On Minimaxity of Follow the Leader Strategy in the Stochastic Setting

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 "admissible" 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 [4,5], the learner sequentially decides on one of K experts to follow, and suffers loss associated with the chosen expert. The difference between the learner'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. A prediction strategy 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: for 0/1 losses on the binary labels [4,5], for unit losses with fixed loss budget [2], and when $K = 2$ [9]. Interestingly, all these algorithms share a similar strategy of playing against a maximin adversary which assigns losses uniformly at random. They also have the *equalization* 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 is not adversarially generated[1].

* This research was supported by the Polish National Science Centre under grant no. 2013/11/D/ST6/03050.

[1] There are various algorithms which combine almost optimal worst-case performance with good performance on "easy" sequences [12,6,13,10,11]; 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. The motivation behind our assumption is the practical usefulness of the stochastic setting: the data encountered in practice are rarely adversarial and can often be modeled as generated from a fixed (yet unknown) distribution. That is why we believe 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 [4,5] or sequential minimax duality [1,9], 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 "admissible" 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* [8], 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 [5]. On the contrary, in the stochastic setting with binary losses, FL has 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 (which results in the *binarized FL* strategy), we obtain the minimax strategy in the continuous case.

We note that when the excess risk is used as a performance metric, our setup falls into the framework of statistical decision theory [7,3], and the question we pose can be reduced to the problem of finding the minimax decision rule for a properly constructed loss function, which matches the excess risk on expectation. In principle, one could try to solve our problem by using the complete class theorem and search for the minimax rule within the class of (generalized) Bayesian decision rules. We initially followed this approach, but it turned out to be futile, as the class of distributions we are considering are all distributions in the range $[0, 1]$, and exploring prior distributions over such classes becomes very difficult. On the other hand, the analysis presented in this paper is relatively

simple, and works not only for the excess risk, but also for the expected regret and the expected redundancy. To the best of our knowledge, both the results and the analysis presented here are novel.

The paper is organized as follows. In Section 2 we formally define the problem. The binary case is solved in Section 3, while Section 4 concerns continuous case. Section 5, concludes the paper and discusses an open problem.

2 Problem Setting

2.1 Prediction with Expert Advice in the Stochastic Setting

In the game of prediction with expert advice, at each trial $t = 1, \ldots, T$, the learner predicts with a distribution $\boldsymbol{w}_t = (w_{t,1}, \ldots, w_{t,K})$ over K experts. Then, the loss vector $\boldsymbol{\ell}_t = (\ell_{t,1}, \ldots, \ell_{t,K}) \in \mathcal{X}^K$ is revealed (where $\mathcal{X}$ is either $\{0, 1\}$ or $[0, 1]$), and the learner suffers loss:

$$\boldsymbol{w}_t \cdot \boldsymbol{\ell}_t = \sum_{k=1}^{K} w_{t,k} \ell_{t,k},$$

which can be interpreted as the expected loss the learner suffers by following one of the experts chosen randomly according to $\boldsymbol{w}_t$. Let $L_{t,k}$ denote the cumulative loss of expert k at the end of iteration t, $L_{t,k} = \sum_{q \leq t} \ell_{q,k}$. Let $\boldsymbol{\ell}^t$ abbreviate the sequence of losses $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_t$. We will also use $\boldsymbol{\omega} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T)$ to denote the whole prediction strategy of the learner, having in mind that each distribution $\boldsymbol{w}_t$ is a function of the past $t-1$ outcomes $\boldsymbol{\ell}^{t-1}$. The performance of the strategy is measured by means of *regret*:

$$\sum_{t=1}^{T} \boldsymbol{w}_t \cdot \boldsymbol{\ell}_t - \min_k L_{T,k},$$

which is a difference between the algorithm's cumulative loss and the cumulative loss of the best expert. In the worst-case (adversarial) formulation of the problem, no assumption is made on the way the sequence of losses is generated, and hence the goal is then to find an algorithm which minimizes the worst-case regret over all possible sequences $\boldsymbol{\ell}^T$.

