Supplementary Material

Mustafa A. Kocak, *Student Member, IEEE,* David Ramirez, *Member, IEEE,* Elza Erkip, *Fellow, IEEE,* and Dennis E. Shasha, *Fellow, ACM*

This supplementary document is organized as follows. In Section A we present the proof of the equivalence lemma (Lemma 4.1 from the main text). The extension of the numerical experiments from the main text is given in Section B.

A Proof of The Equivalence Lemma (Lemma 4.1)

We first repeat the statement of the lemma for the sake completeness.

Lemma A.1. *(Equivalence lemma) Suppose we have a base predictor P. Consider an ensemble of 2^T experts, $\mathcal{P} = \{P_0, P_1, \ldots, P_{2^T-1}\}$, defined as follows:*

- *Denote the t^{th} bit of the binary expansion of integer i with $b_{i,t}$, and define the notation $\bar{x} = 1 - x$.*
- *Fix the predictions and the losses of each expert P_i as follows:*

$$\hat{y}_{P_i,t} = \begin{cases} \varnothing & b_{i,t} = 0 \\ \hat{y}_{P,t} & b_{i,t} = 1 \end{cases} \quad \text{and } l_{P_i,t} = \begin{cases} \epsilon & b_{i,t} = 0 \\ l_{P,t} & b_{i,t} = 1 \end{cases}.$$

- *Set the initial weights for each expert as*

$$w_{P_i,1} = w_{P,1}^{b_{i,1}} w_{D,1}^{\bar{b}_{i,1}} \cdots \prod_{t=1}^{T-1} \alpha_t^{\bar{b}_{i,t} b_{i,t+1}} \bar{\alpha}_t^{\bar{b}_{i,t} \bar{b}_{i,t+1}} \beta_t^{b_{i,t} b_{i,t+1}} \bar{\beta}_t^{b_{i,t} \bar{b}_{i,t+1}}.$$

Then the EWAF algorithm (Alg. 1) using the expert ensemble $\mathcal{P}$ with the learning rate η is equivalent to Adaptive SafePredict (Alg. 4) using the base predictor P, in terms of the prediction probability

$$w_{P,t} = \sum_{i:b_{i,t}=1} w_{P_i,t}.$$

To prove the lemma, we first need the following auxiliary proposition.

Proposition A.2. *The total weight at time t over the experts that follow P both at time t and $t+1$ can be represented as*

$$\sum_{i:b_{i,t}b_{i,t+1}=1} w_{P_i,t} = \beta_t \sum_{i:b_{i,t}=1} w_{P_i,t}. \tag{1}$$

Similarly, the sum of the weights over the experts that follow the dummy at t but follow P at $t+1$ can be written as

$$\sum_{i:\bar{b}_{i,t}b_{i,t+1}=1} w_{P_i,t} = \alpha_t \sum_{i:\bar{b}_{i,t}=1} w_{P_i,t}. \tag{2}$$

Proof of Proposition A.2. First fix the time index t and define a conjugate predictor $P_{i'}$ for each expert P_i in the ensemble such that

$$b_{i',\tau} = \begin{cases} \bar{b}_{i,\tau} & \tau = t+1 \\ b_{i,\tau} & \text{otherwise} \end{cases}$$

and note that

$$\frac{w_{P_i,t}}{w_{P_i,1}} = \frac{w_{P'_i,t}}{w_{P'_i,1}} \tag{3}$$

since both P_i and $P_{i'}$ suffers the same losses in the first t rounds.

Finally eq. (1) follows from,

$$\frac{\sum_{i:b_{i,t}b_{i,t+1}=1} w_{P_i,t}}{\sum_{i:b_{i,t}=1} w_{P_i,t}} = \frac{\sum_{i:b_{i,t}b_{i,t+1}=1} w_{P_i,t}}{\sum_{i:b_{i,t}b_{i,t+1}=1} w_{P_i,t} + w_{P'_i,t}} \tag{4}$$

$$= \frac{\sum_{i:b_{i,t}b_{i,t+1}=1} w_{P_i,1}}{\sum_{i:b_{i,t}b_{i,t+1}=1} w_{P_i,1} + w_{P'_i,1}} \tag{5}$$

$$= \beta_t \tag{6}$$

where (4) is from the definition of the conjugate predictor $P_{i'}$, (5) follows from eq. (3), and (6) by plugging in the initial weights and straightforward algebra. Eq. (2) also follow from a similar argument for $\bar{b}_{i,t} = 1$. □

Next, armed with this proposition, the proof of the lemma simply follows by an induction argument over t.

