DELORO Yonatan.
Prediction of individual sequences.
Homework 1.

Part I - Theory - Sleeping experts.

1 - The prod algorithm

1) a) let $\varphi : [-\frac{1}{2}, +\infty) \longrightarrow \mathbb{R}$

$$x \longmapsto \log(1+x) - x + x^2$$

$$\varphi'(x) = \frac{1}{1+x} - 1 + 2x = \frac{1}{1+x}\left[1 - 1 - x + 2x + 2x^2\right]$$

$$= \frac{2x^2 + x}{1+x}$$

thus, $\varphi'(x)$ has the same sign as $2x^2 + x = (2x+1)x$ since $1+x \geqslant 0$ for any $x \geqslant -\frac{1}{2}$. Thus:

x	$-\frac{1}{2}$		0		$+\infty$
$\varphi'(x)$		$-$		$+$	
$\varphi(x)$		↘	$\varphi(0)$	↗	

$\varphi(0) = \log(1) = 0$. Thus: $\varphi(x) \geqslant 0$. $\forall x \geqslant -\frac{1}{2}$.

We just showed that:

$$\forall x \geqslant -\frac{1}{2}, \ \log(1+x) \geqslant x - x^2$$

---

1) b) let $k \in \mathcal{X}$

- We just can write : $W_{T+1} = \sum_{i=1}^{K} w_{T+1}(i) \geqslant w_{T+1}(k)$

Indeed, for any $i \in [\![1, k]\!]$, $w_{T+1}(i) \geq 0$. Let us prove it by recursion:

- $w_1(i) = 1 \geq 0$
- Let's assume $w_t(i) \geq 0$ for a given $t \geq 1$.

$$w_{t+1}(i) = w_t(i)\left(1 + \eta(i)\left(p_t \cdot l_t - l_t(i)\right)\right)$$

→ $p_t \cdot l_t = \sum_{j \in X} \underbrace{p_t(i)}_{\geq 0} \underbrace{l_t(i)}_{\geq 0} \geq 0$.

And $l_t(i) \leq 1$, thus $p_t \cdot l_t - l_t(i) \geq -1$.

→ as $\eta(i) \geq 0$, $\eta(i)\left(p_t \cdot l_t - l_t(i)\right) \geq -\eta(i) \geq -\frac{1}{2}$ $\left(\eta(i) \leq \frac{1}{2}\right)$

thus, $1 + \eta(i)\left(p_t \cdot l_t - l_t(i)\right) \geq \frac{1}{2} \geq 0$

this shows $w_{t+1}(i) \geq 0$.

• Then, composing by the non-decreasing "log" function, we get:

$$\begin{aligned}\log(W_{T+1}) &\geq \log\left(w_{T+1}(k)\right)\\ &= \sum_{s=1}^{T} \log\left(1 + \eta(k)\left(p_s \cdot l_s - l_s(k)\right)\right)\end{aligned}$$

We showed above that $\eta(k)\left(p_s \cdot l_s - l_s(k)\right) \geq -\frac{1}{2}$. Thus, by applying question 1)a), we get:

$$\begin{aligned}\log(W_{T+1}) &\geq \sum_{s=1}^{T} \eta(k)\left(p_s \cdot l_s - l_s(k)\right) - \sum_{s=1}^{T} \eta(k)^2 \left(p_s \cdot l_s - l_s(k)\right)^2\\ &= \eta(k) \sum_{s=1}^{T} \left(p_s \cdot l_s - l_s(k)\right) - \eta(k)^2 \sum_{s=1}^{T} \left(p_s \cdot l_s - l_s(k)\right)^2\end{aligned}$$

for any $k \in X$

---

1)c) * Let $t \geq 1$.

$$W_{t+1} = \sum_{k=1}^{K} w_{T+1}(k)$$

But, for any $k \in [\![1, K]\!]$, $w_{t+1}(k) = w_t(k)\left(1 + \eta(k)\left(p_t \cdot l_t - l_t(k)\right)\right)$

