Minimax strategy for prediction with expert advice under stochastic assumptions

Wojciech Kotłowski
Poznań University of Technology, Poland
wkotlowski@cs.put.poznan.pl

Abstract

We consider the setting of prediction with expert advice with an additional assumption that each expert generates its losses i.i.d. according to some distribution. We first identify a class of "admissable" strategies, which we call permutation invariant, and show that every strategy outside this class will perform not better than some permutation invariant strategy. We then show that when the losses are binary, a simple Follow the Leader (FL) algorithm is the minimax strategy for this game, where minimaxity is simultaneously achieved for the expected regret, the expected redundancy, and the excess risk. Furthermore, FL has also the smallest regret, redundancy, and excess risk over all permutation invariant prediction strategies, simultaneously for all distributions over binary losses. When the losses are continuous in $[0, 1]$, FL remains minimax only when an additional trick called "loss binarization" is applied.

1 Introduction

In the game of prediction with expert advice [1, 2], the algorithm sequentially decides on one of K experts to follow, and suffers loss associated with the chosen expert. The difference between the algorithm's cumulative loss and the cumulative loss of the best expert is called *regret*. The goal is to minimize the regret in the worst case over all possible loss sequences. An algorithm which achieves this goal (i.e., minimizes the worst-case regret) is called *minimax*. While there is no known solution to this problem in the general setting, it is possible to derive minimax algorithms for some special variants of this game [1, 2, 3, 4]. Interestingly, all these algorithms share a similar strategy of playing against a maximin adversary which assigns losses uniformly at random. They often have the *equlization* property: all data sequences lead to the same value of the regret. While this property makes them robust against the worst-case sequence, it also makes them over-conservative, preventing them from exploiting the case, when the actual data are "easy". There are various algorithm which combine almost optimal worst-case performance with good performance on easy sequences [5, 6, 7, 8, 9]; these algorithms, however, are not motivated from the minimax principle.

In this paper, we drop the analysis of worst-case performance entirely, and explore the minimax principle in a more constrained setting, in which the adversary is assumed to be *stochastic*. In particular, we associate with each expert k a fixed distribution P_k over loss values, and assume the observed losses of expert k are generated independently from P_k. We believe this setting might be practically useful, and that it is interesting to determine the minimax algorithm under this assumption. We immediately face two difficulties here. First, due to stochastic nature of the adversary, it is no longer possible to follow standard approaches of minimax analysis, such as backward induction[1, 2] or sequential minimax duality [10, 3], and we need to resort to a different technique. We define the notion of *permutation invariance* of prediction strategies. This let us identify a class of "admissable" strategies (which we call permutation invariant), and show that every strategy outside this class will perform not better than some permutation invariant strategy. Secondly, while the regret is a single, commonly used performance metric in the worst-case setting, the situation is different in the stochastic case. We know at least three potentially useful metrics in the stochastic setting: the *expected regret*, the *expected redundancy*, and the *excess risk* [11], and it is not clear, which of them should be used to define the minimax strategy.

Fortunately, it turns out that there exists a single strategy which is minimax with respect to all three metrics simultaneously. In the case of *binary* losses, which take out values from $\{0,1\}$, this strategy turns out to be the *Follow the Leader* (FL) algorithm, which chooses an expert with the smallest cumulative loss at a given trial (with ties broken randomly). Interestingly, FL is known to perform poorly in the worst-case, as its worst-case regret will grow linearly with T [2]. On the contrary, in the stochastic setting with binary losses, FL has also the smallest regret, redundancy, and excess risk over all permutation invariant prediction strategies, *simultaneously for all distributions over binary losses!* In a more general case of continuous losses in the range $[0,1]$, FL is provably suboptimal. However, by applying *binarization trick* to the losses [6], i.e. randomly setting them to $\{0,1\}$ such that the expectation matches the actual loss, and using FL on the binarized sequence, we obtain the minimax strategy in the continuous case.

2 Problem Setting

In the game of prediction with expert advice, at each trial $t = 1,\ldots,T$, the algorithm predicts with a distribution $\boldsymbol{w}_t \in \Delta^K$ over K experts. Then the loss vector $\boldsymbol{\ell}_t \in \mathcal{X}^K$ is revealed (where $\mathcal{X}$ is either $\{0,1\}$ or $[0,1]$), and the algorithm suffers loss $\boldsymbol{w}_t \cdot \boldsymbol{\ell}_t$. The sequence of outcomes $\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_t$ is abbreviated as $\boldsymbol{\ell}^t$. We let $L_{t,k}$ denote the cumulative loss of expert k at iteration t, $L_{t,k} = \sum_{q \le t} \ell_{t,k}$. We assume there are K distributions $\mathcal{P} = (P_1,\ldots,P_K)$ over $\mathcal{X}$, such that for each k, the losses $\ell_{t,k}$, $t = 1,\ldots,T$, are generated i.i.d. from P_k. Note that this implies that $\ell_{t,k}$ is independent from $\ell_{t',k'}$ whenever $t' \neq t$ or $k \neq k'$. We formally define the prediction strategy as a sequence of T functions $\boldsymbol{\omega} = (\boldsymbol{w}_1,\ldots,\boldsymbol{w}_T)$, such that $\boldsymbol{w}_t \colon \mathcal{X}^{t-1} \to \Delta^K$. The performance of the strategy $\boldsymbol{\omega}$ on the set of distributions $\mathcal{P}$ can be measured by one of the three following metrics:

$$\begin{aligned}
&\text{the } \textit{expected regret}: && R_{\mathrm{eg}}(\boldsymbol{\omega},\mathcal{P}) \;=\; \mathbb{E}\Bigg[\sum_{t=1}^T \boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t - \min_k L_{T,k}\Bigg], \\
&\text{the } \textit{expected redundancy}: && R_{\mathrm{ed}}(\boldsymbol{\omega},\mathcal{P}) \;=\; \mathbb{E}\Bigg[\sum_{t=1}^T \boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\Bigg] - \min_k \mathbb{E}\left[L_{T,k}\right], \\
&\text{the } \textit{excess risk}: && R_{\mathrm{isk}}(\boldsymbol{\omega},\mathcal{P}) \;=\; \mathbb{E}\Big[\boldsymbol{w}_T(\boldsymbol{\ell}^{T-1}) \cdot \boldsymbol{\ell}_T\Big] - \min_k \mathbb{E}\left[\ell_{T,k}\right],
\end{aligned}$$

