A Additional Results

A.1 Multiclass Prediction

For multiclass prediction, we suppose that Y can take K distinct values. We denote Δ^K as the K-dimensional probability simplex. For notational convenience we represent Y as a one-hot vector in $\mathbb{R}^K$, so $\mathcal{Y} = \{(1, 0, \cdots), (0, 1, \cdots), \cdots\}$.

Protocol 3: Decision Making with Bets, Multiclass At time $t = 1, \cdots, T$

1. Nature reveals $x_t \in \mathcal{X}$ and chooses $\mu_t^* \in \Delta^K$ without revealing it
2. Forecaster reveals $\mu_t \in \Delta^K$ and $c_t \in R_+^K$
3. Agent t has loss $l_t : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ and chooses action a_t and $g_t \in \mathbb{R}^K$
4. Sample $y_t \sim \text{Categorical}(\mu_t^*)$ and reveal y_t
5. Agent total loss is $l_t(y_t, a_t) - \langle g_t, y_t - \mu_t \rangle + \langle |g_t|, c_t \rangle$, forecaster loss is $\langle g_t, y_t - \mu_t \rangle - \langle |g_t|, c_t \rangle$

As before we require the regularity condition that $\mu_c + c_t \in [0, 1]^K$ and $\mu_t - c_t \in [0, 1]^K$ (even though these are no longer on Δ^K, hence not probabilities.

Similar to Section 3 we can denote the agent's maximum / minimum expected loss under the forecasted probability as

$$L_t^{\max} = \max_{\tilde{\mu} \in \Delta^K, \tilde{\mu} \in \mu_t \pm c_t} \mathbb{E}_{Y \sim \tilde{\mu}}[l_t(a_t, Y)]$$
$$L_t^{\min} = \min_{\tilde{\mu} \in \Delta^K, \tilde{\mu} \in \mu_t \pm c_t} \mathbb{E}_{Y \sim \tilde{\mu}}[l_t(a_t, Y)]$$

and true expected loss as $L_t^* = \mathbb{E}_{Y \sim \mu_t^*}[l_t(a_t, Y)]$. As before denote

$$\mathrm{L}_t^{\text{pay}} = L_t^* + \mathbb{E}_{\mu^*}[\langle g_t, \mu_t - Y \rangle + \langle |g_t|, c_t \rangle]$$

Proposition 3. *If* $g_t = l(\cdot, a_t) - \inf_{\gamma \in \mathbb{R}} \langle c_t, |l - \gamma 1| \rangle$ *then* $L_t^{\text{pay}} = L_t^{\max}$

Proof of Proposition 3. As a notation shorthand we denote $l_t(a_t, Y)$ with the vector l, such that $l_i = l_t(a_t, Y = i)$. We first show a closed form solution for $L_t^{\max}$ which can be written as

$$\begin{aligned}
L_t^{\max} &= \sup_{\tilde{\mu} \in \Delta^K, \tilde{\mu} \in \mu_t \pm c_t} \mathbb{E}_{Y \sim \tilde{\mu}}[l_t(a_t, Y)] \\
&= \sup_{\tilde{\mu} \in \Delta^K, \tilde{\mu} \in \mu_t \pm c_t} \langle \tilde{\mu}, l \rangle && \text{Notation Change} \\
&= \langle \mu_t, l \rangle + \sup_{\delta\mu \in [-c_t, c_t], \langle \delta\mu, 1 \rangle = 0} \langle \delta\mu, l \rangle && \text{Algebric Manipulation} \\
&= \langle \mu_t, l \rangle + \sup_{\delta\mu \in [-c_t, c_t]} \inf_{\gamma \in \mathbb{R}} \langle \delta\mu, l \rangle - \gamma \langle \delta\mu, 1 \rangle && \text{Lagrangian} \\
&= \langle \mu_t, l \rangle + \inf_{\gamma \in \mathbb{R}} \sup_{\delta\mu \in [-c_t, c_t]} \langle \delta\mu, l \rangle - \gamma \langle \delta\mu, 1 \rangle && \text{Sion Minimax Theorem} \\
&= \langle \mu_t, l \rangle + \inf_{\gamma \in \mathbb{R}} \langle c_t, |l - \gamma 1| \rangle
\end{aligned}$$

Similarly we have

$$L_t^{\min} = \langle \mu_t, l \rangle - \inf_{\gamma \in \mathbb{R}} \langle c_t, |l - \gamma 1| \rangle$$

Denote the γ that achieves the infimum as γ^*. Comparing with L_t^{pay} we have

$$\begin{aligned}
L_t^{\text{pay}} &= L_t^* + \mathbb{E}_{\mu^*}[\langle g_t, \mu_t - Y \rangle + \langle |g_t|, c_t \rangle] \\
&= \langle \mu^*, l \rangle - \langle l - \gamma^* 1, \mu_t - \mu_t^* \rangle + \langle c_t, |l - \gamma^* 1| \rangle \\
&= \langle l, \mu_t \rangle + \langle c_t, |l - \gamma^* 1| \rangle && \langle \mu_t, 1 \rangle = 0, \langle \mu_t^*, 1 \rangle = 0 \\
&= L^{\max}
\end{aligned}$$

□

A.2 Offline Calibration

For this section we restrict to the i.i.d. setup, where we assume there are random variables X, Y with some distribution p^*_{XY} such that at each time step,

$$x_t \sim X \qquad \mu^*_t = \mathbb{E}[Y \mid x_t]$$

We also assume that the forecaster 's choice μ_t, c_t and the agent's choice b_t in Protocol 2 are computed by functions of x_t

$$\mu : x_t \mapsto \mu_t \quad c : x_t \mapsto c_t \quad b : x_t \mapsto b_t$$

In other words, given the input x_t all the players choose their actions based on fixed functions of x_t.

The following definition is the equivalent of asymptotic soundness in the i.i.d. setup

Definition 1. *We say that the functions $\mu, c : \mathcal{X} \to [0,1]$ are sound with respect to some set of functions $\mathcal{B} \subset \{\mathcal{X} \to [-M, M]\}$ if*

$$\sup_{b \in \mathcal{B}} \mathbb{E}[b(X)(\mu(X) - \mathbb{E}[Y \mid X]) - |b(X)|c(X)] \leq 0$$

If $c(x) \equiv 0$ we say μ is $\mathcal{B}$-calibrated.