Thus, $W_{t+1} = \sum_{k=1}^{K} w_t(k) + \sum_{k=1}^{K} w_t(k)\eta(k)\left(p_t\cdot l_t - l_t(k)\right)$

$$= W_t + \sum_{k=1}^{K} p_t(k)\left(\sum_{j=1}^{K}\eta(j)\,w_t(j)\right)\left(p_t\cdot l_t - l_t(k)\right)$$

$$= W_t + \left(\sum_{j=1}^{K}\eta(j)w_t(j)\right)\underbrace{\left(p_t\cdot l_t \underbrace{\left(\sum_{k=1}^{K} p_t(k)\right)}_{=1} - \underbrace{\sum_{k=1}^{K} p_t(k)l_t(k)}_{=p_t\cdot l_t}\right)}_{=0}$$

Which shows that, for any $t \geq 1$,

$$\underline{W_{t+1} = W_t}$$

- Therefore, $\underline{\log(W_{T+1}) = \log(W_1) = \log\left(\sum_{k=1}^{K} w_1(k)\right) = \log(K)}$

1) d) Thus, we have, for any $k \in [\![1, K]\!]$,

$$\log(K) \geqslant \eta(k)\left(\sum_{t=1}^{T} p_t\cdot l_t - l_t(k)\right) - \eta(k)^2\left(\sum_{t=1}^{T}\left(p_t\cdot l_t - l_t(k)\right)^2\right)$$

$$\Leftrightarrow \log(K) \geqslant \eta(k)\,R_T(k) - \eta(k)^2\sum_{t=1}^{T}\left(p_t\cdot l_t - l_t(k)\right)^2$$

The left term of the inequality is maximal when:

$$\eta(k) = \frac{R_T(k)}{2\sum_{t=1}^{T}\left(p_t\cdot l_t - l_t(k)\right)^2}.$$

Thus, when we choose such η, we get:

$$\log(K) \geqslant R_T(k)^2\left(\frac{1}{2\sum_{t=1}^{T}\left(p_t\cdot l_t - l_t(k)\right)^2} - \frac{1}{4\sum_{t=1}^{T}\left(p_t\cdot l_t - l_t(k)\right)^2}\right)$$

Which leads to:

$$\log(K) \geqslant \frac{1}{4}\,\frac{R_T(k)^2}{\sum_{t=1}^{T}\left(p_t\cdot l_t - l_t(k)\right)^2}$$

Or in other words:

$$\underline{R_T(k) \leqslant 2\sqrt{\log(K)\sum_{t=1}^{T}\left(p_t\cdot l_t - l_t(k)\right)^2}}$$

<u>for any $k \in [\![1, K]\!]$</u>

2- Sleeping experts.

a) let $t \geq 1$ and $k \in \mathcal{X}$

- First of all,

$$\tilde{p}_t \cdot \tilde{l}_t = \sum_{k\in A_t} \tilde{p}_t(k)\,\tilde{l}_t(k) + \sum_{k\notin A_t} \tilde{p}_t(k)\,\tilde{l}_t(k)$$

$$= \sum_{k\in A_t} \tilde{p}_t(k)\, l_t(k) + \sum_{k\notin A_t} \tilde{p}_t(k)\,(p_t\cdot l_t)$$

But, by definition, $p_t(k) = \dfrac{\tilde{p}_t(k)\,\mathbb{1}_{k\in A_t}}{\sum_{j\in A_t}\tilde{p}_t(j)}$

Thus, $p_t\cdot l_t = \sum_{k\in A_t} p_t(k)\, l_t(k) = \left(\sum_{k\in A_t}\tilde{p}_t(k)\, l_t(k)\right)\left(\dfrac{1}{\sum_{k\in A_t}\tilde{p}_t(k)}\right)$

Therefore, we can rewrite :

$$\tilde{p}_t\cdot\tilde{l}_t = \left(\sum_{k\in A_t}\tilde{p}_t(k)\right)(p_t\cdot l_t) + (p_t\cdot l_t)\sum_{k\notin A_t}\tilde{p}_t(k)$$

$$= (p_t\cdot l_t)\underbrace{\left(\sum_{k\in\mathcal{X}}\tilde{p}_t(k)\right)}_{=1}$$

$$= p_t\cdot l_t$$

- Thus, if $k \in A_t$,

$$\tilde{p}_t\cdot\tilde{l}_t - \tilde{l}_t(k) = p_t\cdot l_t - l_t(k)$$

and if $k \notin A_t$,

$$\tilde{p}_t\cdot\tilde{l}_t - \tilde{l}_t(k) = p_t\cdot l_t - p_t\cdot l_t = 0 .$$