where in each case, the expectation is over the sequence $\boldsymbol{\ell}^T$ with respect to distributions $P_1,\ldots,P_K$. Given performance measure R, we say that a strategy $\boldsymbol{\omega}^*$ is *minimax* with respect to R, if:

$$\sup_{\mathcal{P}} R(\boldsymbol{\omega}^*,\mathcal{P}) = \inf_{\boldsymbol{\omega}} \sup_{\mathcal{P}} R(\boldsymbol{\omega},\mathcal{P}),$$

where the supremum is over all K-sets of distributions $(P_1,\ldots,P_K)$ on $\mathcal{X}$, and the infimum is over all prediction strategies.

We say that a strategy $\boldsymbol{\omega}$ is *permutation invariant* if for any $t = 1,\ldots,T$, and any permutation $\sigma \in S_K$, where S_K denotes the group of permutations over $\{1,\ldots,K\}$, $\boldsymbol{w}_t(\sigma(\boldsymbol{\ell}^{t-1})) = \sigma(\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}))$, where for any vector $\boldsymbol{v} = (v_1,\ldots,v_K)$, we defined $\sigma(\boldsymbol{v}) = (v_{\sigma(1)},\ldots,v_{\sigma(K)})$ and abbreviated $\sigma(\boldsymbol{\ell}^{t-1}) = \sigma(\boldsymbol{\ell}_1),\ldots,\sigma(\boldsymbol{\ell}_{t-1})$. In words, if we σ-permute the indices of all past loss vectors, the resulting weight vector will be the σ-permutation of the original weight vector. Permutation invariant strategies are natural, as they only rely on the observed outcomes, not on the expert indices.

Lemma 1. *Let $\boldsymbol{\omega}$ be permutation invariant. Then, for any permutation $\sigma \in S_K$, $\mathbb{E}_{\sigma(\mathcal{P})}\left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right] = \mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right]$, and moreover $R(\boldsymbol{\omega},\sigma(\mathcal{P})) = R(\boldsymbol{\omega},\mathcal{P})$, where R is the expected regret, expected redundancy, or excess risk, and $\sigma(\mathcal{P}) = (P_{\sigma(1)},\ldots,P_{\sigma(K)})$.*

We now show that permutation invariant strategies are "admissable" in the following sense:

Theorem 2. *For any strategy $\boldsymbol{\omega}$, there exists permutation invariant strategy $\widetilde{\boldsymbol{\omega}}$, such that $\sup_{\mathcal{P}} R(\widetilde{\boldsymbol{\omega}},\mathcal{P}) \le \sup_{\mathcal{P}} R(\boldsymbol{\omega},\mathcal{P})$, where R is either the expected regret, the expected redundancy or the excess risk.*

Theorem 2 statest that it suffices to search for minimax strategy only within the set of permutation invariant strategies.

Given loss sequence $\boldsymbol{\ell}^{t-1}$, let $N = |\operatorname{argmin}_{j=1,\ldots,K} L_{t-1,j}|$ be the size of the leader set at the beginning of trial t. We define *Follow the Leader* (FL) strategy $\boldsymbol{w}_t^{\text{fl}}$ such that $w_{t,k}^{\text{fl}} = \frac{1}{N}$ if $k \in \operatorname{argmin}_j L_{t-1,j}$ and $w_{t,k}^{\text{fl}} = 0$ otherwise. In other words, FL predicts with the current leader, breaking ties uniformly at random. It is easy to show that FL strategy is permutation invariant.

3 Binary losses

In this section, we set $\mathcal{X} = \{0,1\}$, so that all losses are binary. In this case, each P_k is a Bernoulli distribution. Take any permutation invariant strategy $\boldsymbol{\omega}$. Using Lemma 1 $K!$ on all $\sigma \in S_K$:

$$\mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right] = \underbrace{\frac{1}{K!}\sum_{\sigma} \mathbb{E}_{\sigma(\mathcal{P})}\left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right]}_{=:\, \overline{\text{loss}}_t(\boldsymbol{w}_t, \mathcal{P})}. \tag{1}$$

We now show the main result of this paper, a surprisingly strong property of FL strategy, which states that FL minimizes $\overline{\text{loss}}_t(\boldsymbol{w}_t, \mathcal{P})$ over all K-sets of distributions *simultaneously*. Hence, FL is not only optimal in the worst case, but is actually optimal for permutation-averaged expected loss for any $\mathcal{P}$, which implies by (1) that *FL has the smallest expected loss among all permutation invariant strategies for any $\mathcal{P}$.*

Theorem 3. *Let $\boldsymbol{\omega}^{\text{fl}} = (\boldsymbol{w}_1^{\text{fl}}, \ldots, \boldsymbol{w}_T^{\text{fl}})$ be the FL strategy. Then, for any K-set of distributions $\mathcal{P} = (P_1, \ldots, P_K)$ over binary losses, for any strategy $\boldsymbol{\omega} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T)$, and any $t = 1, \ldots, T$:*

$$\overline{\text{loss}}_t(\boldsymbol{w}_t^{\text{fl}}, \mathcal{P}) \leq \overline{\text{loss}}_t(\boldsymbol{w}_t, \mathcal{P}).$$

The consequence of Theorem 3 is the following corollary which states the minimaxity of FL strategy for binary losses:

Corollary 4. *Let $\boldsymbol{\omega}^{\text{fl}} = (\boldsymbol{w}_1^{\text{fl}}, \ldots, \boldsymbol{w}_T^{\text{fl}})$ be the FL strategy. Then, for any $\mathcal{P}$ over binary losses, and any permutation invariant strategy $\boldsymbol{\omega}$:*

$$R(\boldsymbol{\omega}^{\text{fl}}, \mathcal{P}) \leq R(\boldsymbol{\omega}, \mathcal{P}).$$

where R is the expected regret, expected redundancy, or excess risk. This implies:

$$\sup_{\mathcal{P}} R(\boldsymbol{\omega}^{\text{fl}}, \mathcal{P}) = \inf_{\boldsymbol{\omega}} \sup_{\mathcal{P}} R(\boldsymbol{\omega}, \mathcal{P}),$$

where the supremum is over all distributions on binary losses, and the infimum over all (not necessarily permutation invariant) strategies.