Intuitively if μ, c are sound with respect to $\mathcal{B}$ then if the decision making agents chooses a strategy in $b \in \mathcal{B}$ we can guarantee that the forecaster will not lose (on average). In other words, if μ, c are sound according to Definition 1, and if $b_t = b(x_t)$ for some $b \in \mathcal{B}$, then the forecaster is almost surely asymptotically sound as defined in Eq.(4).

A.2.1 Examples and Special Cases

Standard Calibration Standard calibration is defined as: for any $u \in [0,1]$, among the X where $\mu(X) = u$ it is indeed true that Y is 1 with u probability. Formally this can be written as

$$\mathbb{E}[Y \mid \mu(X) = u] = u, \forall u \in [0,1]$$

Deviation from this ideal situation is measured by the maximum calibration error (MCE).

$$\mathrm{MCE}(\mu) = \max_{u \in [0,1]} |\mathbb{E}[Y \mid \mu(X) = u] - u|$$

Note that the MCE may be ill-defined if there is an interval $(u_0, u_1) \subset [0,1]$ such that $\mu(X) \in (u_0, u_1)$ with zero probability. We are going to avoid the technical subtlety by assuming that this does not happen, i.e. the distribution of $\mu(X)$ is supported on the entire set $[0,1]$.

When $\mathcal{B}$ is the set of all possible functions $\mu(x) \to \mathbb{R}$ (i.e. it only depends on the probability forecast $\mu(x)$ but not x itself), we obtain the standard definition of calibration (Dawid, 1985; Guo et al., 2017), as shown by the following proposition

Proposition 4. *The forecaster function $\mu : \mathcal{X} \to [0,1], c : x \mapsto c_0$ is sound with respect to $\mathcal{B} = \{x \mapsto \tilde{b}(\mu(x)), \tilde{b} : \mathbb{R} \to \mathbb{R}\}$ if and only if the MCE error of μ is less than c_0.*

Proof. See Appendix D □

We remark that this proposition (intentionally) does not involve the upper bound M on b; it holds even when $M \to \infty$.

Multi-Calibration Multi-calibration (Hébert-Johnson et al., 2017) achieves standard calibration for all subsets S in some collection of sets $\mathcal{S}$. The following proposition shows that a forecaster that's sound with respect to any function that only depends on $\mu(x)$ and takes zero value whenever $x \notin S$ is also multicalibrated.

Proposition 5. *Let $\mathcal{S} \subset 2^{\mathcal{X}}$. If a forecaster function $\mu : \mathcal{X} \to [0,1], c : x \mapsto c_0$ is sound with respect to $\mathcal{B} = \{x \mapsto \tilde{b}(\mu(x))\mathbb{I}(x \in S), S \in \mathcal{S}, \tilde{b} : \mathbb{R} \to \mathbb{R}\}$, then it is $(\mathcal{S}, c_0)$-multicalibrated.*

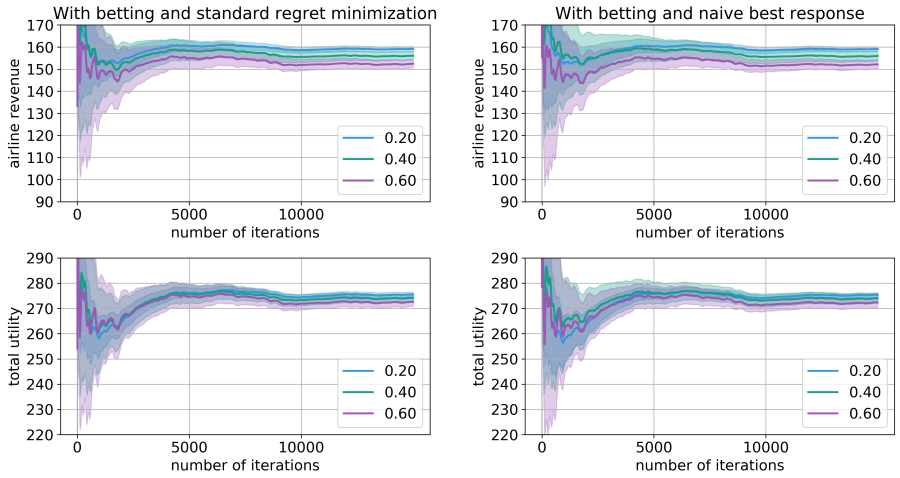

Figure 4: This plot extends Figure 1. We compare with additional Alternatives to Algorithm 3.

B Experiment Details and Additional Results

B.1 Airline Delay

Negative c_t In Protocol 2 c_t must be non-negative for its interpretation as a probability interval $[\mu_t - c_t, \mu_t + c_t]$. However if we only consider the flight delay insurance interpretation: airline pay passenger b_t^1 if flight delays and passenger pays airline $b_t^0 := \frac{b_t^1(\mu_t + c_t)}{1 - \mu_t - c_t}$ if flight doesn't delay. These payments are meaningful for both positive and negative c_t; the passenger utility (with insurance) can be computed as $r^{\text{trip}} - c^{\text{ticket}} - (\mu_t + c_t)c^{\text{delay}}$, which is also meaningful for both positive and negative c_t. We find that allowing negative c_t improves the stability of the algorithm.

Passenger Model We sample r^{alt} as $\text{Uniform}(0, 200)$ and sample r^{trip} from $\text{Uniform}(0, 400)$. We assume the cost of delay can be more varied, so we sample it from the following process: $z \sim \text{Uniform}(4, 9)$ and $c^{\text{delay}} = 0.2e^z$. This gives us a cost of delay between $[10, 1600]$, but large values are less likely.

Additional Results We show additional comparison with other alternatives to Algorithm 3 in Figure 4. For details about these alternatives see Section 7.

B.2 Additional Experiments

Decision Loss For each data point we associate an extra feature z used to define decision loss. For MNIST this is the digit label and for UCI Adult this is the age (binned by quantile into 10 bins). We simulate three kinds of decision losses; for each type of decision loss we randomly sample a few instantiations.

1. One-sided: we assume that $a \in [0, 1]$ and each decision loss $l(z, y, a)$ is large if $y \neq a$ and small if $y = a$. For different values of z there are different stakes (i.e. how much does the loss when $y = a$ differ from $y \neq a$).

