A Derivation of the Minimax Forecaster

In this appendix, we outline how the Minimax Forecaster is derived, as well as its associated guarantees. This outline closely follows the exposition in [10, Chapter 8], to which we refer the reader for some of the technical derivations.

First, we note that the Minimax Forecaster as presented in [10] actually refers to a slightly different setup than ours, where the outcome space is $\mathcal{Y} = \{0, 1\}$ and the prediction space is $\mathcal{P} = [0, 1]$, rather than $\mathcal{Y} = \{-1, +1\}$ and $\mathcal{P} = [-1, +1]$. We will first derive the forecaster for the first setting, and then show how to convert it to the second setting.

Our goal is to find a predictor which minimizes the worst-case regret,

$$\max_{\mathbf{y} \in \{0,1\}^T} \left(L(\mathbf{p}, \mathbf{y}) - \inf_{\mathbf{f} \in \mathcal{F}} L(\mathbf{f}, \mathbf{y}) \right)$$

where $\mathbf{p} = (p_1, \ldots, p_T)$ is the prediction sequence.

For convenience, in the following we sometimes use the notation $\mathbf{y}^t$ to denote a vector in $\{0, 1\}^t$. The idea of the derivation is to work backwards, starting with computing the optimal prediction at the last round T, then deriving the optimal prediction at round $T - 1$ and so on. In the last round T, the first $T - 1$ outcomes $\mathbf{y}^{T-1}$ have been revealed, and we want to find the optimal prediction p_T. Since our goal is to minimize worst-case regret with respect to the absolute loss, we just need to compute p_T which minimizes

$$\max\left\{ L(\mathbf{p}^{T-1}, \mathbf{y}^{T-1}) + p_T - \inf_{\mathbf{f} \in \mathcal{F}} L(\mathbf{f}, \mathbf{y}^{T-1}0) \ , \ L(\mathbf{p}^{T-1}, \mathbf{y}^{T-1}) + (1 - p_T) - \inf_{\mathbf{f} \in \mathcal{F}} L(\mathbf{f}, \mathbf{y}^{T-1}1) \right\} .$$

In our setting, it is not hard to show that $\left| \inf_{\mathbf{f} \in \mathcal{F}} L(\mathbf{f}, \mathbf{y}^{t-1}0) - \inf_{\mathbf{f} \in \mathcal{F}} L(\mathbf{f}, \mathbf{y}^{t-1}1) \right| \leq 1$ (see [10, Lemma 8.1]). Using this, we can compute the optimal p_T to be

$$p_T = \frac{1}{2} \Big(A_T(\mathbf{y}^{T-1}1) - A_T(\mathbf{y}^{T-1}0) + 1 \Big) \tag{5}$$

where $A_T(\mathbf{y}^T) = -\inf_{\mathbf{f} \in \mathcal{F}} L(\mathbf{f}, \mathbf{y}^T)$.

Having determined p_T, we can continue to the previous prediction p_{T-1}. This is equivalent to minimizing

$$\max\left\{ L(\mathbf{p}^{T-2}, \mathbf{y}^{T-2}) + p_{T-1} + A_{T-1}(\mathbf{y}^{T-2}0) \ , \ L(\mathbf{p}^{T-1}, \mathbf{y}^{T-1}) + (1 - p_{T-1}) - \inf_{\mathbf{f} \in \mathcal{F}} L(\mathbf{f}, \mathbf{y}^{T-1}1) \right\}$$

where

$$A_{t-1}(\mathbf{y}^{t-1}) = \min_{p_t \in [0,1]} \max \left\{ p_t - \inf_{\mathbf{f} \in \mathcal{F}} L(\mathbf{f}, \mathbf{y}^{t-1}0) \ , \ (1 - p_t) - \inf_{\mathbf{f} \in \mathcal{F}} L(\mathbf{f}, \mathbf{y}^{t-1}1) \right\} . \tag{6}$$