Proof. The second statement immediately follows from the first statement and Theorem 2. For the first statement, note that the "loss of the best expert" part of each measure only depends on $\mathcal{P}$. Hence, we only need to show that for any $t = 1, \ldots, T$,

$$\mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t^{\text{fl}} \cdot \boldsymbol{\ell}_t\right] \leq \mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t \cdot \boldsymbol{\ell}_t\right].$$

Since $\boldsymbol{w}_t^{\text{fl}}$ and $\boldsymbol{w}_t$ are permutation invariant, Lemma 1 shows that $\mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t^{\text{fl}} \cdot \boldsymbol{\ell}_t\right] = \overline{\text{loss}}_t(\boldsymbol{w}_t^{\text{fl}}, \mathcal{P})$, and similarly, $\mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t \cdot \boldsymbol{\ell}_t\right] = \overline{\text{loss}}_t(\boldsymbol{w}_t, \mathcal{P})$. Application of Theorem 3 finishes the proof. □

4 Continuous losses

In this section, we consider the general case $\mathcal{X} = [0,1]$ of continuous loss vectors. We give a modification of FL and prove its minimaxity. In the appendix, we justify the modification by arguing that the vanilla FL strategy is not minimax for continuous losses.

The modification of FL is based on the procedure we call *binarization*. A similar trick has already been used in [6] to deal with non-integer losses in a different context. We define a binarization of any loss value $\ell_{t,k} \in [0,1]$ as a Bernoulli random variable $b_{t,k}$ which takes out value 1 with probability $\ell_{t,k}$ and value 0 with probability $1 - \ell_{t,k}$. In other words, we replace each non-binary loss $\ell_{t,k}$ by a random binary outcome $b_{t,k}$, such that $\mathbb{E}[b_{t,k}] = \ell_{t,k}$. Note that if $\ell_{t,k} \in \{0,1\}$, then $b_{t,k} = \ell_{t,k}$, i.e. binarization has no effect on losses which are already binary. Let us also define $\boldsymbol{b}_t = (b_{t,1}, \ldots, b_{t,K})$, where all K Bernoulli random variables $b_{t,k}$ are independent. Similarly, $\boldsymbol{b}^t$ will denote a binary loss sequence $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_t$, where the binarization procedure was applied independently (with a new set of Bernoulli variables) for each trial t. Now, given the loss sequence $\boldsymbol{\ell}^{t-1}$, we define the *binarized FL* strategy $\boldsymbol{\omega}^{\mathrm{bfl}}$ by:

$$\boldsymbol{w}_t^{\mathrm{bfl}}(\boldsymbol{\ell}^{t-1}) = \mathbb{E}_{\boldsymbol{b}^{t-1}}\left[\boldsymbol{w}_t^{\mathrm{fl}}(\boldsymbol{b}^{t-1})\right],$$

where $\boldsymbol{w}_t^{\mathrm{fl}}(\boldsymbol{b}^{t-1})$ is the standard FL strategy applied to binarized losses $\boldsymbol{b}^{t-1}$, and the expectation is over internal randomization of the algorithm (binarization variables).

Note that if the set of distributions $\mathcal{P}$ has support only on $\{0,1\}$, then $\boldsymbol{w}_t^{\mathrm{bfl}} \equiv \boldsymbol{w}_t^{\mathrm{fl}}$. On the other hand, these two strategies may differ significantly for non-binary losses. However, we will show that for any K-set of distributions $\mathcal{P}$ (with support in $[0,1]$), $\boldsymbol{w}_t^{\mathrm{bfl}}$ will behave in the same way as $\boldsymbol{w}_t^{\mathrm{fl}}$ would behave on some particular K-set of distributions over binary losses. To this end, we introduce *binarization of a K-set of distributions $\mathcal{P}$*, defined as $\mathcal{P}^{\mathrm{bin}} = (P_1^{\mathrm{bin}}, \ldots, P_K^{\mathrm{bin}})$, where P_k^{bin} is a distribution with support $\{0,1\}$ such that:

$$\mathbb{E}_{P_k^{\mathrm{bin}}}[\ell_{t,k}] = P_k^{\mathrm{bin}}(\ell_{t,k} = 1) = \mathbb{E}_{P_k}[\ell_{t,k}].$$

In other words, P_k^{bin} is a Bernoulli distribution which has the same expectation as the original distribution (over continuous losses) P_k. We now show the following results:

Lemma 5. *For any K-set of distributions $\mathcal{P} = (P_1, \ldots, P_K)$ with support on $\mathcal{X} = [0,1]$,*

$$\mathbb{E}_{\boldsymbol{\ell}^t \sim \mathcal{P}^t}\left[\boldsymbol{w}_t^{\mathrm{bfl}}(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right] = \mathbb{E}_{\boldsymbol{\ell}^t \sim (\mathcal{P}^{\mathrm{bin}})^t}\left[\boldsymbol{w}_t^{\mathrm{fl}}(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right].$$

Lemma 6. *For any K-set of distributions $\mathcal{P} = (P_1, \ldots, P_K)$ with support on $\mathcal{X} = [0,1]$,*

$$R(\boldsymbol{\omega}^{\mathrm{bfl}}, \mathcal{P}) \leq R(\boldsymbol{\omega}^{\mathrm{fl}}, \mathcal{P}^{\mathrm{bin}}),$$

where R is either the expected regret, the expected redundancy, or the excess risk.

