t\in[T]}(q_t - r_t)^2.$$

□

Finally, we can conclude with our main lower bound.

Theorem 8. *In the setting of Construction 1, no $p_{[T]}$ can be $(\varepsilon, \varepsilon^2)$-recalibrated against $(q_{[T]}, y_{[T]})$ under ℓ_{sq}, unless, for an appropriate constant, $T = \Omega(\varepsilon^{-3})$.*

Proof. We will instead prove that, following notation (32), it is impossible to obtain

$$\mathbb{E}\left[\mathcal{K}_1\left(r_{[T]}, y_{[T]}\right)\right] \le 2\varepsilon, \quad \mathbb{E}\left[\mathcal{B}_2\left(r_{[T]}, y_{[T]}\right) - \mathcal{B}_2\left(q_{[T]}, y_{[T]}\right)\right] \le 4\varepsilon^2, \tag{38}$$

unless $T = \Omega(\varepsilon^{-3})$, from which the conclusion follows by applying Lemmas 11 and 12. Assume for contradiction that (38) were possible. Under the second bound in (38), Lemma 16 gives

$$\sum_{v \in \mathcal{N}} \mathbb{E}\left[\nu_v^2\right] \le 128\left(\varepsilon T^2 + \frac{1}{\varepsilon} + \varepsilon T^2 + T\right) \le 400\left(\varepsilon T^2 + \frac{1}{\varepsilon}\right).$$

where we used $2T \le \varepsilon T^2 + \frac{1}{\varepsilon}$. Plugging this into Lemma 15,

$$\frac{1}{T}\sum_{v \in \mathcal{N}} \mathbb{E}\left[|\mu_v|\right] \ge \frac{\sqrt{T}}{960\sqrt{\varepsilon T^2 + \frac{2}{\varepsilon}}}. \tag{39}$$

On the other hand, Lemma 13 and the bounds in (38) imply

$$\frac{1}{T}\sum_{v \in \mathcal{N}} \mathbb{E}\left[|\mu_v|\right] \le 2\varepsilon + \sqrt{4\varepsilon^2} = 4\varepsilon.$$

We conclude by casework on the denominator of (39). If $\varepsilon T^2 \le \frac{1}{\varepsilon} \iff T \le \frac{1}{\varepsilon}$, then

$$4\varepsilon \ge \frac{\sqrt{T}}{960\sqrt{\frac{3}{\varepsilon}}} \implies \varepsilon \ge \frac{T}{3(3840)^2}$$

is lower bounded by a constant, at which point every $T = \Omega(\varepsilon^{-3})$. Otherwise, if $\varepsilon T^2 \ge \frac{1}{\varepsilon}$,

$$4\varepsilon \ge \frac{\sqrt{T}}{960\sqrt{3\varepsilon T^2}} \implies T = \Omega(\varepsilon^{-3}).$$

□

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: We claimed and formally proved improvements for calibeating and recalibration in the online learning setting. We also run experiments to validate our results.

   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: We discussed our limits and potential improvements in the open problems paragraph in Section 3.

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: We provide self-contained proofs for all results.

   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: [Yes]

   Justification: We give a full description of our experimental settings and provide code in the supplementary materials for reproducibility.

   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: [Yes]

   Justification: We provide code in supplementary material and are ready to provide a (deanonymized) GitHub repository if our paper is accepted.

   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: [Yes]

   Justification: We include all details in our experiment section.

   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: [Yes]

   Justification: All of our experiments report variance statistics across repeated runs.

   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: [Yes]

   Justification: We discussed the computation resources in the experiment section.

   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: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

   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: [N/A]

    Justification: Our work is inspired by an interpretability definition, calibration, so there may be an ultimately positive social impact. However, our work is mainly theoretical, and there is no immediate social impact of our work as far as the authors could think of.

    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: We used CIFAR-10 and CIFAR-100 which are both publicly accessible datasets.

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: [Yes]

Justification: We cited our datasets used.

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: We do 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: [Yes]

Justification: In Section 3, we declared our LLM usage for aiding in the proofs in Appendix E, and assume full responsibility for their correctness. All other ideas are fully original.

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.