In this paper, we drop the analysis of worst-case performance and explore the minimax principle in the *stochastic* setting, defined as follows. 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 $k \neq k'$. The prediction strategy is then evaluated by means of *expected regret*:

$$R_{\mathrm{eg}}(\boldsymbol{\omega}, \mathcal{P}) = \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t - \min_k L_{T,k} \right],$$

Expected regret:	$R_{\mathrm{eg}}(\boldsymbol{\omega}, \mathcal{P}) = \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t - \min_k L_{T,k}\right]$
Expected redundancy:	$R_{\mathrm{ed}}(\boldsymbol{\omega}, \mathcal{P}) = \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right] - \min_k \mathbb{E}\left[L_{T,k}\right]$
Excess risk:	$R_{\mathrm{isk}}(\boldsymbol{\omega}, \mathcal{P}) = \mathbb{E}\left[\boldsymbol{w}_T(\boldsymbol{\ell}^{T-1}) \cdot \boldsymbol{\ell}_T\right] - \min_k \mathbb{E}\left[\ell_{T,k}\right]$

Table 1. Performance measures.

where the expectation over the loss sequences $\boldsymbol{\ell}^T$ with respect to distribution $\mathcal{P} = (P_1, \ldots, P_k)$, and we explicitly indicate the dependency of $\boldsymbol{w}_t$ on $\boldsymbol{\ell}^{t-1}$. However, the expected regret is not the only performance metric one can use in the stochastic setting. Instead of comparing the algorithm's loss to the loss of the best expert on the actual outcomes, one can choose the *best expected* expert as a comparator, which leads to a metric:

$$R_{\mathrm{ed}}(\boldsymbol{\omega}, \mathcal{P}) = \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right] - \min_k \mathbb{E}\left[L_{T,k}\right],$$

which we call the *expected redundancy*, as it closely resembles a measure used in information theory to quantify the excess codelength of a prequential code[8]. Note that from Jensen's inequality it holds that $R_{\mathrm{ed}}(\boldsymbol{\omega}, \mathcal{P}) \geq R_{\mathrm{eg}}(\boldsymbol{\omega}, \mathcal{P})$ for any $\boldsymbol{\omega}$ and any $\mathcal{P}$, and the difference $R_{\mathrm{ed}}(\boldsymbol{\omega}, \mathcal{P}) - R_{\mathrm{eg}}(\boldsymbol{\omega}, \mathcal{P})$ is independent of $\boldsymbol{\omega}$ given fixed $\mathcal{P}$. This does not, however, imply that these metrics are equivalent in the minimax analysis, as the set of distributions $\mathcal{P}$ is chosen by the adversary *against* strategy $\boldsymbol{\omega}$ played by learner, and this choice will in general be different for the expected regret and the expected redundancy. Finally, the stochastic setting permits us to evaluate the prediction strategy by means of the *individual* rather than cumulative losses. Thus, it is reasonable to define the *excess risk* of the prediction strategy at time T:

$$R_{\mathrm{isk}}(\boldsymbol{\omega}, \mathcal{P}) = \mathbb{E}\left[\boldsymbol{w}_T(\boldsymbol{\ell}^{T-1}) \cdot \boldsymbol{\ell}_T\right] - \min_k \mathbb{E}\left[\ell_{T,k}\right],$$

a metric traditionally used in statistics to measure the accuracy of statistical procedures. Contrary to the expected regret and redundancy defined by means of cumulative losses of the prediction strategy, the excess risk concerns only a single prediction at a given trial; hence, without loss of generality, we can choose the last trial T in the definition. For the sake of clarity, we summarize the three measures in Table 1.

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.

2.2 Permutation Invariance

In this section, we identify a class of "admissible" prediction strategies, which we call permutation invariant. The name comes from the fact that the performance of these strategies remains invariant under any permutation of the distributions $\mathcal{P} = (P_1, \ldots, P_K)$. We show that that for every prediction strategy, there exists a corresponding permutation invariant strategy with not worse expected regret, redundancy and excess risk in the worst-case with respect to all permutations of $\mathcal{P}$.