Theorem 7. *Let $\boldsymbol{\omega}^{\mathrm{bfl}} = (\boldsymbol{w}_1^{\mathrm{bfl}}, \ldots, \boldsymbol{w}_T^{\mathrm{bfl}})$ be the binarized FL strategy. Then:*

$$\sup_{\mathcal{P}} R(\boldsymbol{\omega}^{\mathrm{bfl}}, \mathcal{P}) = \inf_{\boldsymbol{\omega}} \sup_{\mathcal{P}} R(\boldsymbol{\omega}, \mathcal{P}),$$

where R is the expected regret, expected redundancy, or excess risk, the supremum is over all K-sets of distributions on $[0,1]$, and the infimum is over all prediction strategies.

In the appendix, we argue that vanilla FL is not the minimax strategy for continuous losses, so that the binarization procedure is justified.

5 Open Problem

The setting considered in this paper is quite limited even in the stochastic case, as it does not consider distributions over loss vectors which are i.i.d. between trials, but not necessarily i.i.d. between experts. It would be interesting to determined the minimax strategy in this more general setting. Preliminary computational experiment suggest that FL is not minimax even for binary losses, when dependencies between experts are allowed.

Acknowledgments. The author was supported by the Polish National Science Centre under grant no. 2013/11/D/ST6/03050.

References

[1] Nicolò Cesa-Bianchi, Yaov Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

[2] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[3] Wouter M. Koolen. *Combining Strategies Efficiently: High-quality Decisions from Conflicting Advice*. PhD thesis, ILLC, University of Amsterdam, 2011.

[4] Jacob Abernethy, Manfred K. Warmuth, and Joel Yellin. When random play is optimal against an adversary. In *COLT*, pages 437–445, July 2008.

[5] Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(1):1281–1316, 2014.

[6] Tim van Erven, Wojciech Kotłowski, and Manfred K. Warmuth. Follow the leader with dropout perturbations. In *COLT*, pages 949–974, 2014.

[7] Amir Sani, Gergely Neu, and Alessandro Lazaric. Exploiting easy data in online optimization. In *NIPS*, pages 810–818. 2014.

[8] Wouter M. Koolen and Tim van Erven. Second-order quantile methods for experts and combinatorial games. In *COLT*, pages 1155–1175, 2015.

[9] Haipeng Luo and Robert E. Schapire. Achieving all with no parameters: AdaNormalHedge. In *COLT*, pages 1286–1304, 2015.

[10] Jacob Abernethy, Alekh Agarwal, Peter L. Bartlett, and Alexander Rakhlin. A stochastic view of optimal regret through minimax duality. In *COLT*, 2009.

[11] Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.

Proof of Lemma 1

We first show that the expected loss of the algorithm at any iteration $t = 1, \ldots, T$, is the same for both $\sigma(\mathcal{P})$ and $\mathcal{P}$:

$$\begin{aligned} \mathbb{E}_{\sigma(\mathcal{P})}\left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right] &= \mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t(\sigma(\boldsymbol{\ell}^{t-1})) \cdot \sigma(\boldsymbol{\ell}_t)\right] = \mathbb{E}_{\mathcal{P}}\left[\sigma(\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1})) \cdot \sigma(\boldsymbol{\ell}_t)\right] \\ &= \mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right], \end{aligned}$$

where the first equality is due to the fact, that permuting the distributions is equivalent to permuting the coordinates of the losses (which are random variables with respect to these distributions), the second equality exploits the permutation invariance of $\boldsymbol{\omega}$, while the third inequality uses a simple fact that the dot product is invariant under permuting both arguments. Therefore, the "loss of the algorithm" part of any of the three measures (regret, redundancy, risk) remains the same. To show that the "loss of the optimal" part of each measure is the same, note that for any $t = 1, \ldots, T$, $k = 1, \ldots, K$, $\mathbb{E}_{\sigma(\mathcal{P})}\left[\ell_{t,k}\right] = \mathbb{E}_{\mathcal{P}}\left[\ell_{t,\sigma(k)}\right]$, which implies:

$$\begin{aligned} \min_k \mathbb{E}_{\sigma(\mathcal{P})}\left[\ell_{T,k}\right] &= \min_k \mathbb{E}_{\mathcal{P}}\left[\ell_{T,\sigma(k)}\right] = \min_k \mathbb{E}_{\mathcal{P}}\left[\ell_{T,k}\right], \\ \min_k \mathbb{E}_{\sigma(\mathcal{P})}\left[L_{T,k}\right] &= \min_k \mathbb{E}_{\mathcal{P}}\left[L_{T,\sigma(k)}\right] = \min_k \mathbb{E}_{\mathcal{P}}\left[L_{T,k}\right], \\ \mathbb{E}_{\sigma(\mathcal{P})}\left[\min_k L_{T,k}\right] &= \mathbb{E}_{\mathcal{P}}\left[\min_k L_{T,\sigma(k)}\right] = \mathbb{E}_{\mathcal{P}}\left[\min_k L_{T,k}\right], \end{aligned}$$

so that the "loss of the best expert" parts of all measures are also the same for both $\sigma(\mathcal{P})$ and $\mathcal{P}$.

Proof of Theorem 2

Define $\widetilde{\boldsymbol{\omega}} = (\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_T)$ as:

$$\widetilde{\boldsymbol{w}}_t(\boldsymbol{\ell}^{t-1}) = \frac{1}{K!} \sum_{\tau \in S_K} \tau^{-1}\Big(\boldsymbol{w}_t(\tau(\boldsymbol{\ell}^{t-1}))\Big).$$