2. Different Stakes: Each value of the decision loss $l(z, y, a)$ is a draw from $\mathcal{N}(0, z)$, which is used to capture the feature that certain groups of people have larger stakes

3. Random. Each value of the decision loss $l(z, y, a)$ is a draw from $\mathcal{N}(0, 10)$ but clipped to be within $[-10, 10]$.

Forecasted Loss vs. True Loss In Figure 6 we plot the relationship between the expected loss under the forecasted probability and the expected loss under the true probability (we can compute this for the MNIST dataset because the true probability is known as explained in Section 7). Even if we apply histogram binning recalibration (explained in Section 7), the individual probabilities are almost always incorrect.

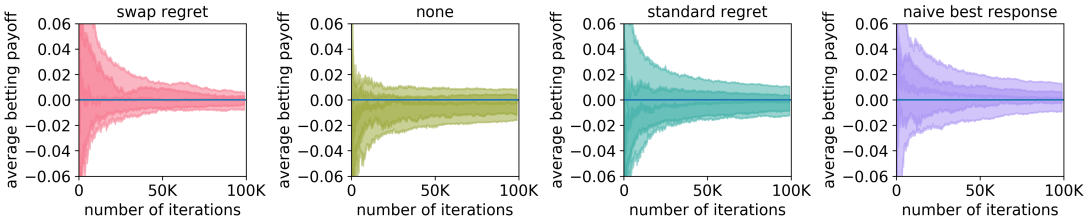

Figure 5: This plot is identical to Figure 2 but for the Adult dataset

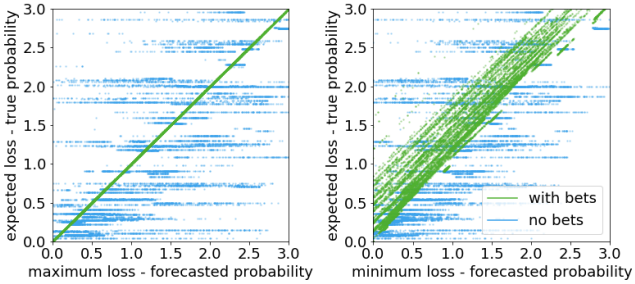

Figure 6: The expected loss under the forecaster utility vs. expected loss under the true probability. Each dot represents an individual probability forecast with a particular choice of loss function. We use histogram binning on the entire validation set to recalibrate the forecaster. Even though the forecaster is calibrated, the individual probabilities are often incorrect. Therefore, the expected loss under the forecasted probability often differs from the expected loss under the true probability (blue dots). On other hand, with additional payment from the bets, the expected total loss under true probability is always bounded between the minimum loss under the forecasted probability, and the maximum loss under the forecasted probability.

Asymptotic Exactness In Figure 2 and Figure 5 we plot the average betting loss of the forecaster. Algorithm 1 consistently achieve better asymptotic exactness compared to alternatives.

Average Interval Size In Figure 3 we plot the interval size c_t. A small c_t satisfies desideratum 2 in Section 3 and makes the guarantee in Proposition 2 useful for decision makers. We observe that most interval sizes are small, and larger intervals are exponentially unlikely.

C Proof of Theorem 1

Algorithm 3 is the core reason why our forecasting algorithm achieves Theorem 1, so before we prove Theorem 1 we first understand Algorithm 3. The goal of Algorithm 3 is to select a sequence of λ_t to minimize the loss $\sum_{t=1}^{T}(r_t + s_t\lambda_t)^2$ for any choice of $r_t, s_t \in \mathbb{R}$. More specifically the goal is to minimize the swap regret defined by

$$R_T^{\text{swap}} = \underbrace{\sum_{t=1}^{T}(r_t + s_t\lambda_t)^2}_{\text{Loss incurred by Algorithm 3}} - \inf_{\psi \in L^1[-1,1]} \underbrace{\sum_{t=1}^{T}(r_t + s_t\psi(\lambda_t))^2}_{\text{Loss incurred by ``alternative'' } \psi(\lambda_t)} \tag{7}$$

where $L^1[-1,1]$ denotes the set of 1-Lipshitz functions $\mathbb{R} \to [-1,1]$. Intuitively, $\sum_{t=1}^{T}(r_t + s_t\psi(\lambda_t))^2$ is the loss of an alternative algorithm: whenever Algorithm 3 selects λ_t, select $\psi(\lambda_t)$ instead. Swap regret measures the additional loss compared to the best alternative algorithm. We remark that if instead Algorithm 3 minimizes the standard regret, we can no longer guarantee Theorem 1. For intuition on the reason we refer interested readers to a counter-example in (Cesa-Bianchi and Lugosi, 2006) Section 4.5 (for a related calibration problem).

We now prove that Algorithm 3 indeed achieve its goal of minimizing the swap regret.

Theorem 2. *If there exists M_1, M_2 such that $\forall t, |s_t| \le M_1, |r_t/s_t| \le M_2$, then there exists a constant $C(M_1, M_2) > 0$, such that for any choice of $K > 1$, the regret of Algorithm 3 is bounded by*

$$R_T^{\text{swap}} \le C(M_1, M_2) K^2 \log T + \frac{1}{K^2} \sum_{t=1}^{T} s_t^2$$

In particular, if we choose $K^2 = \sqrt{T/\log T}$ then the swap regret R_T^{swap} is bounded by $O(\sqrt{T \log T})$.

Before we prove Theorem 2 we show how to use it to prove Theorem 1 restated below.

Theorem 1. *Suppose there is a constant $M > 0$ such that $\forall t$, $|b_t| \le M$, there exists an algorithm to output μ_t, c_t in Protocol 2 that is asymptotically exact for μ_t^*, b_t generated by any strategy of nature and agent. In particular, Algorithm 1 satisfies*

$$\left(\frac{1}{T} \sum_{t=1}^{T} b_t(\mu_t - y_t) - |b_t| c_t \right)^2 = O\left(\sqrt{\frac{\log T}{T}} \right) \tag{6}$$

Proof of Theorem 1. To prove this theorem we need the following inequality that relates the LHS in Eq.(6) to the swap regret R_T^{swap}

Lemma 2. *For any choice of $r_t, s_t, \lambda_t, t = 1, \cdots, T$ we have*

$$\left(\frac{1}{T} \sum_{t=1}^{T} s_t(r_t + s_t \lambda_t) \right)^2 \le \frac{R_T^{\text{swap}}}{T^2} \sum_{t=1}^{T} s_t^2$$