Note that by plugging in the value of p_T from Eq. (5), we also get the following equivalent formulation for $A_{T-1}(\mathbf{y}^{T-1})$:

$$A_{T-1}(\mathbf{y}^{T-1}) = \frac{1}{2} \Big(A_T(\mathbf{y}^{T-1}0) + A_T(\mathbf{y}^{T-1}1) + 1 \Big).$$

Again, it is possible to show that the optimal value of p_{T-1} is

$$p_{T-1} = \frac{1}{2} \Big(A_{T-1}(\mathbf{y}^{T-2}1) - A_T(\mathbf{y}^{T-2}0) + 1 \Big).$$

Repeating this procedure, one can show that at any round t, the minimax optimal prediction is

$$p_t = \frac{1}{2} \Big(A_t(\mathbf{y}^{t-1}1) - A_t(\mathbf{y}^{t-1}0) + 1 \Big) \tag{7}$$

where A_t is defined recursively as $A_T(\mathbf{y}^T) = -\inf_{\mathbf{f} \in \mathcal{F}} L(\mathbf{f}, \mathbf{y}^T)$ and

$$A_{t-1}(\mathbf{y}^{t-1}) = \frac{1}{2} \Big(A_t(\mathbf{y}^{t-1}0) + A_t(\mathbf{y}^{t-1}1) + 1 \Big). \tag{8}$$

for all t.

At first glance, computing p_t from Eq. (7) might seem tricky, since it requires computing $A_t(\mathbf{y}^t)$ whose recursive expansion in Eq. (8) involves exponentially many terms. Luckily, the recursive expansion has a simple structure, and it is not hard to show that

$$A_t(\mathbf{y}^t) \;=\; \frac{T-t}{2} - \frac{1}{2^T}\sum_{\mathbf{y}\in\{0,1\}^T}\left(\inf_{\mathbf{f}\in\mathcal{F}} L(\mathbf{f},\mathbf{y}^t Y^{T-t})\right) \;=\; \frac{T-t}{2} - \mathbb{E}\Big[\inf_{\mathbf{f}\in\mathcal{F}} L(\mathbf{f},\mathbf{y}^t Y^{T-t})\Big] \quad (9)$$

where Y^{T-t} is a sequence of $T-t$ i.i.d. Bernoulli random variables, which take values in $\{0,1\}$ with equal probability. Plugging this into the formula for the minimax prediction in Eq. (7), we get that[3]

$$p_t = \frac{1}{2}\left(\mathbb{E}\left[\inf_{\mathbf{f}\in\mathcal{F}} L(\mathbf{f},\mathbf{y}^{t-1}0Y^{T-t}) - \inf_{\mathbf{f}\in\mathcal{F}} L(\mathbf{f},\mathbf{y}^{t-1}1Y^{T-t})\right] + 1\right). \quad (10)$$

This prediction rule constitutes the Minimax Forecaster as presented in [10].

After deriving the algorithm, we turn to analyze its regret performance. To do so, we just need to note that A_0 equals the worst-case regret —see the recursive definition at Eq. (6). Using the alternative explicit definition in Eq. (9), we get that the worst-case regret equals

$$\frac{T}{2} - \mathbb{E}\left[\inf_{\mathbf{f}\in\mathcal{F}}\sum_{t=1}^T |f_t - Y_t|\right] \;=\; \mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}}\sum_{t=1}^T\left(\frac{1}{2} - |f_t - Y_t|\right)\right] \;=\; \mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}}\sum_{t=1}^T\left(f_t - \frac{1}{2}\right)\sigma_t\right]$$

where σ_t are i.i.d. Rademacher random variables (taking values of -1 and $+1$ with equal probability). Recalling the definition of Rademacher complexity, Eq. (2), we get that the regret is bounded by the Rademacher complexity of the shifted class, which is obtained from $\mathcal{F}$ by taking every $\mathbf{f}\in\mathcal{F}$ and replacing every coordinate f_t by $f_t - 1/2$.