Note that $\widetilde{\boldsymbol{\omega}}$ is a valid prediction strategy, since $\widetilde{\boldsymbol{w}}_t$ is a function of $\boldsymbol{\ell}^{t-1}$, and $\widetilde{\boldsymbol{w}}_t \in \Delta^K$ ($\widetilde{\boldsymbol{w}}_t$ is a convex combination of $K!$ distributions, so it is a distribution itself). Moreover, $\widetilde{\boldsymbol{\omega}}$ is permutation invariant:

$$\begin{aligned} \widetilde{\boldsymbol{w}}_t(\sigma(\boldsymbol{\ell}^{t-1})) &= \frac{1}{K!} \sum_{\tau \in S_K} \tau^{-1}\Big(\boldsymbol{w}_t(\tau\sigma(\boldsymbol{\ell}^{t-1}))\Big) = \frac{1}{K!} \sum_{\tau \in S_K} (\tau\sigma^{-1})^{-1}\Big(\boldsymbol{w}_t(\tau(\boldsymbol{\ell}^{t-1}))\Big) \\ &= \frac{1}{K!} \sum_{\tau \in S_K} \sigma\tau^{-1}\Big(\boldsymbol{w}_t(\tau(\boldsymbol{\ell}^{t-1}))\Big) = \sigma(\widetilde{\boldsymbol{w}}_t(\boldsymbol{\ell}^{t-1})), \end{aligned}$$

where the second inequality is from replacing the summation index $\tau \mapsto \tau\sigma$. Now, note that the expected loss of $\widetilde{\boldsymbol{w}}_t$ is:

$$\begin{aligned} \mathbb{E}_{\mathcal{P}}\left[\widetilde{\boldsymbol{w}}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right] &= \frac{1}{K!} \sum_{\tau \in S_K} \mathbb{E}_{\mathcal{P}}\left[\tau^{-1}\left(\boldsymbol{w}_t(\tau(\boldsymbol{\ell}^{t-1}))\right) \cdot \boldsymbol{\ell}_t\right] \\ &= \frac{1}{K!} \sum_{\tau \in S_K} \mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t(\tau(\boldsymbol{\ell}^{t-1})) \cdot \tau(\boldsymbol{\ell}_t)\right] \\ &= \frac{1}{K!} \sum_{\tau \in S_K} \mathbb{E}_{\tau^{-1}(\mathcal{P})}\left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right] \\ &= \frac{1}{K!} \sum_{\sigma \in S_K} \mathbb{E}_{\sigma(\mathcal{P})}\left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right]. \end{aligned}$$

Since the "loss of the best expert" parts of all three measures are invariant under any permutation of $\mathcal{P}$ (see the proof of Lemma 1), we have:

$$R(\widetilde{\boldsymbol{\omega}}, \mathcal{P}) = \frac{1}{K!} \sum_{\sigma \in S_K} R(\boldsymbol{\omega}, \sigma(\mathcal{P})) \leq \max_{\sigma \in S_K} R(\boldsymbol{\omega}, \sigma(\mathcal{P})).$$

Hence,

$$\sup_{\mathcal{P}} R(\widetilde{\boldsymbol{\omega}}, \mathcal{P}) \leq \sup_{\mathcal{P}} \max_{\sigma \in S_K} R(\boldsymbol{\omega}, \sigma(\mathcal{P})) = \sup_{\mathcal{P}} R(\boldsymbol{\omega}, \mathcal{P}).$$

Proof of Theorem 3

For any distribution P_k over binary losses, let $p_k := P_k(\ell_{t,k} = 1) = \mathbb{E}_{P_k}[\ell_{t,k}]$. We have:

$$\begin{aligned}
\overline{\mathrm{loss}}_t(\boldsymbol{w}_t, \mathcal{P}) &= \frac{1}{K!}\sum_{\sigma} \mathbb{E}_{\sigma(\mathcal{P})}\left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right] \qquad (2)\\
&= \frac{1}{K!}\sum_{\sigma} \mathbb{E}_{\sigma(\mathcal{P})}\left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1})\right] \cdot \mathbb{E}_{\sigma(\mathcal{P})}\left[\boldsymbol{\ell}_t\right]\\
&= \frac{1}{K!}\sum_{\sigma}\sum_{\boldsymbol{\ell}^{t-1}} \left(\prod_{k=1}^{K} p_{\sigma(k)}^{\sum_{q=1}^{t-1}\ell_{q,k}} (1-p_{\sigma(k)})^{t-1-\sum_{q=1}^{t-1}\ell_{q,k}}\right)\left(\sum_{k=1}^{K} w_{t,k}(\boldsymbol{\ell}^{t-1}) p_{\sigma(k)}\right)\\
&= \frac{1}{K!}\sum_{\boldsymbol{\ell}^{t-1}} \underbrace{\sum_{k=1}^{K} w_{t,k}(\boldsymbol{\ell}^{t-1}) \left(\sum_{\sigma}\prod_{j=1}^{K} p_{\sigma(j)}^{\sum_{q=1}^{t-1}\ell_{q,j}} (1-p_{\sigma(j)})^{t-1-\sum_{q=1}^{t-1}\ell_{q,j}} p_{\sigma(k)}\right)}_{=: \overline{\mathrm{loss}}_t(\boldsymbol{w}_t, \mathcal{P}|\boldsymbol{\ell}^{t-1})},
\end{aligned}$$

where in the second equality we used the fact that $\boldsymbol{w}_t$ depends on $\boldsymbol{\ell}^{t-1}$ and does not depend on $\boldsymbol{\ell}_t$. Fix $\boldsymbol{\ell}^{t-1}$ and consider the term $\overline{\mathrm{loss}}_t(\boldsymbol{w}_t, \mathcal{P}|\boldsymbol{\ell}^{t-1})$. This term is linear in $\boldsymbol{w}_t$, hence it is minimized by $\boldsymbol{w}_t = \boldsymbol{e}_k$ for some $k = 1, \ldots, K$, where $\boldsymbol{e}_k$ is the k-th standard basis vector with 1 on the k-th coordinate, and zeros on the remaining coordinates. We will use a shorthand notation $L_j = \sum_{q=1}^{t-1} \ell_{q,j}$, for $j = 1, \ldots, K$, and $\boldsymbol{L} = (L_1, \ldots, L_K)$. In this notation, we rewrite $\overline{\mathrm{loss}}_t(\boldsymbol{w}_t, \mathcal{P}|\boldsymbol{\ell}^{t-1})$ as:

$$\overline{\mathrm{loss}}_t(\boldsymbol{w}_t, \mathcal{P}|\boldsymbol{\ell}^{t-1}) = \sum_{k=1}^{K} w_{t,k}(\boldsymbol{\ell}^{t-1}) \left(\sum_{\sigma}\prod_{j=1}^{K} p_{\sigma(j)}^{L_j}(1-p_{\sigma(j)})^{t-1-L_j} p_{\sigma(k)}\right), \qquad (3)$$