- Therefore, we can conclude that :

$$\tilde{p}_t\cdot\tilde{l}_t - \tilde{l}_t(k) = (p_t\cdot l_t - l_t(k))\,\mathbb{1}_{k\in A_t}$$

for any $k \in [\![1,K]\!]$ and $t \geq 1$

2) b) Hence,

$$R_T(k) = \sum_{t=1}^{T} (p_t \cdot l_t - l_t(k)) \mathbb{1}_{\{k \in A_t\}}$$

$$= \sum_{t=1}^{T} (\tilde{p}_t \cdot \tilde{l}_t - \tilde{l}_t(k)) \quad \text{with 2) a)}$$

$$\leq 2\sqrt{(\log K)\sum_{t=1}^{T}(\tilde{p}_t \cdot \tilde{l}_t - \tilde{l}_t(k))^2} \quad \text{with 1) d)}$$

$$= 2\sqrt{\log K}\sqrt{\sum_{t=1}^{T}(p_t \cdot l_t - l_t(k))^2 \mathbb{1}_{k \in A_t}^2}$$

But $\forall t \in [\![1, T]\!]$,

$$p_t \cdot l_t = \sum_{k \in A_t} \underbrace{\underbrace{p_t(k)}_{\geq 0} \underbrace{l_t(k)}_{\leq 1}}_{\leq p_t(k)} \leq \sum_{k \in A_t} p_t(k) = 1$$

and $l_t(k) \geq 0$, thus $p_t \cdot l_t - l_t(k) \leq 1$

We also showed in 1) b) that $p_t \cdot l_t - l_t(k) \geq -1$.

Thus, $(p_t \cdot l_t - l_t(k))^2 \leq 1$. Therefore, this allows to write:

$$R_T(k) \leq 2\sqrt{\log K}\sqrt{\sum_{t=1}^{T} \mathbb{1}_{k \in A_t}}$$

$$\Leftrightarrow \quad R_T(k) \leq 2\sqrt{\log(K) T_k}$$

with T_k the number of times arm k is active $\left(T_k = \sum_{t=1}^{T} \mathbb{1}_{k \in A_t}\right)$

→ Part II

Part 2. Experiments – predict votes of surveys

3. The loss l(ŷt , yt) = (1 − ŷt)yt + ŷt (1 − yt) is very adequate. It is equal to the predicted probability of player 1's defeat (1-ŷt) if player 1 wins (yt=1), and to the predicted probability of player 1's win (ŷt) if player 1 looses (yt=0). The chosen loss is also linear, hence convex.

4. See the Python code (Jupyter Notebook). I implemented the Exponentially Weighted Average Forecaster (EWA), the Online Gradient Descent (OGD) and the Prod Forecaster (PROD).

Note : I first answer to question 5 on the basis of the experiments carried on the "politicians" dataset. I give the results obtained on the "ideas" dataset in last page.

5.

a) Below (Figure 1), are plotted the final cumulated loss at time T=15,000 for each forecaster against the choice of the eta parameter (for the PROD algorithm, I looked for a common value eta for each expert). I plot here the results for a fine-grained grid of eta (range was chosen after a first search based on a log-scale coarse grid).

It seemed that (i) the OGD final loss varied in the most significant way around the optimal eta we found (0.002), (ii) the EWA loss varied quite smoothly below its optimal value (0.15) but quite significantly above it, and (iii) the PROD algorithm finally led to the largest region of interesting eta : final cumulated loss did not vary so much around optimal eta=0.3 (below or above), and that's why I would give preference to the PROD algorithm.