Finally, it remains to show how to convert the forecaster and analysis above to the setting discussed in this paper, where the outcomes are in $\{-1,+1\}$ rather than $\{0,1\}$ and the predictions are in $[-1,+1]$ rather than $[0,1]$. To do so, consider a learning problem in this new setting, with some class $\mathcal{F}$. For any vector $\mathbf{y}$, define $\widetilde{\mathbf{y}}$ to be the shifted vector $(\mathbf{y}+\mathbf{1})/2$, where $\mathbf{1}=(1,\ldots,1)$ is the all-ones vector. Also, define $\widetilde{\mathcal{F}}$ to be the shifted class $\widetilde{\mathcal{F}} = \{(\mathbf{f}+\mathbf{1})/2 \,:\, \mathbf{f}\in\mathcal{F}\}$. It is easily seen that $L(\mathbf{f},\mathbf{y}) = 2L(\widetilde{\mathbf{f}},\widetilde{\mathbf{y}})$ for any $\mathbf{f},\mathbf{y}$. As a result, if we look at the prediction p_t given by our forecaster in Eq. (3), then $\widetilde{p_t} = (p_t+1)/2$ is the minimax optimal prediction given by Eq. (10) with respect to the class $\widetilde{\mathcal{F}}$ and the outcomes $\widetilde{\mathbf{y}}^T$. So our analysis above applies, and we get that

$$\begin{aligned}\max_{\mathbf{y}\in\{-1,+1\}^T}\left(L(\mathbf{p},\mathbf{y}) - \inf_{\mathbf{f}\in\mathcal{F}} L(\mathbf{f},\mathbf{y})\right) &= \max_{\widetilde{\mathbf{y}}\in[0,1]^T} 2\left(L(\widetilde{\mathbf{p}},\widetilde{\mathbf{y}}) - \inf_{\widetilde{\mathbf{f}}\in\widetilde{\mathcal{F}}} L(\widetilde{\mathbf{f}},\widetilde{\mathbf{y}})\right)\\ &= 2\mathbb{E}\left[\sup_{\widetilde{\mathbf{f}}\in\widetilde{\mathcal{F}}}\sum_{t=1}^T\left(\widetilde{f}_t - \frac{1}{2}\right)\sigma_t\right]\\ &= \mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}}\sum_{t=1}^T \sigma_t f_t\right]\end{aligned}$$

which is exactly the Rademacher complexity of the class $\mathcal{F}$.

B Proof of Thm. 3

Let $Y(t)$ denote the set of Bernoulli random variables chosen at round t. Let $\mathbb{E}_{z_t}$ denote expectation with respect to z_t, conditioned on $z_1, Y(1), \ldots, z_{t-1}, Y(t-1)$ as well as $Y(t)$. Let $\mathbb{E}_{Y(t)}$ denote the expectation with respect to the random drawing of $Y(t)$, conditioned on $z_1, Y(1), \ldots, z_{t-1}, Y(t-1)$.

We will need two simple observations. First, by convexity of the loss function, we have that for any p_t, f_t, y_t, $\ell(p_t,y_t) - \ell(f_t,y_t) \le (p_t - f_t)\,\partial_{p_t}\ell(p_t,y_t)$. Second, by definition of r_t and

[3]This fact appears in an implicit form in [9] —see also [10, Exercise 8.4].

z_t, we have that for any fixed p_t, f_t,

$$\begin{aligned}
\frac{1}{\rho b}(p_t - f_t)\partial_{p_t}\ell(p_t, y_t) &= \frac{1}{b}(p_t - f_t)(1 - 2r_t) \\
&= \frac{1}{b}r_t(f_t - p_t) + \frac{1}{b}(1 - r_t)(p_t - f_t) \\
&= r_t(\widetilde{f}_t - \widetilde{p}_t) + (1 - r_t)(\widetilde{p}_t - \widetilde{f}_t) \\
&= r_t\left((1 - \widetilde{p}_t) - \left(1 - \widetilde{f}_t\right)\right) + (1 - r_t)\left((\widetilde{p}_t + 1) - \left(\widetilde{f}_t + 1\right)\right) \\
&= \mathbb{E}_{z_t}\left[|\widetilde{p}_t - z_t| - \left|\widetilde{f}_t - z_t\right|\right] .
\end{aligned}$$