Proof of Lemma A.1. The base case, $t = 1$, follows from a sequence of straightforward algebraic manipulations

$$w_{P,1} = \sum_{i:b_{i,1}=1} w_{P_i,1}.$$

Next, assume the induction hypothesis holds for t, i.e.

$$w_{P,t} = \sum_{i:b_{i,t}=1} w_{P_i,t}. \tag{7}$$

- *M. A. Kocak, D. Ramirez and E. Erkip are with the Department of Electrical and Computer Engineering, NYU Tandon School of Engineering, Brooklyn, NY, 11201.*
  E-mail: {kocak, dar550, elza}@nyu.edu
- *D. E. Shasha is with Courant Institute of Mathematical Sciences New York University, New York, NY, 10012.*
  E-mail: shasha@courant.nyu.edu

Finally the induction step is executed as follows:

$$\sum_{i:b_{i,t+1}=1} w_{P_i,t+1} = \sum_{i:b_{i,t}b_{i,t+1}=1} w_{P_i,t+1} + \sum_{i:\bar{b}_{i,t}b_{i,t+1}=1} w_{P_i,t+1}$$

$$= \frac{\sum_{i:b_{i,t}b_{i,t+1}=1} w_{P_i,t}e^{-\eta l_{P,t}} + \sum_{i:\bar{b}_{i,t-1}b_{i,t}=1} w_{P_i,t-1}e^{-\eta\epsilon}}{\sum_i w_{P_i,t}e^{-\eta l_{P_i,t}}} \tag{8}$$

$$= \frac{\sum_{i:b_{i,t}b_{i,t+1}=1} w_{P_i,t}e^{-\eta l_{P,t}} + \sum_{i:\bar{b}_{i,t-1}b_{i,t}=1} w_{P_i,t-1}e^{-\eta\epsilon}}{\sum_{i:b_{i,t}=1} w_{P_i,t}e^{-\eta l_{P,t}} + \sum_{i:\bar{b}_{i,t}=1} w_{P_i,t}e^{-\eta\epsilon}} \tag{9}$$

$$= \frac{\beta_t w_{P,t}e^{-\eta l_{P,t}} + \alpha_t w_{D,t}e^{-\eta\epsilon}}{w_{P,t}e^{-\eta l_{P,t}} + w_{D,t}e^{-\eta\epsilon}} \tag{10}$$

$$= \alpha_t + (\beta_t - \alpha_t)\frac{w_{P,t}e^{-\eta l_{P,t}}}{w_{P,t}e^{-\eta l_{P,t}} + w_{D,t}e^{-\eta\epsilon}} \tag{11}$$

$$= w_{P,t+1}. \tag{12}$$

Note (8) follows from the EWAF update rule, (9) follows from the the choice of $l_{P_i,t}$ (see the body of the lemma), (10) follows from Proposition A.2 and the induction hypothesis eq. (7), (11) follows from simple algebra, and finally (12) follows from the update rule for Adaptive SafePredict, i.e.

$$w_{P,t+1} \quad = \quad \alpha_t + (\beta_t - \alpha_t)\frac{w_{P,t}e^{-\eta l_{P,t}}}{w_{P,t}e^{-\eta l_{P,t}} + w_{D,t}e^{-\eta\epsilon}}.$$

□

B Experimental Results

For the sake of reproducibility, Python scripts used to generate our results are available at https://tinyurl.com/yagw3xzx.