Because at each iteration Algorithm 1 selects $r_t = \frac{b_t}{\sqrt{|b_t|}}(\mu_t - y_t) - \sqrt{|b_t|}\hat{c}_t$ and $s_t = -\sqrt{|b_t|}$ we can plug this into Lemma 2 and conclude that for any sequence of λ_t (which includes any λ_t chosen by Algorithm 3), Algorithm 1 must satisfy

$$\left(\frac{-1}{T} \sum_{t=1}^{T} b_t(\mu_t - y_t) + |b_t|\hat{c}_t + |b_t|\lambda_t \right)^2 \le \frac{R_T^{\text{swap}}}{T} \frac{1}{T} \sum_{t=1}^{T} |b_t| \le \frac{M R_T^{\text{swap}}}{T}$$

In addition we have

$$\left| \frac{r_t}{s_t} \right| = \left| -\frac{b_t}{|b_t|}(\mu_t - y_t) + \hat{c}_t \right| \le 2$$

So the conditions of Theorem 2 is satisfied (i.e. $|s_t|$ and $|r_t/s_t|$ are bounded), and we can apply Theorem 2 to conclude $R_T^{\text{swap}} = O(\sqrt{T \log T})$. Combined we have

$$\left(\frac{1}{T} \sum_{t=1}^{T} b_t(\mu_t - y_t) - |b_t|(\hat{c}_t + \lambda_t) \right)^2 = O(M\sqrt{T \log T}/T) = O(\sqrt{\log T/T})$$

□

Now we proceed to prove Theorem 2

Proof of Theorem 2. To prove this theorem we first need the following Lemma, which bounds the standard regret (rather than swap regret)

Lemma 3. *If there exists some $M_1, M_2 > 0$ such that $\forall t$, $|\beta_t| \le M_1$ and $|\alpha_t/\beta_t| \le M_2$, choosing $\lambda_t = \arg\inf_{\lambda \in \mathbb{R}} \sum_{\tau=1}^{t-1} (\alpha_\tau + \beta_\tau \lambda)^2$ satisfies for some constant $C(M_1, M_2) > 0$*

$$\sum_{t=1}^{T} (\alpha_t + \beta_t \lambda_t)^2 \le \inf_{\lambda} \sum_{t=1}^{T} (\alpha_t + \beta_t \lambda)^2 + C(M_1, M_2) \log T$$

To prove Theorem 2 we first bound the discretized swap regret, defined as follows

$$\tilde{R}_T^{\text{swap}} = \sum_{t=1}^{T}(r_t + s_t\lambda_t)^2 - \sum_{k=1}^{K}\inf_{\lambda}\sum_{t=1}^{T}\mathbb{I}(\lambda_t \in [v_k, v_{k+1}))(r_t + s_t\lambda)^2$$

Intuitively, this is the regret with respect to the alternative algorithm: whenever the Algorithm 3 chooses some λ_t that falls with in a bin $[v_k, v_{k+1})$, choose a different λ.

To bound the discretized swap regret our proof strategy is similar to (Blum and Mansour, 2007): As a notation shorthand we denote $c^t(\lambda) = (r_t+s_t\lambda)^2$ and $\mathcal{I}_k = [v_k, v_{k+1})$. For each k, we apply Lemma 3 with $\alpha_t = r_t\mathbb{I}(\lambda_t \in \mathcal{I}_k)$ and $\beta_t = s_t\mathbb{I}(\lambda_t \in \mathcal{I}_k)$ with the convention that $0/0 = 0$. By the assumptions in Theorem 2 we know that $|\beta_t| \le M_1$ and $|\alpha_t/\beta_t| \le M_2$ so we can guarantee by Lemma 3 that

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{I}(\lambda_t \in \mathcal{I}_k)c^t(\lambda_t^k) &= \sum_{t=1}^{T}(\alpha_t + \beta_t\lambda_t^k)^2 \\ &\le \inf_{\lambda}\sum_{t=1}^{T}(\alpha_t + \beta_t\lambda)^2 + C(M_1, M_2)\log T = \inf_{\lambda}\sum_{t=1}^{T}\mathbb{I}(\lambda_t \in \mathcal{I}_k)c^t(\lambda) + C(M_1, M_2)\log T\end{aligned}$$

The total loss is given by

$$\sum_{t=1}^{T}c^t(\lambda_t) = \sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{I}(\lambda_t \in \mathcal{I}_k)c^t(\lambda_t^k) \le \sum_{k=1}^{K}\inf_{\lambda}\sum_{t=1}^{T}\mathbb{I}(\lambda_t \in \mathcal{I}_k)c^t(\lambda) + C(M_1, M_2)K\log T$$

We can conclude that

$$\tilde{R}_T^{\text{swap}} := \sum_{t=1}^{T}c^t(\lambda_t) - \sum_{k=1}^{K}\inf_{\lambda}\sum_{t=1}^{T}\mathbb{I}(\lambda_t \in \mathcal{I}_k)c^t(\lambda) \le C(M_1, M_2)K\log T$$

Finally we conclude the proof of the theorem with the following Lemma that bounds the difference between the discretized swap regret and the continuous swap regret.

Lemma 4. *In Algorithm 3,* $R_T^{\text{swap}} \le \tilde{R}_T^{\text{swap}} + \sum_{t=1}^{T} s_t^2\frac{v_K - v_0}{K}$

□

Proof of Lemma 4. Denote $\mathcal{I}_k = [v_k, v_{k+1})$ and denote $\delta v = \max_k v_{k+1} - v_k$. In addition denote $\lambda_k^* = \arg\inf_\lambda \sum_{t=1}^{T}\mathbb{I}(\lambda_t \in \mathcal{I}_k)(r_t + s_t\lambda)^2$