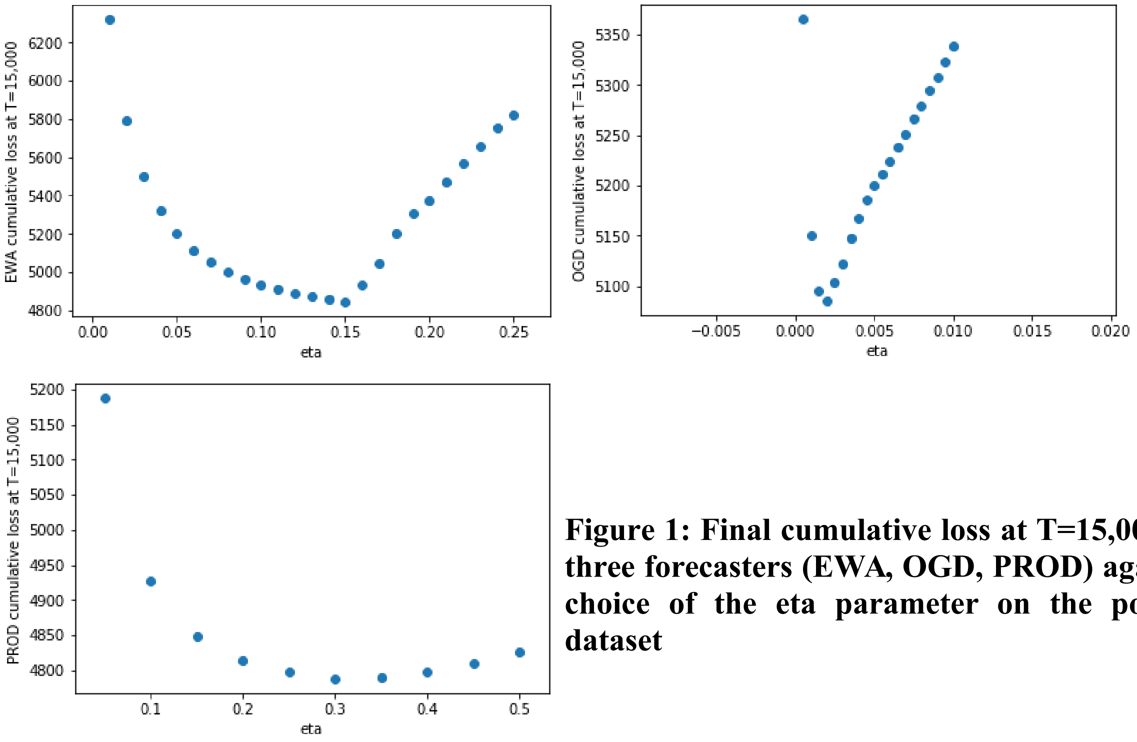

Figure 1: Final cumulative loss at T=15,000 of the three forecasters (EWA, OGD, PROD) against the choice of the eta parameter on the politicians dataset

Note : In addition, we see that the optimal eta found for EWA (0.15) differs from the advised theoretic value sqrt(log(K)/(K*T))= 0.002 (K= 78 experts and T=15,000 rounds), which is not completely surprising as the theoretic value is such that it minimizes an upper bound of the regret. Similarly, the theoretic eta value for OGD is D/(G*sqrt(T))=0.008 (the difference between two predicted probabilities is lower than D=1, and the loss gradient is lower than G=1 too) while the best obtained one is 0.002.

b) We see (Figure 2) that, fortunately, the three forecasters significantly beat the random predictions if we choose for them the best empirical etas (cf. 5)a)). The average expected loss of the EWA, OGD, and PROD forecasters with empirically optimal eta converged respectively towards 0.323, 0.339 and 0.319.