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 denote $\sigma(\boldsymbol{v}) = (v_{\sigma(1)}, \ldots, v_{\sigma(K)})$ and $\sigma(\boldsymbol{\ell}^{t-1}) = \sigma(\boldsymbol{\ell}_1), \ldots, \sigma(\boldsymbol{\ell}_{t-1})$. 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. The performance of these strategies remains invariant under any permutation of the distributions from $\mathcal{P}$:

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 $\sigma(\mathcal{P}) = (P_{\sigma(1)}, \ldots, P_{\sigma(K)})$.*

Proof. 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 best expert" 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 $\mathcal{P}$. □

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

Theorem 1. *For any strategy $\boldsymbol{\omega}$, there exists permutation invariant strategy $\widetilde{\boldsymbol{\omega}}$, such that for any set of distributions $\mathcal{P}$,*

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

where R is either the expected regret, the expected redundancy or the excess risk. In particular, this implies that: $\sup_{\mathcal{P}} R(\widetilde{\boldsymbol{\omega}}, \mathcal{P}) \leq \sup_{\mathcal{P}} R(\boldsymbol{\omega}, \mathcal{P})$.

Proof. This first equality in the theorem immediately follows from Lemma 1. 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 a distribution over K experts ($\widetilde{\boldsymbol{w}}_t$ 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 equality 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:

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

This implies that:

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

□

Theorem 1 states that strategies which are not permutation-invariant do not give any advantage over permutation-invariant strategies even when the set of distributions $\mathcal{P}$ is fixed (and even possibly known to the learner), but the adversary can permute the distributions to make the learner incur the most loss. We also note that one can easily show a slightly stronger version of Theorem 1: if strategy $\boldsymbol{\omega}$ is not permutation invariant, and it holds that $R(\boldsymbol{\omega}, \mathcal{P}) \neq R(\boldsymbol{\omega}, \tau(\mathcal{P}))$ for some set of distributions and permutation τ, then $R(\widetilde{\boldsymbol{\omega}}, \mathcal{P}) < \max_{\sigma \in S_K} R(\boldsymbol{\omega}, \sigma(\mathcal{P}))$. This follows from the fact that the inequality in (1) becomes sharp.

2.3 Follow the Leader Strategy

Given loss sequence $\boldsymbol{\ell}^{t-1}$, let $N = |\operatorname{argmin}_{j=1,\dots,K} L_{t-1,j}|$ be the size of the leader set at the beginning of trial t. We define the *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 straightforward to show that such defined 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 $\boldsymbol{\omega}$. It follows from Lemma 1 that for any $\mathcal{P}$, and any permutation $\sigma \in S_K$, $\mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right] = \mathbb{E}_{\sigma(\mathcal{P})}\left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right]$. Averaging this equality over all permutations $\sigma \in S_K$ gives:

$$\mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right] = \underbrace{\frac{1}{K!}\sum_{\sigma} \mathbb{E}_{\sigma(\mathcal{P})}\left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t\right]}_{=:\ \overline{\text{loss}}_t(\boldsymbol{w}_t, \mathcal{P})}, \tag{2}$$

where we defined $\overline{\text{loss}}_t(\boldsymbol{w}_t, \mathcal{P})$ to be permutation-averaged expected loss at trial t. We now show the main result of this paper, a surprisingly strong property of FL strategy, which states that FL minimizes $\overline{\text{loss}}_t(\boldsymbol{w}_t, \mathcal{P})$ *simultaneously* over all K-sets of distributions. Hence, FL is not only optimal in the worst case, but is actually optimal for permutation-averaged expected loss *for any* $\mathcal{P}$, even if $\mathcal{P}$ is known to the learner! The consequence of this fact (by (2)) is that *FL has the smallest expected loss among all permutation invariant strategies for any* $\mathcal{P}$ (again, even if $\mathcal{P}$ is known to the learner).

Theorem 2. *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 $t = 1, \ldots, T$:*

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

Proof. 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:

$$\begin{aligned}
\overline{\text{loss}}_t(\boldsymbol{w}_t, \mathcal{P}) &= \frac{1}{K!} \sum_{\sigma} \mathbb{E}_{\sigma(\mathcal{P})} \left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \cdot \boldsymbol{\ell}_t \right] \qquad (3) \\
&= \frac{1}{K!} \sum_{\sigma} \mathbb{E}_{\sigma(\mathcal{P})} \left[\boldsymbol{w}_t(\boldsymbol{\ell}^{t-1}) \right] \cdot \mathbb{E}_{\sigma(\mathcal{P})} \left[\boldsymbol{\ell}_t \right] \\
&= \frac{1}{K!} \sum_{\sigma} \sum_{\boldsymbol{\ell}^{t-1}} \left(\prod_{k=1}^{K} p_{\sigma(k)}^{L_{t-1,k}} (1 - p_{\sigma(k)})^{t-1-L_{t-1,k}} \right) \left(\sum_{k=1}^{K} w_{t,k}(\boldsymbol{\ell}^{t-1}) p_{\sigma(k)} \right) \\
&= \frac{1}{K!} \sum_{\boldsymbol{\ell}^{t-1}} \underbrace{\sum_{k=1}^{K} w_{t,k}(\boldsymbol{\ell}^{t-1}) \left(\sum_{\sigma} \prod_{j=1}^{K} p_{\sigma(j)}^{L_{t-1,j}} (1 - p_{\sigma(j)})^{t-1-L_{t-1,j}} p_{\sigma(k)} \right)}_{=: \overline{\text{loss}}_t(\boldsymbol{w}_t, \mathcal{P} | \boldsymbol{\ell}^{t-1})},
\end{aligned}$$

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 $\overline{\text{loss}}_t(\boldsymbol{w}_t, \mathcal{P} | \boldsymbol{\ell}^{t-1})$. 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 drop the trial index and use a shorthand notation $L_j = L_{t-1,j}$, for $j = 1, \ldots, K$, and $\boldsymbol{L} = (L_1, \ldots, L_K)$. In this notation, we rewrite $\overline{\text{loss}}_t(\boldsymbol{w}_t, \mathcal{P} | \boldsymbol{\ell}^{t-1})$ as:

$$\overline{\text{loss}}_t(\boldsymbol{w}_t, \mathcal{P} | \boldsymbol{\ell}^{t-1}) \;=\; \sum_{k=1}^{K} w_{t,k}(\boldsymbol{\ell}^{t-1}) \left(\sum_{\sigma} \prod_{j=1}^{K} p_{\sigma(j)}^{L_j} (1 - p_{\sigma(j)})^{t-1-L_j} p_{\sigma(k)} \right). \quad (4)$$

We will show that for any $\mathcal{P}$, and any $\boldsymbol{\ell}^{t-1}$ (and hence, any $\boldsymbol{L}$), $\overline{\text{loss}}_t(\boldsymbol{w}_t, \mathcal{P} | \boldsymbol{\ell}^{t-1})$ 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 $k = 1, \ldots, K$,

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

or equivalently, using (4), that for any $\mathcal{P}$, $\boldsymbol{L}$, $k^* \in \operatorname{argmin}_j L_j$, and $k = 1, \ldots, K$,

$$\sum_{\sigma} \prod_{j=1}^{K} p_{\sigma(j)}^{L_j} (1 - p_{\sigma(j)})^{t-1-L_j} p_{\sigma(k^*)} \leq \sum_{\sigma} \prod_{j=1}^{K} p_{\sigma(j)}^{L_j} (1 - p_{\sigma(j)})^{t-1-L_j} p_{\sigma(k)}. \quad (5)$$

We proceed by induction on K. Take $K = 2$ and note that when $k^* = k$, there is nothing to prove, as both sides of (5) are identical. Therefore, without loss of generality, assume $k^* = 1$ and $k = 2$, which implies $L_1 \leq L_2$. Then, (5) reduces to:

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

After rearranging the terms, it amounts to show that:

$$\begin{aligned}
(p_1 p_2)^{L_1} \Big((1-p_1)(1-p_2)\Big)^{t-1-L_2} (p_1 - p_2) \\
\times \Big((p_2(1-p_1))^{L_2-L_1} - (p_1(1-p_2))^{L_2-L_1}\Big) \leq 0.
\end{aligned}$$

But this will hold if:

$$(p_1 - p_2)\Big((p_2(1-p_1))^{L_2-L_1} - (p_1(1-p_2))^{L_2-L_1}\Big) \leq 0. \tag{6}$$