First we present the complete numerical results for the experiments presented in Section 5.1 in Table 2 (see the caption and the description in the main text for the details). Further, we performed experiments similar to the ones we performed on the MNIST dataset on other datasets from UCI Machine Learning repository [1]. The datasets used are listed in Table 1 and corresponding results are shown in Figures 1-7.

For each dataset we used $T = 10000$ randomly chosen data points for our experiments. The remaining the data points are used to choose the target error rates, shown in the last column of Table 1. In particular, we trained a random forest on $T/4 = 2500$ of these unused data points, and measured the error rate on the rest. Note, neither this random forest nor the data is used for the experiments other than to choose ϵ.

Results for MNIST dataset from the main text are reproduced in Fig. 1. Results obtained in the other datasets are presented in Fig. 2 - 9, and yield the same conclusions as those drawn from the MNIST dataset in the main text.

TABLE 1: Datasets from UCI ML Repository [1]

Data-Set	Description	Features	Classes	Target (ϵ)
MNIST	Scanned hand-written digits	784	10	0.08
SENSIT	Vehicle types from wireless sensor nets	100	3	0.20
COD-RNA	Coding/Non-coding parts of RNA	8	2	0.07
COVER	Dominant forest types based on images	54	7	0.25
CONNECT-4	Outcome of a multi-player game	126	2	0.22
LETTER	Letter recognition from pixel displays	16	26	0.12
MAGIC	Simulated data to register high energy gamma particles	11	2	0.14