$$\begin{aligned}&R_T^{\text{swap}} - \tilde{R}_T^{\text{swap}} \\ &= \sum_{k=1}^{K}\inf_{\lambda}\sum_{t=1}^{T}\mathbb{I}(\lambda_t \in \mathcal{I}_k)(r_t + s_t\lambda)^2 - \inf_{\psi\in L^1}\sum_{t=1}^{T}(r_t + s_t\psi(\lambda_t))^2 && \text{Definition} \\ &= \sum_{k=1}^{K}\inf_{\lambda}\sum_{t=1}^{T}\mathbb{I}(\lambda_t \in \mathcal{I}_k)(r_t + s_t\lambda)^2 - \inf_{\psi\in L^1}\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{I}(\lambda_t \in \mathcal{I}_K)(r_t + s_t\psi(\lambda_t))^2 && \text{Decompose 2nd term} \\ &\le \sum_{k=1}^{K}\left(\inf_{\lambda}\sum_{t=1}^{T}\mathbb{I}(\lambda_t \in \mathcal{I}_k)(r_t + s_t\lambda)^2 - \inf_{\psi\in L^1}\sum_{t=1}^{T}\mathbb{I}(\lambda_t \in \mathcal{I}_K)(r_t + s_t\psi(\lambda_t))^2\right) && \text{Jensen} \\ &= \sum_{k=1}^{K}\left(\sum_{t=1}^{T}\mathbb{I}(\lambda_t \in \mathcal{I}_k)(r_t + s_t\lambda_k^*)^2 - \inf_{\delta\psi\in L^1}\sum_{t=1}^{T}\mathbb{I}(\lambda_t \in \mathcal{I}_K)(r_t + s_t(\lambda_k^* + \delta\psi(\lambda_t)))^2\right) && \text{Change of variable} \\ &\le \sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{I}(\lambda_t \in \mathcal{I}_k)s_t^2\delta_v^2 && \text{1-Lipschitzness}\end{aligned}$$

$$= \sum_{t=1}^{T} s_t^2 \delta_v^2$$

For Algorithm 3 we know that $\delta v = \frac{v_K - v_0}{K}$ because of the equal width partition. □

Lemma 3. *If there exists some $M_1, M_2 > 0$ such that $\forall t$, $|\beta_t| \leq M_1$ and $|\alpha_t/\beta_t| \leq M_2$, choosing $\lambda_t = \arg\inf_{\lambda \in \mathbb{R}} \sum_{\tau=1}^{t-1} (\alpha_\tau + \beta_\tau \lambda)^2$ satisfies for some constant $C(M_1, M_2) > 0$*

$$\sum_{t=1}^{T} (\alpha_t + \beta_t \lambda_t)^2 \leq \inf_{\lambda} \sum_{t=1}^{T} (\alpha_t + \beta_t \lambda)^2 + C(M_1, M_2) \log T$$

Proof of Lemma 3. The proof strategy is similar to Chapter 4 of (Cesa-Bianchi and Lugosi, 2006). Define $\lambda_t^* = \arg\inf_\lambda \sum_{\tau=1}^{t} (\alpha_\tau + \beta_\tau \lambda)^2$. In words the only difference between λ_t^* and λ_t is that λ_t^* can look one step into the future. Then by Lemma 3.1 of (Cesa-Bianchi and Lugosi, 2006) we have

$$R_T := \sum_{t=1}^{T} (\alpha_t + \beta_t \lambda_t)^2 - \inf_{\lambda} \sum_{t=1}^{T} (\alpha_t + \beta_t \lambda)^2 \leq \sum_{t=1}^{T} (\alpha_t + \beta_t \lambda_t)^2 - (\alpha_t + \beta_t \lambda_t^*)^2 \tag{8}$$

We introduce simplified notation $r_t(\lambda) = \sum_{\tau=1}^{t} (\alpha_\tau + \beta_\tau \lambda)^2$. So with the new notation $\lambda_t = \inf_\lambda r_{t-1}(\lambda)$ and $\lambda_t^* = \inf_\lambda r_t(\lambda)$. We can compute

$$r'_{t-1}(\lambda_t) = 0, \qquad r''_{t-1}(\lambda_t) = 2 \sum_{\tau=1}^{t-1} \beta_\tau^2, \qquad r'''_{t-1}(\lambda) = 0 \tag{9}$$

Also denote $\delta \lambda_t = \lambda_t^* - \lambda_t$ we have

$$\begin{aligned}
\delta\lambda_t &= \arg\inf_{\delta\lambda} r_t(\lambda_t + \delta\lambda) \\
&= \arg\inf_{\delta\lambda} r_{t-1}(\lambda_t + \delta\lambda) + (\alpha_t + \beta_t\lambda_t + \beta_t\delta\lambda)^2 && \text{By definition} \\
&= \arg\inf_{\delta\lambda} r_{t-1}(\lambda_t) + r'_{t-1}(\lambda_t)\delta\lambda + \frac{1}{2} r''_{t-1}(\lambda_t)\delta\lambda^2 + \\
&\qquad (\alpha_t + \beta_t\lambda_t)^2 + 2(\alpha_t + \beta_t\lambda_t)\beta_t\delta\lambda + \beta_t^2\delta\lambda^2 && \text{Taylor expansion} \\
&= \arg\inf_{\delta\lambda} \sum_{\tau=1}^{t-1} \beta_\tau^2 \delta\lambda^2 + 2(\alpha_t + \beta_t\lambda_t)\beta_t\delta\lambda + \beta_t^2\delta\lambda^2 && \text{Apply Eq.(9) and remove irrelevant terms} \\
&= -\frac{2(\alpha_t + \beta_t\lambda_t)\beta_t}{2\sum_{\tau=1}^{t-1}\beta_\tau^2 + 2\beta_t^2} = -\frac{(\alpha_t + \beta_t\lambda_t)\beta_t}{\sum_{\tau=1}^{t}\beta_\tau^2} && \text{Minimizer of quadratic}
\end{aligned}$$