We will show that for any $\mathcal{P}$, and any $\boldsymbol{\ell}^{t-1}$ (and hence, any $\boldsymbol{L}$), $\overline{\mathrm{loss}}_t(\boldsymbol{w}_t, \mathcal{P}|\boldsymbol{\ell}^{t-1})$ is minimized by setting $w_t = \boldsymbol{e}_{k^*}$ for any $k^* \in \operatorname{argmin}_j L_j$. In other words, we will show that for any $\mathcal{P}$, $\boldsymbol{L}$, any $k^* \in \operatorname{argmin}_j L_j$, and any $k = 1, \ldots, K$,

$$\overline{\mathrm{loss}}_t(\boldsymbol{e}_{k^*}, \mathcal{P}|\boldsymbol{\ell}^{t-1}) \leq \overline{\mathrm{loss}}_t(\boldsymbol{e}_k, \mathcal{P}|\boldsymbol{\ell}^{t-1}).$$

or equivalently, using (3), that for any $\mathcal{P}$, $\boldsymbol{L}$, $k^* \in \operatorname{argmin}_j L_j$, and $k = 1, \ldots, K$,

$$\sum_{\sigma}\prod_{j=1}^{K} p_{\sigma(j)}^{L_j}(1-p_{\sigma(j)})^{t-1-L_j} p_{\sigma(k^*)} \leq \sum_{\sigma}\prod_{j=1}^{K} p_{\sigma(j)}^{L_j}(1-p_{\sigma(j)})^{t-1-L_j} p_{\sigma(k)} \qquad (4)$$

We proceed by induction on K. Take $K = 2$ and note that when $k^* = k$, there is nothing to prove, as both sides of (4) are identical. Therefore, without loss of generality, assume $k^* = 1$ and $k = 2$, which implies $L_1 \leq L_2$. Then, (4) reduces to:

$$\begin{aligned}
&p_1^{L_1} p_2^{L_2}(1-p_1)^{t-1-L_1}(1-p_2)^{t-1-L_2} p_1 + p_2^{L_1} p_1^{L_2}(1-p_2)^{t-1-L_1}(1-p_1)^{t-1-L_2} p_2\\
&\quad \leq p_1^{L_1} p_2^{L_2}(1-p_1)^{t-1-L_1}(1-p_2)^{t-1-L_2} p_2 + p_2^{L_1} p_1^{L_2}(1-p_2)^{t-1-L_1}(1-p_1)^{t-1-L_2} p_1,
\end{aligned}$$

After rearranging the terms, it amounts to show that:

$$(p_1 p_2)^{L_1}\big((1-p_1)(1-p_2)\big)^{t-1-L_2}(p_1 - p_2)\left((p_2(1-p_1))^{L_2-L_1} - (p_1(1-p_2))^{L_2-L_1}\right) \leq 0.$$

But this will hold if:

$$(p_1 - p_2)\left((p_2(1-p_1))^{L_2-L_1} - (p_1(1-p_2))^{L_2-L_1}\right) \leq 0. \qquad (5)$$

If $L_1 = L_2$, (5) clearly holds; therefore assume $L_1 < L_2$. We prove the validity of (5) by noticing that:

$$p_2(1-p_1) > p_1(1-p_2) \geq 0 \quad \Longleftrightarrow \quad p_2 > p_1,$$

which means that the two factors of the product on the left-hand side of (5) have the opposite sign (when $p_1 \neq p_2$) or are zero at the same time (when $p_1 = p_2$). Hence, we proved (5), which implies

(4) when $k^* = 1$ and $k = 2$. The opposite case $k^* = 2, k = 1$ with $L_2 \leq L_1$ can be shown with exactly the same line of arguments by simply exchanging the indices 1 and 2.

Now, we assume (4) holds for $K - 1 \geq 2$ experts and any $\mathcal{P} = (P_1, \ldots, P_{K-1})$, any $\boldsymbol{L} = (L_1, \ldots, L_{K-1})$, any $k^* \in \operatorname{argmin}_{j=1,\ldots,K-1} L_j$, and any $k = 1, \ldots, K-1$, and we show that it also holds for K experts. Take any $k^* \in \operatorname{argmin}_{j=1,\ldots,K} L_j$, and any $k = 1, \ldots, K$. Without loss of generality, assume that $k^* \neq 1$ and $k \neq 1$ (it is always possible find expert different than k^* and k, because there are $K \geq 3$ experts). We expand the sum over permutations on the left-hand side of (4) with respect to the value of $\sigma(1)$:

$$\sum_{s=1}^{K} p_s^{L_1} (1 - p_s)^{t-1-L_1} \sum_{\sigma\colon \sigma(1)=s} \prod_{j=2}^{K} p_{\sigma(j)}^{L_j} (1 - p_{\sigma(j)})^{t-1-L_j} p_{\sigma(k^*)},$$

and we also expand the sum on the right-hand side of (4) in the same way. To prove (4), it suffices to show that every term in the sum over s on the left-hand side is not greater than the corresponding term in the sum on the right-hand side, i.e. to show that for any $s = 1, \ldots, K$,

$$\sum_{\sigma\colon \sigma(1)=s} \prod_{j=2}^{K} p_{\sigma(j)}^{L_j} (1 - p_{\sigma(j)})^{t-1-L_j} p_{\sigma(k^*)} \leq \sum_{\sigma\colon \sigma(1)=s} \prod_{j=2}^{K} p_{\sigma(j)}^{L_j} (1 - p_{\sigma(j)})^{t-1-L_j} p_{\sigma(k)}. \quad (6)$$

We now argue that this inequality follows directly from the inductive assumption by dropping L_1 and P_s, and applying (4) to such a $(K-1)$-expert case. More precisely, note that the sum on both sides of (6) goes over all permutations on indices $(1, \ldots, s-1, s+1, \ldots, K)$ and since $k, k^* \neq 1$, $k^* \in \operatorname{argmin}_{j=2,\ldots,K} L_j$ and $k \geq 2$. Hence, applying (4) to $K-1$ expert case with $K-1$ distributions $(P_1, P_2, \ldots, P_{s-1}, P_{s+1}, \ldots, P_K)$ (or any permutation thereof), and $K-1$ integers $(L_2, \ldots, L_K)$ immediately implies (6).