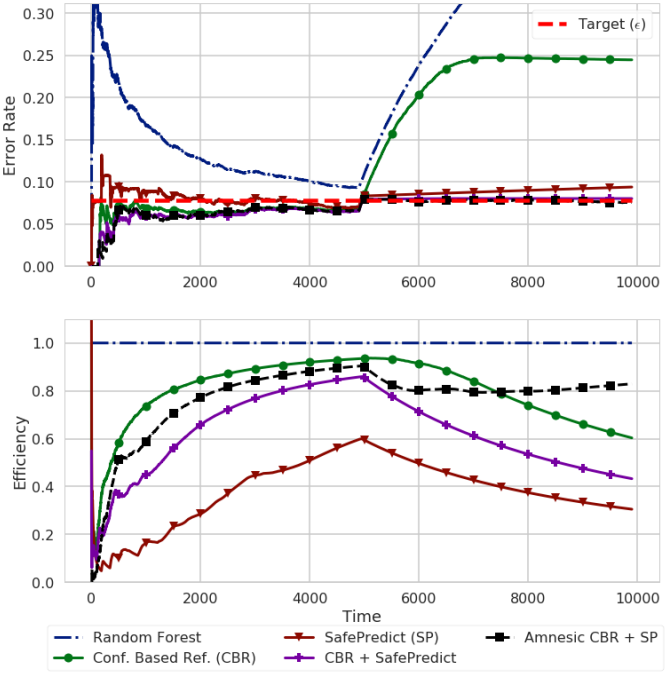
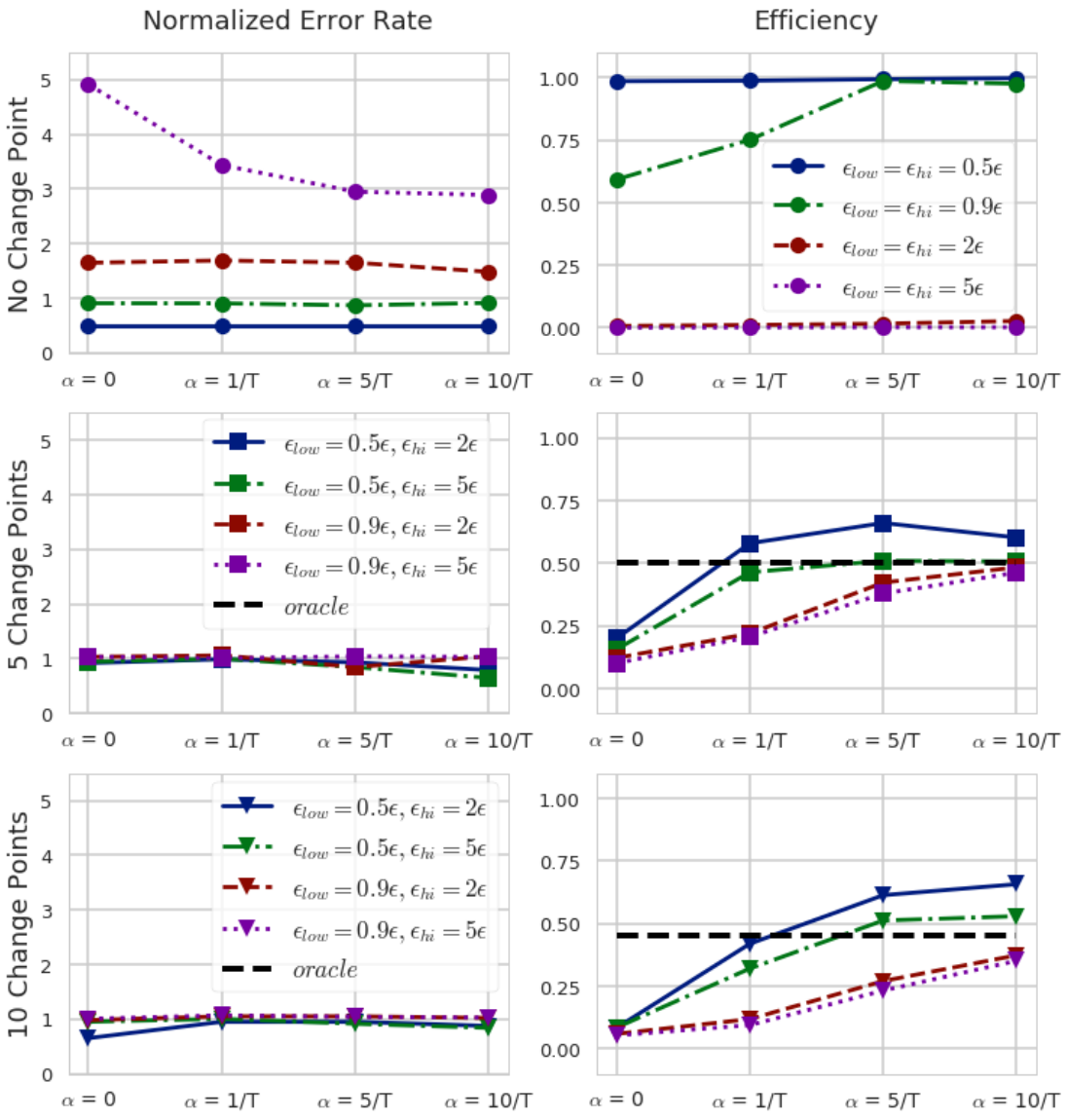


Fig. 1: *MNIST Dataset ($\epsilon = 0.08$, reproduced from Fig. 4 of the main text): Efficiency is 1.0 for the base predictor but lower for the various refusing meta-algorithms. Validity is measured as a fraction of the target error rate. So the base predictor has a poor error rate (way over ϵ). All the SafePredict variants rapidly approach a normalized error rate value of 1 though the error rate increases at the change point at time t = 5000. The confidence based competition cannot guarantee an asymptotic validity. Two forms of adaptivity help reduce the number of refusals: weight-shifting especially with a high α value and amnesic adaptivity. Combining both leads to the highest efficiency while preserving validity.*