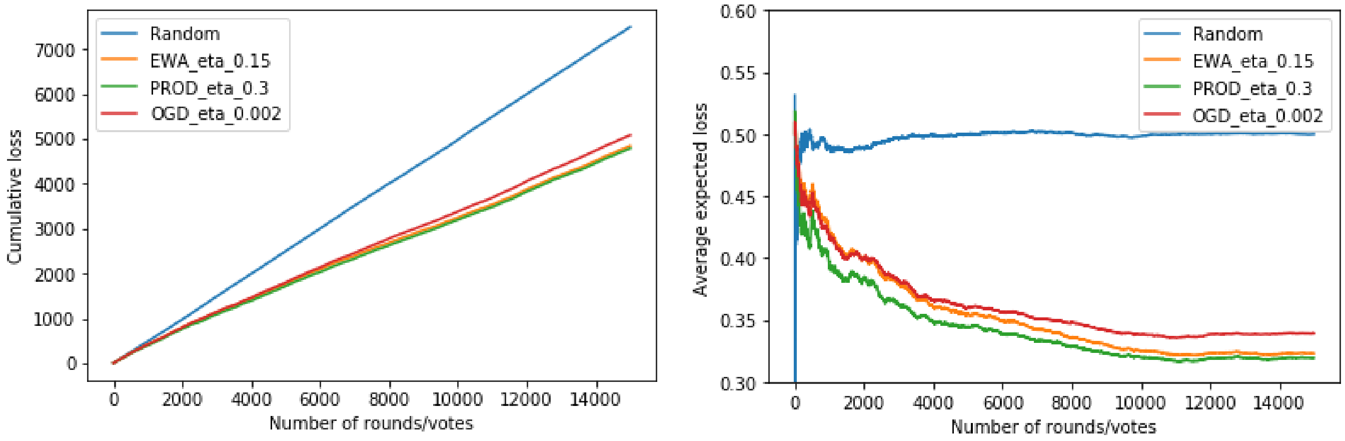

Figure 2: Evolution of cumulative expected loss, and average expected loss throughout the rounds of the three forecasters (EWA, OGD, PROD) with empirically optimal eta parameters, on the politicians dataset.

c) First we observe (Figure 3), as expected, that the average true loss vary more than the expected one in the first rounds (at the beginning, we can expect that the expected probability of player 1's win is close from 0.5 since we initialized weights of experts to a constant (1/K), and sampling from a Bernoulli of parameter 0.5 to vote leads to loss varying a lot from one round to the other).

The average true loss of the EWA and PROD forecasters with empirically optimal eta converged towards 0.318, while the average true loss of the OGD forecaster converged towards 0.339, which are very similar to the average expected loss convergence values.

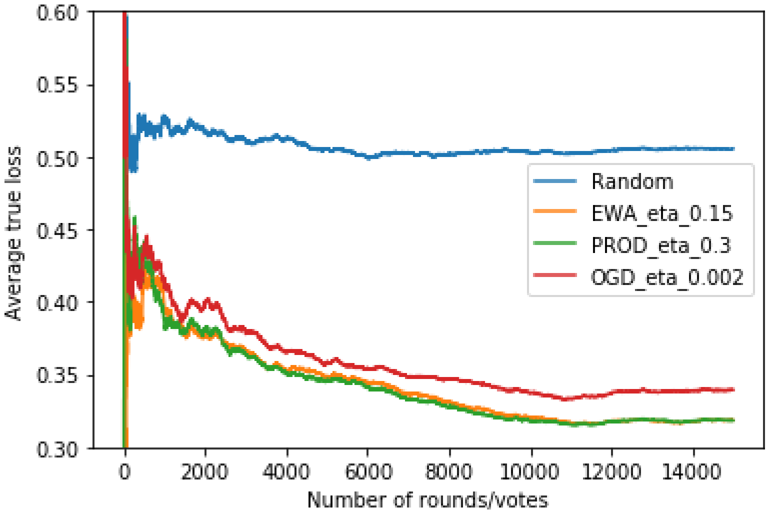

Figure 3: Evolution of average true loss throughout the rounds of the three forecasters (EWA, OGD, PROD) with empirically optimal eta parameters, on the politicians dataset.

Note : Let's also note that if we chose instead the theoretic eta values (expressions given at the bottom of page 6/8), the average expected (resp. true) loss converged to 0.353 (resp. 0.352) for OGD, and to only 0.476 (resp. 0.48) for EWA, but they both still beat the random predictions. *[Setting boolean "empirical_etas" to "False" in the code, one can visualize the curbs and convergence values for theoretic etas for EWA/OGD)]*

Results on the ideas dataset :

5)a) As the experiments were more time consuming on this dataset (261 decisions, 522 experts), I opted only for a coarse log-scale grid search for the eta parameter. On this dataset, we again see that the PROD Forecaster leads to a larger range of good etas than the EWA/OGD Forecasters.