The last transition uses the fact that $\widetilde{p}_t, \widetilde{f}_t \in [-1, +1]$. By these two observations, we have

$$\sum_{t=1}^{T}\ell(p_t, y_t) - L(\mathbf{f}, \mathbf{y}) \le \sum_{t=1}^{T}(p_t - f_t)\,\partial_{p_t}\ell(p_t, y_t) = \rho\, b\,\sum_{t=1}^{T}\mathbb{E}_{z_t}\left[|\widetilde{p}_t - z_t| - \left|\widetilde{f}_t - z_t\right|\right] . \tag{11}$$

Now, note that $|\widetilde{p}_t - z_t| - |\widetilde{f}_t - z_t| - \mathbb{E}_{z_t}\left[|\widetilde{p}_t - z_t| - |\widetilde{f}_t - z_t|\right]$ for $t = 1, \ldots, T$ is a martingale difference sequence: for any values of $z_1, Y(1), \ldots, z_{t-1}, Y(t-1), Y(t)$ (which fixes $\widetilde{p}_t$), the conditional expectation of this expression over z_t is zero. Using Azuma's inequality, we can upper bound Eq. (11) with probability at least $1 - \delta/2$ by

$$\rho\, b\,\sum_{t=1}^{T}\left(|\widetilde{p}_t - z_t| - |\widetilde{f}_t - z_t|\right) + \rho\, b\sqrt{8T\ln(2/\delta)}. \tag{12}$$

The next step is to relate Eq. (12) to $\rho\, b\sum_{t=1}^{T}\left(\left|\mathbb{E}_{Y(t)}[\widetilde{p}_t] - z_t\right| - |\widetilde{f}_t - z_t|\right)$. It might be tempting to appeal to Azuma's inequality again. Unfortunately, there is no martingale difference sequence here, since z_t is itself a random variable whose distribution is influenced by $Y(t)$. Thus, we need to turn to coarser methods. Eq. (12) can be upper bounded by

$$\rho\, b\,\sum_{t=1}^{T}\left(\left|\mathbb{E}_{Y(t)}[\widetilde{p}_t] - z_t\right| - |\widetilde{f}_t - z_t|\right) + \rho\, b\,\sum_{t=1}^{T}\left|\widetilde{p}_t - \mathbb{E}_{Y(t)}[\widetilde{p}_t]\right| + \rho\, b\sqrt{8T\ln(2/\delta)}. \tag{13}$$

Recall that $\widetilde{p}_t$ is an average over ηT i.i.d. random variables, with expectation $\mathbb{E}_{Y(t)}[\widetilde{p}_t]$. By Hoeffding's inequality, this implies that for any $t = 1, \ldots, T$, with probability at least $1 - \delta/2T$ over the choice of $Y(t)$, $\left|\widetilde{p}_t - \mathbb{E}_{Y(t)}[\widetilde{p}_t]\right| \le \sqrt{2\ln(2T/\delta)/(\eta T)}$. By a union bound, it follows that with probability at least $1 - \delta/2$ over the choice of $Y(1), \ldots, Y(T)$,

$$\sum_{t=1}^{T}\left|\widetilde{p}_t - \mathbb{E}_{Y(t)}[\widetilde{p}_t]\right| \le \sqrt{\frac{2T\ln(2T/\delta)}{\eta}} .$$