References

[1] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml



			Efficiency: T^*/T					Normalized Error Rate: $L^*_{P,T}/T^*/\epsilon$			
$numChange$	ϵ_{low}/ϵ	ϵ_{hi}/ϵ	oracle	$\alpha = 0$	$\alpha = 1/T$	$\alpha = 5/T$	$\alpha = 10/T$	$\alpha = 0$	$\alpha = 1/T$	$\alpha = 5/T$	$\alpha = 10/T$
0	0.5	0.5	1.000	0.985	0.987	0.993	0.997	0.502	0.502	0.502	0.502
0	0.9	0.9	1.000	0.592	0.751	0.986	0.975	0.907	0.905	0.869	0.913
0	2.0	2.0	0.000	0.007	0.011	0.016	0.027	1.648	1.690	1.650	1.482
0	5.0	5.0	0.000	0.000	0.001	0.002	0.002	4.914	3.437	2.949	2.890
5	0.5	2.0	0.500	0.199	0.579	0.660	0.602	0.919	0.988	0.931	0.790
5	0.5	5.0	0.500	0.155	0.463	0.509	0.506	0.943	0.998	0.844	0.647
5	0.9	2.0	0.500	0.121	0.218	0.422	0.483	1.031	1.059	0.845	1.043
5	0.9	5.0	0.500	0.100	0.205	0.378	0.462	1.034	1.006	1.041	1.031
10	0.5	2.0	0.545	0.083	0.419	0.613	0.657	0.650	0.950	0.950	0.875
10	0.5	5.0	0.545	0.084	0.320	0.512	0.529	0.951	1.004	0.914	0.839
10	0.9	2.0	0.545	0.059	0.118	0.270	0.373	0.981	1.050	1.051	1.021
10	0.9	5.0	0.545	0.053	0.094	0.232	0.351	1.005	1.068	1.040	1.028

TABLE 2: *Experimental results on synthetic data:* The first three columns gives the characteristics of each loss sequence. The next group of columns report on the efficiency (fraction of predictions made) for the oracle and increasing values of α. The final group of columns report on the error rate normalized with respect to the target error rate $\epsilon = 0.05$ of SafePredict with the same increasing values of α. The same values are also represented in the figures above. One can draw two main conclusions: i.) a small α does better when the error rate of the base predictor is high relative to the target, e.g. see the efficiency plot for the error rate 2ϵ or 5ϵ, whereas a large α adapts better when the underlying error rate at least sometimes falls below the target error rate. ii.) As the number of change points increases, the error bound behaves more like $1/\sqrt{T^*}$ rather than $1/T^*$.

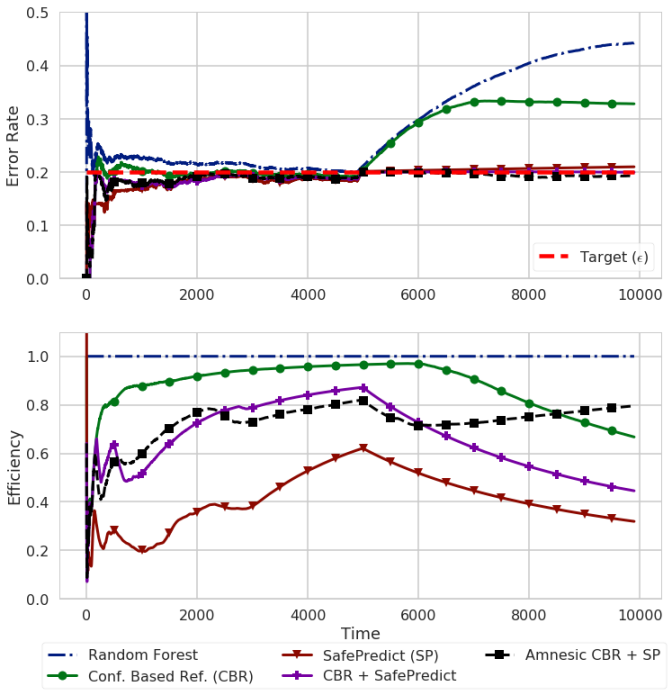


Fig. 2: *SENSIT Dataset ($\epsilon = 0.20$): Note one can draw similar conclusions as with MNIST Dataset (Fig. 1).*