Applying the new result to Eq.(8) we have

$$\begin{aligned}
R_T &\leq \sum_{t=1}^{T} (\alpha_t + \beta_t\lambda_t)^2 - (\alpha_t + \beta_t\lambda_t^*)^2 \\
&= \sum_{t=1}^{T} (2\alpha_t + \beta_t\lambda_t + \beta_t\lambda_t^*)(\beta_t\lambda_t - \beta_t\lambda_t^*) && \text{By } (a+b)(a-b) = a^2 - b^2 \\
&= \sum_{t=1}^{T} \left(|\alpha_t + \beta_t\lambda_t| + |\alpha_t + \beta_t\lambda_t^*|\right) |\beta_t\delta\lambda_t| && \text{Cauchy schwarz} \\
&= \sum_{t=1}^{T} 2|\alpha_t + \beta_t\lambda_t||\beta_t\delta\lambda_t| && \text{By } (\alpha_t + \beta_t\lambda_t^*)^2 \leq (\alpha_t + \beta_t\lambda_t)^2 \\
&= \sum_{t=1}^{T} 2|\alpha_t + \beta_t\lambda_t| \left|\beta_t \frac{(\alpha_t + \beta_t\lambda_t)\beta_t}{\sum_{\tau=1}^{t}\beta_\tau^2}\right| && \text{Insert expression for } \delta\lambda_t
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{t=1}^{T} 2(\alpha_t + \beta_t\lambda_t)^2 \frac{\beta_t^2}{\sum_{\tau=1}^{t}\beta_\tau^2} && \text{Cauchy schwarz}\\
&= \sum_{t=1}^{T} 2(\alpha_t/\beta_t + \lambda_t)^2 \frac{\beta_t^4}{\sum_{\tau=1}^{t}\beta_\tau^2}\\
&\leq \left(\max_t 2(\alpha_t/\beta_t + \lambda_t)^2\right) \sum_{t=1}^{T} \frac{\beta_t^4}{\sum_{\tau=1}^{t}\beta_\tau^2} && \text{Holder inequality}\\
&\leq 8M_2^2 \sum_{t=1}^{T} \frac{\beta_t^4}{\sum_{\tau=1}^{t}\beta_\tau^2} && |\lambda_t| \leq M_2
\end{aligned}
$$

Finally we apply the Lemma 5 to conclude that

$$R_T \leq 8M_2^2M_1^2\log(T+1)$$

Lemma 5. *For any sequence $\beta_t, t = 1, \cdots, T$ such that $|\beta_t| \leq M, \forall t$ we have $\sum_{t=1}^{T} \frac{\beta_t^4}{\sum_{\tau=1}^{t}\beta_\tau^2} \leq M^2\log(T+1)$*

□

Finally we prove the remaining unproved Lemmas

Lemma 2. *For any choice of $r_t, s_t, \lambda_t, t = 1, \cdots, T$ we have*

$$\left(\frac{1}{T}\sum_{t=1}^{T} s_t(r_t + s_t\lambda_t)\right)^2 \leq \frac{R_T^{\text{swap}}}{T^2}\sum_{t=1}^{T} s_t^2$$

Proof of Lemma 2. Without loss of generality assume $\frac{1}{T}\sum_{t=1}^{T} s_t(r_t + s_t\lambda_t) > 0$, find some $\epsilon > 0$ such that

$$\sum_{t=1}^{T} s_t(r_t + s_t\lambda_t) = \sum_{t=1}^{T} s_t^2\epsilon$$

Such an ϵ can always be found because the range of the RHS is $[0, +\infty)$ as $\epsilon \in [0, +\infty)$ (unless all the s_t are zero, in which case the Lemma is trivially true). Therefore, there must be a solution to the equality. Because the function $\lambda_t \mapsto \lambda_t + \lambda$ is 1-Lipshitz, we have

$$
\begin{aligned}
R_T^{\text{swap}} &\geq \sum_{t=1}^{T}(r_t + s_t\lambda_t)^2 - \inf_{\lambda}\sum_{t=1}^{T}(r_t + s_t(\lambda_t + \lambda))^2 && \text{Choose a particular } \psi\\
&\geq \sum_{t=1}^{T}(r_t + s_t\lambda_t)^2 - \sum_{t=1}^{T}(r_t + s_t(\lambda_t - \epsilon))^2 && \text{Choose a particular } \lambda\\
&= \sum_{t=1}^{T}(2r_t + 2s_t\lambda_t - s_t\epsilon)s_t\epsilon = 2\left(\sum_{t=1}^{T} s_t(r_t + s_t\lambda_t)\right)\epsilon - \sum_t s_t^2\epsilon^2 = \sum_t s_t^2\epsilon^2
\end{aligned}
$$

Therefore we have

$$\left(\frac{1}{T}\sum_{t=1}^{T} s_t(r_t + s_t\lambda_t)\right)^2 = \frac{1}{T^2}\left(\sum_{t=1}^{T} s_t^2\right)^2\epsilon^2 \leq \frac{R_T^{\text{swap}}}{T^2}\sum_{t=1}^{T} s_t^2$$

□

Lemma 5. *For any sequence $\beta_t, t = 1, \cdots, T$ such that $|\beta_t| \leq M, \forall t$ we have $\sum_{t=1}^{T} \frac{\beta_t^4}{\sum_{\tau=1}^{t}\beta_\tau^2} \leq M^2\log(T+1)$*

Proof of Lemma 5. First observe that for any j if we fix the values of $\beta_t, t \neq j$, then choosing $\beta_j = M$ always maximizes $\sum_{t=1}^{T} \frac{\beta_t^4}{\sum_{\tau=1}^{t} \beta_\tau^2}$. Therefore, we have

$$\sum_{t=1}^{T} \frac{\beta_t^4}{\sum_{\tau=1}^{t} \beta_\tau^2} \leq \sum_{t=1}^{T} \frac{M^4}{\sum_{\tau=1}^{t} M^2} = M^2 \sum_{t=1}^{T} \frac{1}{t} \leq M^2 \int_{t=1}^{T+1} \frac{1}{t} = M^2 \log(T+1)$$

□

Corollary 1. *Under the assumptions of Theorem 1 if additionally $b_t \geq 0, \forall t$, there exists an algorithm to output μ_t in Protocol 2 with $c_t \equiv 0$ that is asymptotically exact for μ_t^*, b_t generated by any strategy of nature and agent.*

Proof of Corollary 1. We make a small modification in Algorithm 1. Originally line 5 of Algorithm 1 outputs $\mu_t = \hat{\mu}_t$ and $c_t = \hat{c}_t + \lambda_t$; instead we output $\mu_t' = \hat{\mu}_t - (\hat{c}_t + \lambda_t)$ and $c_t' = 0$.