If $L_1 = L_2$, (6) clearly holds; therefore assume $L_1 < L_2$. We prove the validity of (6) by noticing that:

$$p_2(1-p_1) > p_1(1-p_2) \quad \iff \quad p_2 > p_1,$$

which means that the two factors of the product on the left-hand side of (6) 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 (6), which implies (5) 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 (5) 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 (5) 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 (5) in the same way. To prove (5), 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 $s = 1, \dots, K$,

$$\sum_{\sigma\colon \sigma(1)=s} \prod_{j=2}^{K} p_{\sigma(j)}^{L_j} (1-p_{\sigma(j)})^{t-1-L_j} p_{\sigma(k^*)} \le \sum_{\sigma\colon \sigma(1)=s} \prod_{j=2}^{K} p_{\sigma(j)}^{L_j} (1-p_{\sigma(j)})^{t-1-L_j} p_{\sigma(k)}. \tag{7}$$

We now argue that this inequality follows directly from the inductive assumption by dropping L_1 and P_s, and applying (5) to such a $(K-1)$-expert case. More precisely, note that the sum on both sides of (7) goes over all permutations on indices $(1, \dots, s-1, s+1, \dots, K)$ and since $k, k^* \neq 1$, $k^* \in \operatorname{argmin}_{j=2,\dots,K} L_j$ and $k \geq 2$. Hence, applying (5) to $K-1$ expert case with $K-1$ distributions $(P_1, P_2, \dots, P_{s-1}, P_{s+1}, \dots, P_K)$ (or any permutation thereof), and $K-1$ integers $(L_2, \dots, L_K)$ immediately implies (7).

Thus, we proved (5) 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 $\boldsymbol{w}_t^{\mathrm{fl}}$, 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 (3) 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.

The consequence of Theorem 2 is the following corollary which states the minimaxity of FL strategy for binary losses:

Corollary 1. *Let $\boldsymbol{\omega}^{\mathrm{fl}} = (\boldsymbol{w}_1^{\mathrm{fl}}, \dots, \boldsymbol{w}_T^{\mathrm{fl}})$ be the FL strategy. Then, for any $\mathcal{P}$ over binary losses, and any permutation invariant strategy $\boldsymbol{\omega}$:*

$$R(\boldsymbol{\omega}^{\mathrm{fl}}, \mathcal{P}) \le R(\boldsymbol{\omega}, \mathcal{P}).$$

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

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

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 1. 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 $t = 1, \dots, T$,

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

Since $\boldsymbol{w}_t^{\mathrm{fl}}$ and $\boldsymbol{w}_t$ are permutation invariant, Lemma 1 shows that $\mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t^{\mathrm{fl}} \cdot \boldsymbol{\ell}_t\right] = \overline{\operatorname{loss}}_t(\boldsymbol{w}_t^{\mathrm{fl}}, \mathcal{P})$, and similarly, $\mathbb{E}_{\mathcal{P}}\left[\boldsymbol{w}_t \cdot \boldsymbol{\ell}_t\right] = \overline{\operatorname{loss}}_t(\boldsymbol{w}_t, \mathcal{P})$. Application of Theorem 2 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. We later justify the modification by arguing that the plain FL strategy is not minimax for continuous losses.

4.1 Binarized FL

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 $\boldsymbol{\omega}^{\text{bfl}}$ by:

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

where $\boldsymbol{w}_t^{\text{fl}}(\boldsymbol{b}^{t-1})$ 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 $\boldsymbol{w}_t^{\text{bfl}} \equiv \boldsymbol{w}_t^{\text{fl}}$. 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 $[0, 1]$), $\boldsymbol{w}_t^{\text{bfl}}$ will behave in the same way as $\boldsymbol{w}_t^{\text{fl}}$ 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 $\mathcal{P}^{\text{bin}} = (P_1^{\text{bin}}, \ldots, P_K^{\text{bin}})$, where P_k^{bin} is a distribution with support $\{0, 1\}$ such that:

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

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

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

Proof. 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 $\boldsymbol{p} = (p_1, \ldots, p_K)$. Similarly,

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

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 the binary outcomes resulting from the joint distribution of $\mathcal{P}$ and the distribution of binarization variables:

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

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

$$\begin{aligned} P(b_{t,k}=1) &= \int_{[0,1]} P(b_{t,k}=1|\ell_{t,k})P_k(\ell_{t,k})\mathrm{d}\ell_{t,k} \\ &= \int_{[0,1]} \ell_{t,k}P_k(\ell_{t,k})\mathrm{d}\ell_{t,k} \;=\; p_t = P^{\text{bin}}(\ell_{t,k}=1). \end{aligned} \tag{8}$$