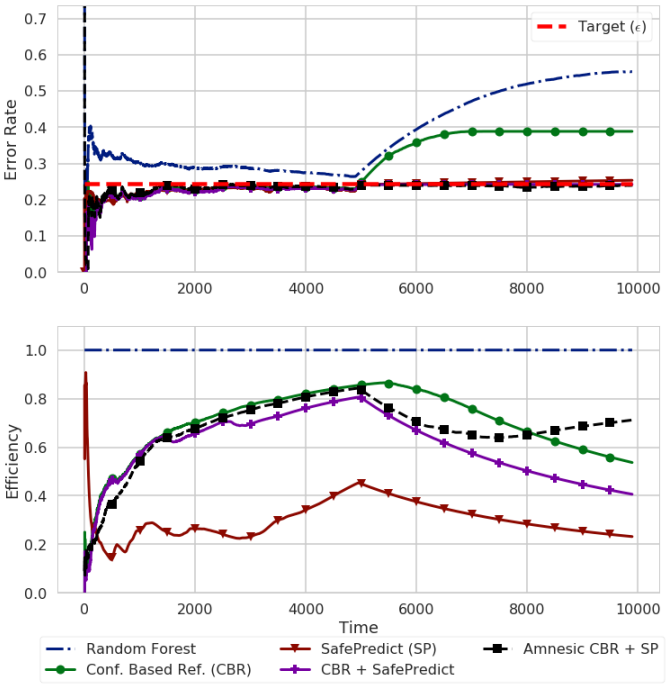


Fig. 4: *COVER Dataset ($\epsilon = 0.25$): Note one can draw similar conclusions as with MNIST Dataset (Fig. 1).*

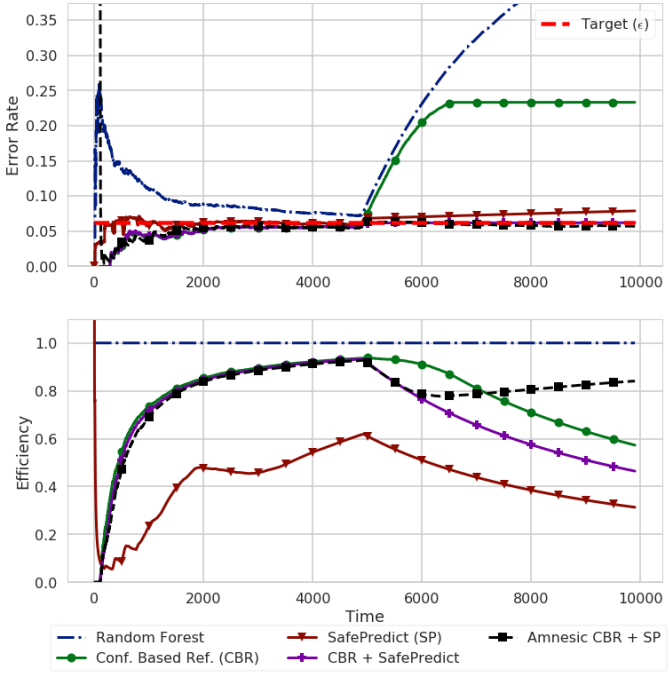


Fig. 3: *COD-RNA Dataset ($\epsilon = 0.07$): Note one can draw similar conclusions as with MNIST Dataset (Fig. 1).*