Thus, we proved (4) which states that $\overline{\operatorname{loss}}_t(\boldsymbol{w}_t, \mathcal{P} | \boldsymbol{\ell}^{t-1})$ is minimized by any leader $k^* \in \operatorname{argmin}_j L_j$, where $L_j = \sum_{q=1}^{t-1} \ell_{q,j}$. This means $\overline{\operatorname{loss}}_t(\boldsymbol{w}_t, \mathcal{P} | \boldsymbol{\ell}^{t-1})$ is also minimized by the FL strategy $\boldsymbol{w}_t^{\text{fl}}$, which distributes its mass uniformly over all leaders. Since FL minimizes $\overline{\operatorname{loss}}_t(\boldsymbol{w}_t, \mathcal{P} | \boldsymbol{\ell}^{t-1})$ for any $\boldsymbol{\ell}^{t-1}$, by (2) it also minimizes $\overline{\operatorname{loss}}_t(\boldsymbol{w}_t, \mathcal{P})$.

Note that the proof did not require uniform tie breaking over leaders, as any distribution over leaders would work as well. Uniform distribution, however, makes the FL strategy permutation invariant.

Proof of Lemma 5

Let p_k be the expectation of $\ell_{t,k}$ according to either P_k or P_k^{bin}, $p_k := \mathbb{E}_{P_k}[\ell_{t,k}] = \mathbb{E}_{P_k^{\text{bin}}}[\ell_{t,k}]$. Since for any prediction strategy $\boldsymbol{\omega}$, $\boldsymbol{w}_t$ depends on $\boldsymbol{\ell}^{t-1}$ and does not depend on $\boldsymbol{\ell}_t$, we have $\mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t^{\text{bfl}} \cdot \boldsymbol{\ell}_t\right] = \mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t^{\text{bfl}}\right] \cdot \mathbb{E}_{\mathcal{P}}\left[\boldsymbol{\ell}_t\right] = \mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t^{\text{bfl}}\right] \cdot \boldsymbol{p}$, where $\boldsymbol{p} = (p_1, \ldots, p_K)$. Simiarly, $\mathbb{E}_{\mathcal{P}^{\text{bin}}}\left[\boldsymbol{w}_t^{\text{fl}} \cdot \boldsymbol{\ell}_t\right] = \mathbb{E}_{\mathcal{P}^{\text{bin}}}\left[\boldsymbol{w}_t^{\text{fl}}\right] \cdot \boldsymbol{p}$. Hence, we only need to show that $\mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t^{\text{bfl}}\right] = \mathbb{E}_{\mathcal{P}^{\text{bin}}}\left[\boldsymbol{w}_t^{\text{fl}}\right]$. This holds because $\boldsymbol{w}_t^{\text{bfl}}$ "sees" only binary outcomes resulting from the joint distribution of $\mathcal{P}$ and the distribution of binarization variables:

$$\mathbb{E}_{\boldsymbol{\ell}^{t-1} \sim \mathcal{P}^{t-1}}\left[\boldsymbol{w}_t^{\text{bfl}}(\boldsymbol{\ell}^{t-1})\right] = \mathbb{E}_{\boldsymbol{\ell}^{t-1} \sim \mathcal{P}^{t-1}, \boldsymbol{b}^{t-1}}\left[\boldsymbol{w}_t^{\text{fl}}(\boldsymbol{b}^{t-1})\right],$$

and for any $b_{t,k}$, the probability (jointly over P_k and binarization variables) of $b_{t,k} = 1$ is the same as probability of $\ell_{t,k} = 1$ over the distribution P_k^{bin}:

$$\begin{aligned} P(b_{t,k} = 1) &= \int_{[0,1]} P(b_{t,k} = 1 | \ell_{t,k}) P_k(\ell_{t,k}) \mathrm{d}\ell_{t,k} \\ &= \int_{[0,1]} \ell_{t,k} P_k(\ell_{t,k}) \mathrm{d}\ell_{t,k} \; = \; p_t = P^{\text{bin}}(\ell_{t,k} = 1). \end{aligned}$$

Hence,

$$\mathbb{E}_{\boldsymbol{\ell}^{t-1} \sim \mathcal{P}^{t-1}, \boldsymbol{b}^{t-1}}\left[\boldsymbol{w}_t^{\text{fl}}(\boldsymbol{b}^{t-1})\right] = \mathbb{E}_{\boldsymbol{\ell}^{t-1} \sim (\mathcal{P}^{\text{bin}})^{t-1}}\left[\boldsymbol{w}_t^{\text{fl}}(\boldsymbol{\ell}^{t})\right].$$

Proof of Lemma 6

Lemma 5 shows that the expected loss of $\boldsymbol{\omega}^{\text{bfl}}$ on $\mathcal{P}$ is the same as the expected loss of $\boldsymbol{\omega}^{\text{fl}}$ on $\mathcal{P}^{\text{bin}}$. Hence, to prove the inequality, we only need to consider the "loss of the best expert" part of each measure. For the expected redundancy, and the expected regret, it directly follows from the definition of $\mathcal{P}^{\text{bin}}$ that for any t, k, $\mathbb{E}_{\mathcal{P}}[\ell_{t,k}] = \mathbb{E}_{\mathcal{P}^{\text{bin}}}[\ell_{t,k}]$, hence $\min_k \mathbb{E}_{\mathcal{P}}[\ell_{T,k}] = \min_k \mathbb{E}_{\mathcal{P}^{\text{bin}}}[\ell_{T,k}]$, and simiarly, $\min_k \mathbb{E}_{\mathcal{P}}[L_{T,k}] = \min_k \mathbb{E}_{\mathcal{P}^{\text{bin}}}[L_{T,k}]$. Thus, for the expected redundancy and the excess risk, the lemma actually holds with equality.