This modified algorithm can achieve asymptotic exactness because

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} b_t(\mu_t' - y_t) - |b_t|c_t' &= \frac{1}{T}\sum_{t=1}^{T} b_t(\mu_t' - y_t) && c_t' \text{ is zero} \\
&= \frac{1}{T}\sum_{t=1}^{T} b_t(\mu_t - c_t - y_t) && \text{Definition of } \mu_t' \\
&= \frac{1}{T}\sum_{t=1}^{T} b_t(\mu_t - y_t) - b_t c_t \\
&= \frac{1}{T}\sum_{t=1}^{T} b_t(\mu_t - y_t) - |b_t|c_t && b_t \geq 0
\end{aligned}$$

The final expression goes to 0 by Theorem 1. □

D Additional Proofs

Proposition 1. *For any $\mu_t, c_t, \mu_t^* \in (0,1)$ where $(\mu_t - c_t, \mu_t + c_t) \subset (0,1)$*

1. If $\mu_t^* \in [\mu_t - c_t, \mu_t + c_t]$ then $\forall l_t : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ we have $L_t^* \in [L_t^{\min}, L_t^{\max}]$

2. If $\mu_t^* \notin [\mu_t - c_t, \mu_t + c_t]$ then $\forall l_t : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, if $\forall a \in \mathcal{A}, \ell_t(a,0) \neq \ell_t(a,1)$, then $L_t^* \notin [L_t^{\min}, L_t^{\max}]$

Proof of Proposition 1. Part I: without loss of generality assume $l_t(a_t, 1) > l_t(a_t, 0)$, denote $L_t = \mathbb{E}_\mu[l_t(a_t, Y)]$ and we also use the notation shorthand $l_t(y)$ to denote $l_t(a_t, y)$. Since $\mu^* \in [\mu_t - c_t, \mu_t + c_t]$ we have

$$\begin{aligned}
|L_t - L_t^*| &\leq \sup_{\mu^* \in \mu_t \pm c_t} |\mathbb{E}_{Y\sim\mu_t}[l_t(Y)] - \mathbb{E}_{Y\sim\mu^*}[l_t(Y)]| \\
&= \sup_{\mu^* \in \mu_t \pm c_t} |\mu_t l_t(1) + (1-\mu_t) l_t(0) - \mu_t^* l_t(1) - (1-\mu_t^*) l_t(0)| \\
&= \sup_{\mu^* \in \mu_t \pm c_t} |(\mu_t - \mu_t^*)(l_t(1) - l_t(0))| \\
&\leq c_t(l_t(1) - l_t(0))
\end{aligned}$$

by similar algebra as above we also have

$$\begin{aligned}
L_t - L_t^{\min} &= c_t(l_t(1) - l_t(0)) \\
L_t^{\max} - L_t &= c_t(l_t(1) - l_t(0))
\end{aligned}$$

therefore it must be that $L_t^* \geq L_t^{\min}$ and $L_t^* \leq L_t^{\max}$.

Part II: Choose $\ell_t(a_t, y) = \alpha y + \beta$ where $\alpha \neq 0$; by choosing α, β this can represent any loss function ℓ where $\ell(a_t, 0) \neq \ell(a_t, 1)$. We prove the case where $\alpha > 0$ and the case where $\alpha < 0$ can be similarly proven. Suppose $\mu_t^* < \mu_t - c_t$

$$L_t^* = \mathbb{E}_{Y\sim\mu^*}[\alpha Y + \beta] = \alpha\mu_t^* + \beta < \alpha(\mu_t - c_t) + \beta$$
$$L_t^{\min} = \min_{\tilde{\mu}\in\mu_t\pm c_t} \mathbb{E}_{Y\sim\tilde{\mu}}[\alpha Y + \beta] = \mathbb{E}_{Y\sim\mu_t - c_t}[\alpha Y + \beta] = \alpha(\mu_t - c_t) + \beta$$

but this would imply that $L_t^* < L_t^{\min}$.

Suppose $\mu_t^* > \mu_t + c_t$

$$L_t^* = \mathbb{E}_{\mu^*}[\alpha Y + \beta] = \alpha\mu_t^* + \beta > \alpha(\mu_t + c_t) + \beta$$
$$L_t^{\max} = \max_{\tilde{\mu}\in\mu_t\pm c_t} \mathbb{E}_{Y\sim\tilde{\mu}}[\alpha Y + \beta] = \mathbb{E}_{Y\sim\mu_t + c_t}[\alpha Y + \beta] = \alpha(\mu_t + c_t) + \beta$$

but this would imply that $L_t^* > L_t^{\max}$. □

Lemma 1. *Let $\mu, c \in (0,1)$ such that $[\mu - c, \mu + c] \subset [0,1]$, then a function $f : \mathcal{Y} \to \mathbb{R}$ satisfies $\forall\tilde{\mu} \in [\mu - c, \mu + c]$, $\mathbb{E}_{Y\sim\tilde{\mu}}[f(Y)] \leq 0$ if and only if for some $b \in \mathbb{R}$ and $\forall y \in \{0,1\}$, $f(y) \leq b(y - \mu) - |b|c$.*

Proof of Lemma 1. If: if for some $b \in \mathbb{R}$ we have $f(y) \leq b(y - \mu) - |b|c$ then for any $\tilde{\mu}$ such that $\tilde{\mu} \in [\mu - c, \mu + c]$ or equivalently $|\tilde{\mu} - \mu| \leq c$ we have

$$\mathbb{E}_{Y\sim\tilde{\mu}}[f(Y)] \leq \mathbb{E}_{Y\sim\tilde{\mu}}[b(Y - \mu) - |b|c] = b(\tilde{\mu} - \mu) - |b|c \leq |b||\tilde{\mu} - \mu| - |b|c \leq 0$$

Only if: If $\mu = 1$ or $\mu = 0$ then the proof is trivial; we consider the case where $\mu \in (0,1)$. Suppose for any $\tilde{\mu} \in [\mu - c, \mu + c]$ we have $\mathbb{E}_{Y\sim\tilde{\mu}}[f(Y)] \leq 0$ we have (by instantiating a few concrete values for $\tilde{\mu}$)

$$f(1)(\mu - c) + f(0)(1 - \mu + c) \leq 0 \tag{10}$$
$$f(1)(\mu + c) + f(0)(1 - \mu - c) \leq 0 \tag{11}$$

Choose some b such that $f(1) = b(1 - \mu) - |b|c$. Such a b must exist because the range of $b \mapsto b(1 - \mu) - |b|c$ is $\mathbb{R}$. If $b < 0$ then by Eq.(10) we have

$$b(1 - \mu + c)(\mu - c) + f(0)(1 - \mu + c) \leq 0, \qquad f(0) \leq -b(\mu - c) = b(0 - \mu) - |b|c$$