Combining this with Eq. (13), we get that with probability at least $1 - \delta$,

$$\rho\, b\sum_{t=1}^{T}\left(\left|\mathbb{E}_{Y(t)}[\widetilde{p}_t] - z_t\right| - |\widetilde{f}_t - z_t|\right) + \rho\, b\sqrt{\frac{2T\ln(2T/\delta)}{\eta}} + \rho\, b\sqrt{8T\ln(2/\delta)} . \tag{14}$$

Finally, by definition of $\widetilde{p}_t = p_t/b$, we have

$$\mathbb{E}_{Y(t)}[\widetilde{p}_t] = \mathbb{E}_{Y(t)}\left[\inf_{\mathbf{f}\in\mathcal{F}} L\left(\widetilde{\mathbf{f}}, z_1 \ldots z_{t-1}\,(-1)\,Y_{t+1} \ldots Y_T\right) - \inf_{\mathbf{f}\in\mathcal{F}} L\left(\widetilde{\mathbf{f}}, z_1 \ldots z_{t-1}\,1\,Y_{t+1} \ldots Y_T\right)\right] .$$

This is exactly the Minimax Forecaster's prediction at round t, with respect to the sequence of outcomes $z_1, \ldots, z_{t-1} \in \{-1, +1\}$, and the class $\widetilde{\mathcal{F}} := \left\{\widetilde{\mathbf{f}} : \mathbf{f} \in \mathcal{F}\right\} \subseteq [-1, 1]^T$. Therefore, using Thm. 1, we can upper bound Eq. (14) by

$$\rho\, b\,\mathcal{R}_T(\widetilde{\mathcal{F}}) + \rho\, b\sqrt{\frac{2T\ln(2T/\delta)}{\eta}} + \rho\, b\sqrt{8T\ln(2/\delta)} .$$

By definition of $\widetilde{\mathcal{F}}$ and Rademacher complexity, it is straightforward to verify that $\mathcal{R}_T(\widetilde{\mathcal{F}}) = \frac{1}{b}\mathcal{R}_T(\mathcal{F})$. Using that to rewrite the bound, and slightly simplifying for readability, the result stated in the theorem follows.

C Proof of Lemma 1

The proof assumes that the infimum and supremum of certain functions over $\mathcal{Y}, \mathcal{F}$ are attainable. If not, the proof can be easily adapted by finding attainable values which are ϵ-close to the infimum or supremum, and then taking $\epsilon \to 0$.

For the purpose of contradiction, suppose there exists a strategy for the adversary and a round $r \leq T$ such that at the end of round r, the forecaster suffers a regret $G' > G$ with probability larger than δ. Consider the following modified strategy for the adversary: the adversary plays according to the aforementioned strategy until round r. It then computes

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{t=1}^{r} \ell(f_t, y_t) \ .$$

At all subsequent rounds $t = r+1, r+2, \ldots, T$, the adversary chooses

$$y_t^* = \operatorname*{argmax}_{y \in \mathcal{Y}} \inf_{p \in \mathcal{P}} \big(\ell(p, y) - \ell(f_t^*, y)\big) \ .$$

By the assumption on the loss function,

$$\ell(p_t, y_t^*) - \ell(f_t^*, y_t^*) \geq \inf_{p \in \mathcal{P}} \big(\ell(p, y_t^*) - \ell(f_t^*, y_t^*)\big) = \sup_{y \in \mathcal{Y}} \inf_{p \in \mathcal{P}} \big(\ell(p, y) - \ell(f_t^*, y)\big) \geq 0 \ .$$

Thus, the regret over all T rounds, with respect to f^*, is

$$\sum_{t=1}^{r} \big(\ell(p_t, y_t) - \ell(f_t^*, y_t)\big) + \sum_{t=r+1}^{T} \big(\ell(p_t, y_t^*) - \ell(f_t^*, y_t^*)\big) \geq \sum_{t=1}^{r} \ell(p_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{r} \ell(f_t, y_t) + 0$$

which is at least G' with probability larger than δ. On the other hand, we know that the learner's regret is at most most G with probability at least $1 - \delta$. Thus we have a contradiction and the proof is concluded.