For the expected regret, we will show that $\mathbb{E}_{\mathcal{P}}[\min_k L_{T,k}] \geq \mathbb{E}_{\mathcal{P}^{\text{bin}}}[\min_k L_{T,k}]$, which will finish the proof. Using the argument from the proof of Lemma 5, and denoting $B_{T,k} = \sum_{t=1}^{T} b_{t,k}$, we have:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\ell}^T \sim (\mathcal{P}^{\text{bin}})^T}[\min_k L_{T,k}] \;&=\; \mathbb{E}_{\boldsymbol{\ell}^T \sim \mathcal{P}^T, \boldsymbol{b}^T}[\min_k B_{T,k}] \;\leq\; \mathbb{E}_{\boldsymbol{\ell}^T \sim \mathcal{P}^T}\left[\min_k \mathbb{E}_{\boldsymbol{b}^T}[B_{T,k} | \boldsymbol{\ell}^T]\right] \\ &=\; \mathbb{E}_{\boldsymbol{\ell}^T \sim \mathcal{P}^T}[\min_k L_{T,k}],\end{aligned}$$

where the inequality follows from Jensen's inequality applied to a concave function $\min(\cdot)$.

Proof of Theorem 7

Since $\boldsymbol{\omega}^{\text{bfl}}$ is the same as $\boldsymbol{\omega}^{\text{fl}}$ when all the losses are binary, $R(\boldsymbol{\omega}^{\text{bfl}}, \mathcal{P}) = R(\boldsymbol{\omega}^{\text{fl}}, \mathcal{P})$ for every $\mathcal{P}$ over binary losses. Furthermore, Lemma 6 states that $R(\boldsymbol{\omega}^{\text{bfl}}, \mathcal{P}) \leq R(\boldsymbol{\omega}^{\text{bfl}}, \mathcal{P}^{\text{bin}})$, i.e. for every $\mathcal{P}$ over continuous losses, there is a corresponding $\mathcal{P}^{\text{bin}}$ over binary losses which incurs at least the same regret/redundancy/risk to $\boldsymbol{\omega}^{\text{bfl}}$. Therefore,

$$\sup_{\mathcal{P} \text{ on } [0,1]} R(\boldsymbol{\omega}^{\text{bfl}}, \mathcal{P}) \;=\; \sup_{\mathcal{P} \text{ on } \{0,1\}} R(\boldsymbol{\omega}^{\text{bfl}}, \mathcal{P}) \;=\; \sup_{\mathcal{P} \text{ on } \{0,1\}} R(\boldsymbol{\omega}^{\text{fl}}, \mathcal{P}).$$

By the second part of Corollary 4, for any prediction strategy $\boldsymbol{\omega}$:

$$\sup_{\mathcal{P} \text{ on } \{0,1\}} R(\boldsymbol{\omega}^{\text{fl}}, \mathcal{P}) \;\leq\; \sup_{\mathcal{P} \text{ on } \{0,1\}} R(\boldsymbol{\omega}, \mathcal{P}) \;\leq\; \sup_{\mathcal{P} \text{ on } [0,1]} R(\boldsymbol{\omega}, \mathcal{P}),$$

which finishes the proof.

A counterexample to FL strategy

We now argue that the vanilla FL is not the minimax strategy for continuous losses, so that the binarization procedure is justified. We will only consider excess risk for simplicity, but we can use similar arguments to show a counterexample for the expected regret and the expected redundancy as well. Take $K = 2$ experts, $T = 2$ iterations and distributions P_1, P_2 on binary losses. Denote $p_1 = P_1(\ell_{t,1} = 1)$ and $p_2 = P_2(\ell_{t,2} = 1)$, and assume $p_1 \leq p_2$. The risk of FL strategy is given by:

$$\underbrace{\big(p_1 p_2 + (1-p_1)(1-p_2)\big)\frac{p_1 + p_2}{2}}_{\text{ties}} + p_1(1-p_2)p_2 + p_2(1-p_1)p_1 - p_1 = \frac{\delta}{2} - \frac{\delta^2}{2},$$

where $\delta = p_2 - p_1$. Maximizing over δ the expression above gives $\delta^* = \frac{1}{2}$ and the maximum risk of FL on binary losses is equal to $\frac{1}{8}$. This is also the minimax risk on continuous losses, as follows from the proof of Theorem 7 (because the binarized FL is the minimax strategy on continuous losses, and it achieves the maximum risk on binary losses). We now show that there exist distributions P_1, P_2 on continuous losses which force FL to suffer more excess risk than $\frac{1}{8}$. We take P_1 with support on two points $\{\epsilon, 1\}$, where ϵ is a very small positive number, and $p_1 = P_1(\ell_{t,1} = 1)$. Note that $\mathbb{E}[\ell_{t,1}] = p_1 + \epsilon(1 - p_1)$. P_2 has support on $\{0, 1-\epsilon\}$, and let $p_2 = P_2(\ell_{t,2} = 1)$. We also assume $p_1 < p_2$, i.e. expert 1 is the "better" expert. The main idea in this counterexample is that by using ϵ values, all "ties" are resolved in favor of expert 2, which makes the FL algorithm suffer more loss. More precisely, this risk of FL is now given by:

$$\underbrace{\big(p_1 p_2 + (1-p_1)(1-p_2)\big)p_2}_{\text{ties}} + p_1(1-p_2)p_2 + p_2(1-p_1)p_1 - p_1 + O(\epsilon).$$

Choosing, e.g. $p_1 = 0$ and $p_2 = 0.5$, this gives $\frac{1}{4} + O(\epsilon)$ excess risk, which is more than $\frac{1}{8}$, given tha we take ϵ sufficiently small.