Hence,

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

□

Lemma 3. *For any K-set of distributions $\mathcal{P} = (P_1, \ldots, P_K)$ with support on $\mathcal{X} = [0,1]$,*

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

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

Proof. Lemma 2 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. Denoting $B_{T,k} = \sum_{t=1}^{T} b_{t,k}$, we have:

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

where the first equality is from the fact that for any $b_{t,k}$, the probability (jointly over P_k and the binarization variables) of $b_{t,k} = 1$ is the same as probability of $\ell_{t,k} = 1$ over the distribution P_k^{bin} (see (8) in the proof of Lemma 2), while the inequality follows from Jensen's inequality applied to the concave function $\min(\cdot)$. □

Theorem 3. *Let $\boldsymbol{\omega}^{\text{bfl}} = (\boldsymbol{w}_1^{\text{bfl}}, \ldots, \boldsymbol{w}_T^{\text{bfl}})$ be the binarized FL strategy. Then:*

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

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.

Proof. Lemma 3 states that for any K-set of distributions $\mathcal{P}$, $R(\boldsymbol{\omega}^{\text{bfl}}, \mathcal{P}) \leq R(\boldsymbol{\omega}^{\text{fl}}, \mathcal{P}^{\text{bin}})$. Furthermore, since $\boldsymbol{\omega}^{\text{bfl}}$ is the same as $\boldsymbol{\omega}^{\text{fl}}$ when all the losses are binary, $R(\boldsymbol{\omega}^{\text{bfl}}, \mathcal{P}^{\text{bin}}) = R(\boldsymbol{\omega}^{\text{fl}}, \mathcal{P}^{\text{bin}})$, and hence $R(\boldsymbol{\omega}^{\text{bfl}}, \mathcal{P}) \leq R(\boldsymbol{\omega}^{\text{bfl}}, \mathcal{P}^{\text{bin}})$, 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 1, 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. □

Theorem 3 states that the binarized FL strategy is the minimax prediction strategy when the losses are continuous on $[0, 1]$. Note that the same arguments would hold for any other loss range $[a, b]$, where the binarization on losses would convert continuous losses to the binary losses with values in $\{a, b\}$.

4.2 Vanilla FL is Not Minimax for Continuous Losses

We introduced the binarization procedure to show that the resulting binarized FL strategy is minimax for continuous losses. So far, however, we did not exclude the possibility that the plain FL strategy (without binarization) could also be minimax in the continuous setup. In this section, we prove (by counterexample) that this is not the case, so that the binarization procedure is justified. We will only consider excess risk for simplicity, but one can use similar arguments to show a counterexample for the expected regret and the expected redundancy as well.

The counterexamples proceeds by choosing the simplest non-trivial setup of $K = 2$ experts and $T = 2$ trials. We will first consider the case of binary losses and determine the minimax excess risk. Take two distributions P_1, P_2 on binary

losses and denote $p_1 = P_1(\ell_{t,1} = 1)$ and $p_2 = P_2(\ell_{t,2} = 1)$, assuming (without loss of generality) that $p_1 \leq p_2$. The excess risk of the FL strategy (its expected loss in the second trial minus the expected loss of the first expert) is given by:

$$P(\ell_{1,1} = 0, \ell_{1,2} = 1)p_1 + P(\ell_{1,2} = 0, \ell_{1,1} = 1)p_2 + P(\ell_{1,1} = \ell_{1,2})\frac{p_1 + p_2}{2} \;-\; p_1,$$

which can be rewritten as:

$$\underbrace{p_2(1-p_1)p_1 + p_1(1-p_2)p_2}_{=2p_1p_2 - p_1p_2(p_1+p_2)} + \underbrace{\big(p_1p_2 + (1-p_1)(1-p_2)\big)\frac{p_1+p_2}{2}}_{=p_1p_2(p_1+p_2) - (p_1+p_2)^2 + \frac{p_1+p_2}{2}} \;-\; p_1$$
$$= \frac{p_2 - p_1}{2} - \frac{(p_2-p_1)^2}{2}.$$

Denoting $\delta = p_2 - p_1$, the excess risk can be concisely written as $\frac{\delta}{2} - \frac{\delta^2}{2}$. Maximizing over δ gives $\delta^* = \frac{1}{2}$ and hence the maximum risk of FL on binary losses is equal to $\frac{1}{8}$.