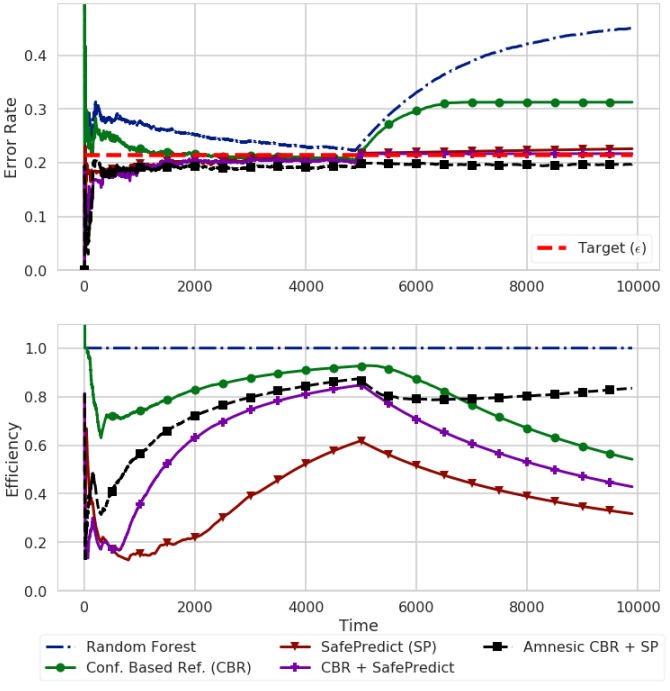


Fig. 5: *CONNECT-4 Dataset ($\epsilon = 0.22$): Note we can draw similar conclusions as with MNIST Dataset (Fig. 1).*

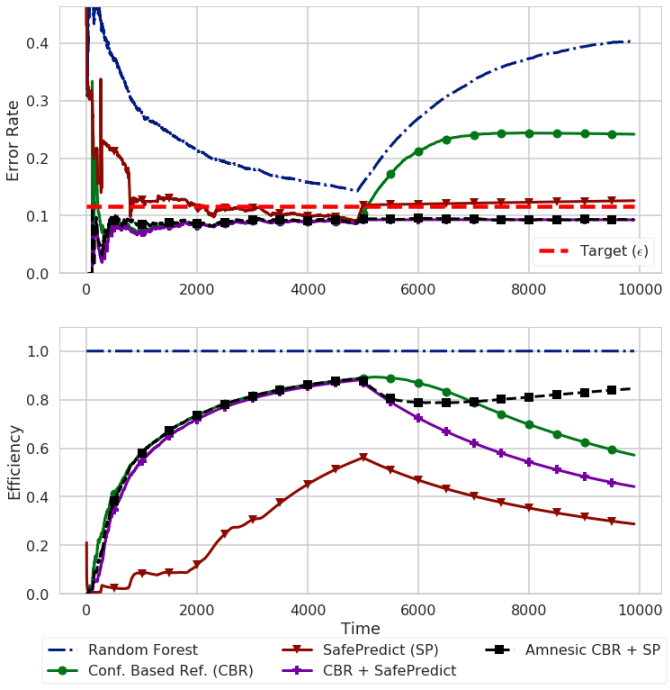


Fig. 6: *LETTER Dataset ($\epsilon = 0.12$): Note we can draw similar conclusions as with MNIST Dataset (Fig. 1).*

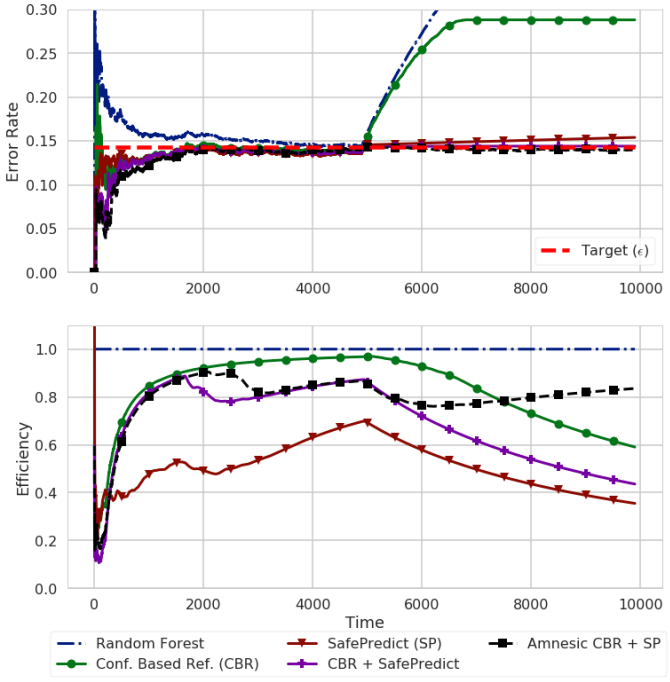


Fig. 7: *MAGIC Dataset ($\epsilon = 0.14$): Note we can draw similar conclusions as with MNIST Dataset (Fig. 1).*