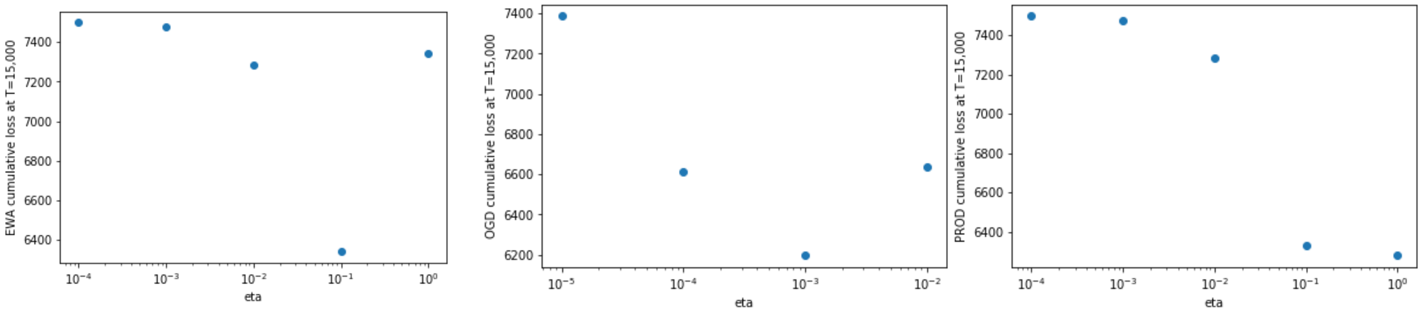

Figure 4: Final cumulative loss at T=15,000 of the three forecasters (EWA, OGD, PROD) against the choice of the eta parameter on the ideas dataset.

5)b) The average expected loss of the EWA, OGD and PROD forecaster converged respectively towards 0.423, 0.413 and 0.419 for the best empirical etas.

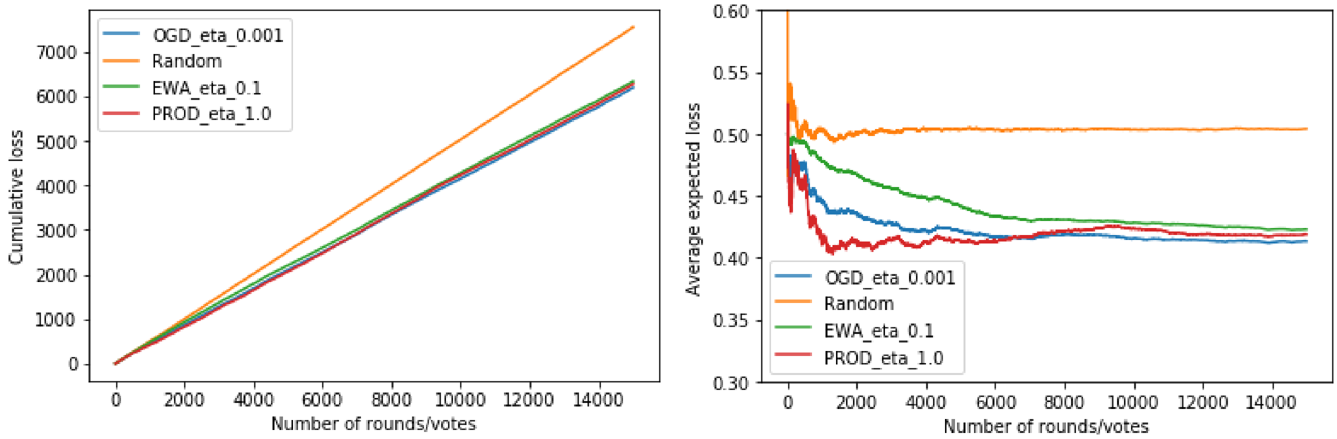

Figure 5: Evolution of cumulative expected loss, and average expected loss throughout the rounds of the three forecasters (EWA, OGD, PROD) with empirically optimal eta parameters, on the ideas dataset.

5)c) The average true loss of the EWA, OGD and PROD forecaster converged respectively towards 0.427, 0.415 and 0.416. On this dataset, the EWA performed worse than the OGD/PROD Forecasters.

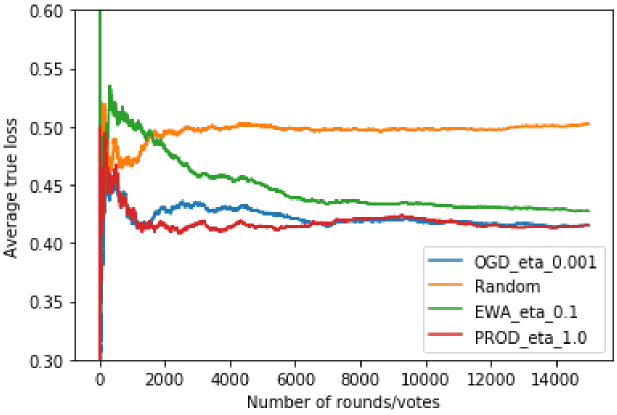

Figure 6: Evolution of average true loss throughout the rounds of the three forecasters (EWA, OGD, PROD) with empirically optimal eta parameters, on the ideas dataset.