Now, the crucial point to note is that this is also the minimax risk on *continuous* losses. This follows because the binarized FL strategy (which is the minimax strategy on continuous losses) achieves the maximum risk on binary losses (for which it is equivalent to the FL strategy), as follows from the proof of Theorem 3. What remains to be shown is 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 $p_2 = P_2(\ell_{t,2} = 1-\epsilon)$, which means that $\mathbb{E}[\ell_{t,2}] = p_2(1-\epsilon)$. We also assume $\mathbb{E}[\ell_{t,1}] < \mathbb{E}[\ell_{t,2}]$ i.e. expert 1 is the "better" expert, which translates to $p_1 + \epsilon(1-p_1) < p_2(1-\epsilon)$. 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:

$$p_2(1-p_1)p_1 + p_1(1-p_2)p_2 + \underbrace{\big(p_1p_2 + (1-p_1)(1-p_2)\big)p_2}_{\text{ties}} \;-\; p_1 + O(\epsilon).$$

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

5 Conclusions and Open Problem

In this paper, we determined the minimax strategy for the stochastic setting of prediction with expert advice in which each expert generates its losses i.i.d. according to some distribution. Interestingly, the minimaxity is achieved by a single strategy, simultaneously for three considered performance measures: the expected regret, the expected redundancy, and the excess risk. We showed that when the losses are binary, the Follow the Leader algorithm is the minimax

strategy for this game, and furthermore, it also has the smallest expected regret, expected redundancy, and excess risk among all permutation invariant prediction strategies for *every* distribution over the binary losses simultaneously, even among (permutation invariant) strategies which know the distributions of the losses. When the losses are continuous in $[0, 1]$, FL remains minimax only when an additional trick called "loss binarization" is applied, which results in the binarized FL strategy.

Open problem. The setting considered in this paper concerns distributions over loss vectors which are i.i.d. between trials and i.i.d. between experts. It would be interesting to determined the minimax strategy in a more general setting, when the adversary can choose any joint distribution over loss vectors (still i.i.d. between trials, but not necessarily i.i.d. between experts). We did some preliminary computational experiment, which showed that that FL is not minimax in this setting, even when the losses are restricted to be binary.

References

1. Abernethy, J., Agarwal, A., Bartlett, P.L., Rakhlin, A.: A stochastic view of optimal regret through minimax duality. In: COLT (2009)
2. Abernethy, J., Warmuth, M.K., Yellin, J.: When random play is optimal against an adversary. In: COLT. pp. 437–445 (July 2008)
3. Berger, J.O.: Statistical decision theory and Bayesian analysis. Springer (1985)
4. Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D.P., Schapire, R.E., Warmuth, M.K.: How to use expert advice. Journal of the ACM 44(3), 427–485 (1997)
5. Cesa-Bianchi, N., Lugosi, G.: Prediction, learning, and games. Cambridge University Press (2006)
6. van Erven, T., Kotłowski, W., Warmuth, M.K.: Follow the leader with dropout perturbations. In: COLT. pp. 949–974 (2014)
7. Ferguson, T.: Mathematical Statistics: A Decision Theoretic Approach. Academic Press (1967)
8. Grünwald, P.D.: The Minimum Description Length Principle. MIT Press, Cambridge, MA (2007)
9. Koolen, W.M.: Combining Strategies Efficiently: High-quality Decisions from Conflicting Advice. Ph.D. thesis, ILLC, University of Amsterdam (2011)
10. Koolen, W.M., van Erven, T.: Second-order quantile methods for experts and combinatorial games. In: COLT. pp. 1155–1175 (2015)
11. Luo, H., Schapire, R.E.: Achieving all with no parameters: AdaNormalHedge. In: COLT. pp. 1286–1304 (2015)
12. de Rooij, S., van Erven, T., Grünwald, P.D., Koolen, W.M.: Follow the leader if you can, hedge if you must. Journal of Machine Learning Research 15(1), 1281–1316 (2014)
13. Sani, A., Neu, G., Lazaric, A.: Exploiting easy data in online optimization. In: NIPS, pp. 810–818 (2014)