Conversely if $b \geq 0$ by Eq.(11) we have

$$b(1 - \mu - c)(\mu - c) + f(0)(1 - \mu + c) \leq 0, \qquad f(0) \leq -b(\mu + c) = b(0 - \mu) - |b|c$$

In either cases this is equivalent to $\forall y \in \{0,1\}$, $f(y) \leq b(y - \mu) - |b|c$. □

Proposition 2. *If the stake $b_t = l_t(a_t, 1) - l_t(a_t, 0)$ then $L_t^{\text{pay}} \in [L_t^{\min}, L_t^{\max}]$*

Proof of Proposition 2. For convenience denote $l(Y) := l_t(Y, a_t)$. Without loss of generality assume $l(1) > l(0)$

$$\begin{aligned} L_t^{\min} &= \min_{\tilde{\mu}\in\mu_t\pm c_t} \mathbb{E}_{\tilde{\mu}}[l(Y)] = (\mu_t - c_t)l(1) + (1 - \mu_t + c_t)l(0) = \mu_t l(1) + (1 - \mu_t)l(0) - (l(1) - l(0))c_t \\ &= \mu_t^* l(1) + (\mu_t - \mu_t^*)l(1) + (1 - \mu_t^*)l(0) - (\mu_t - \mu_t^*)l(0) - (l(1) - l(0))c_t \\ &= \mathbb{E}_{\mu_t^*}[l(Y)] - (l(1) - l(0))\mathbb{E}_{\mu_t^*}[Y - \mu] - (l(1) - l(0))c_t] \leq L_t^{\text{pay}} \end{aligned}$$

and

$$\begin{aligned} L_t^{\max} &= \min_{\tilde{\mu}\in\mu_t\pm c_t} \mathbb{E}_{\tilde{\mu}}[l(Y)] = (\mu_t + c_t)l(1) + (1 - \mu_t - c_t)l(0) = \mu_t l(1) + (1 - \mu_t)l(0) + (l(1) - l(0))c_t \\ &= \mu_t^* l(1) + (\mu_t - \mu_t^*)l(1) + (1 - \mu_t^*)l(0) - (\mu_t - \mu_t^*)l(0) + (l(1) - l(0))c_t \\ &= \mathbb{E}_{\mu_t^*}[l(Y)] - (l(1) - l(0))\mathbb{E}_{\mu_t^*}[Y - \mu] + (l(1) - l(0))c_t = L_t^{\text{pay}} \end{aligned}$$

□

Proposition 4. *The forecaster function $\mu : \mathcal{X} \to [0,1], c : x \mapsto c_0$ is sound with respect to $\mathcal{B} = \{x \mapsto \tilde{b}(\mu(x)), \tilde{b} : \mathbb{R} \to \mathbb{R}\}$ if and only if the MCE error of μ is less than c_0.*

Proof. If the MCE error of μ is less than c_0, denote $U = \mu(X)$ by definition we have, for every $U \in [0,1]$

$$|U - \mathbb{E}[Y \mid U]| \leq c_0 \tag{12}$$

For any $b \in \mathcal{B}$, denote $b(X) := \tilde{b}(\mu(X)) = \tilde{b}(U)$ we have

$$\begin{aligned}
\mathbb{E}[b(X)(\mu(X) - Y) - |b(X)|c(X)] &= \mathbb{E}\left[\mathbb{E}[\tilde{b}(U)(\mu(X) - Y) - |\tilde{b}(U)|c_0 \mid U]\right] && \text{Iterated Expectation} \\
&= \mathbb{E}[\tilde{b}(U)\mathbb{E}[\mu(X) - Y \mid U] - |\tilde{b}(U)|c_0] && \mathbb{E}[UZ \mid U] = U\mathbb{E}[Z \mid U] \\
&= \mathbb{E}[\tilde{b}(U)(U - \mathbb{E}[Y \mid U]) - |\tilde{b}(U)|c_0] && \text{Linearity} \\
&\leq \mathbb{E}[|\tilde{b}(U)||U - \mathbb{E}[Y \mid U]| - |\tilde{b}(U)|c_0] && \text{Cauchy Schwarz} \\
&= \mathbb{E}[|\tilde{b}(U)|\,(|U - \mathbb{E}[Y \mid U]| - c_0)] \leq 0 && \text{By Eq.(12)}
\end{aligned}$$

which shows that μ, c is sound.

Conversely suppose there is some interval (u_0, u_1) such that whenever $U \in (u_0, u_1)$

$$U - \mathbb{E}[Y \mid U] > c_0$$

we can choose $b(X) := \tilde{b}(U) = \mathbb{I}(U \in [u_0, u_1])$ we have

$$\mathbb{E}[b(X)(\mu(X) - Y) - |b(X)|c(X)] = \mathbb{E}[|\tilde{b}(U)|\,(|U - \mathbb{E}[Y \mid U]| - c_0)] > 0$$

so the forecaster is not sound. We can show a similar proof when

$$U - \mathbb{E}[Y \mid U] < -c_0$$

□

Proposition 5. *Let $\mathcal{S} \subset 2^{\mathcal{X}}$. If a forecaster function $\mu : \mathcal{X} \to [0,1], c : x \mapsto c_0$ is sound with respect to $\mathcal{B} = \{x \mapsto \tilde{b}(\mu(x))\mathbb{I}(x \in S), S \in \mathcal{S}, \tilde{b} : \mathbb{R} \to \mathbb{R}\}$, then it is $(\mathcal{S}, c_0)$-multicalibrated.*

Proof. Denote $U = \mu(X)$. Suppose μ, c is not multi-calibrated, then there exists $S \in \mathcal{S}$ and there exists some interval (u_0, u_1) such that whenever $U \in (u_0, u_1)$

$$|\mathbb{I}(X \in S)(U - \mathbb{E}[Y \mid U])| > c_0$$

Suppose $\mathbb{I}(X \in S)(U - \mathbb{E}[Y \mid U]) > c_0$ we can choose $b(X) := \mathbb{I}(U \in [u_0, u_1] \cap X \in S)$ we have

$$\mathbb{E}[b(X)(\mu(X) - Y) - |b(X)|c(X)] = \mathbb{E}[|b(X)|\,(|U - \mathbb{E}[Y \mid U]| - c_0)] > 0$$

We can show a similar proof when $\mathbb{I}(X \in S)(U - \mathbb{E}[Y \mid U]) < -c_0$. □