# EE290S: ADAPTATION II

HAOTIAN GU AND YILING YOU

## 1. Doubling Trick "AdaHedge"

1.1. **The Learning Rate and the Mixability Gap.** Recall that in the sequential prediction setting, we have studied the exponential weights strategy, or the Hedge algorithm: the weights of predicting over 2 outcomes $x \in \{0, 1\}$ at the round $t$, denoted by $w_{t,x}$, is defined to be

$$w_{t,x} = \frac{e^{-\eta_t L_{t-1,x}}}{e^{-\eta_t L_{t-1,0}} + e^{-\eta_t L_{t-1,1}}} \tag{1.1}$$

In the last lecture, we have discussed a simple doubling trick to update the learning rate $\eta$ when the total time $T$ is not known in advance. The basic idea of this simple doubling trick is to keep scaling down $\eta$ by a constant factor $\sqrt{2}$ as $t$ becomes larger. Under this setting, we are in fact choosing our learning rate more and more conservatively. But in practice, it is very possible that as $t$ increases, we become more confident in our predictions. For example, if we are given a $i.i.d$ sequence of Ber(0.7) random variables, $\mathbb{P}(X_t = 1) = 0.7$, then we expect that our learning strategy will give us a much larger weight of predicting 1 as $t$ increases. And in this situation, we may set $\eta$ to be relatively big instead of keeping scaling down $\eta$. In order to fulfill the task, we first need a measurement on how easy it is to predict, or how certain we are about the prediction, at round $t$. This is the reason why we introduce the **mixability**.

**Definition 1.1.** We define the gap between the per-round loss of the Hedge algorithm and the per-round contribution to the fixed loss

$$\delta_t(\eta_t) = h_t - m_t(\eta_t) = \langle w_t, l_t \rangle - (-\frac{1}{\eta_t} \ln (\langle w_t, e^{-\eta_t l_t} \rangle)) \tag{1.2}$$

the mixability gap at round $t$, and $\Delta_T = \sum_{t=1}^{T} \delta_t(\eta_t)$ the cumulative mixability gap.

From the analysis we have done before in Lecture [?], an upper bound for the mixability gap at round $t$ is

$$\delta_t(\eta_t) \leq \frac{\eta_t}{8} \tag{1.3}$$

using Hoeffding's bound on the cumulant generating function. From this expression, we can easily observe that when the weights **concentrate**, which means there exists $x^* \in \{0, 1\}$ such that $w_{t,x^*} \approx 1$, then $h_t \approx l_{t,x^*}$ and $m_t \approx -\frac{1}{\eta} ln(e^{-\eta l_{t,x^*}}) = l_{t,x^*}$. So we have $\delta_t(\eta_t)$ small when the weights concentrate. From the definition of $w_t$, we can see that in order to get the concentrated weights, the following 2 conditions are sufficient:

- **Separation condition**: There exists $x^*$ and $m^*$ such that for all $t \geq t_{m^*}$, $L_{t-1,x} - L_{t-1,x^*} \geq \alpha t$;

- The learning rate $\eta_t$ is not small.

The separation condition is typically satisfied when we consider the sequence is $i.i.d$ Ber(p) with a large $p$. So now intuitively we want to make $\eta_t$ and $\delta_t$ interact with each other in the following way:

- when $\delta_t$ remains small, or equivalently $\Delta_T$ is small, it means it is easy to make predictions and there is no need to make $\eta_t$ smaller;
- when $\delta_t$ remains big, or equivalently $\Delta_T$ is big, we realize that we are not certain about our predictions, and it is better to have a more conservative learning rate.

Here we provide another way to understand the definition of $\delta_t$ (1.2). Originally we get this expression from the analysis we have done for multiplicative weights, but in fact for any random variable $X$ we can define a similar quantity

$$q(X) := \mathbb{E}[X] - (-\frac{1}{\eta}\ln \mathbb{E}[e^{-\eta X}]). \tag{1.4}$$

$$r(X) := e^{\eta q(X)} = \frac{e^{\eta \mathbb{E}[X]}}{\mathbb{E}[e^{\eta X}]} \tag{1.5}$$

In fact this quantity $r(X)$ is related to the variance of $X$. This matches the interpretation we have for the mixability gap before: it reflects how random the sequence is.

From the above discussions, we realize that We can simply run the Hedge algorithm with a fix learning rate $\eta$ until the smallest $T$ such that $\Delta_t(\eta)$ exceeds an appropriate **budget** $b(\eta)$.

1.2. **Doubling Trick AdaHedge.** Now we are ready to introduce the Doubling Trick AdaHedge algorithm. The doubling trick divides the rounds in segments $i = 1, 2, \ldots$ and on each segment restarts the Hedge algorithm with a different learning rate $\eta_i$. To adaptively set the learning rate, we set $\eta_0 = \phi = 2$ initially, and scale down the learning rate by a factor of $\phi$ for every new segment, such that $\eta_i = \phi^{1-i}$. We monitor $\Delta(\eta_i)$, measured only on the losses in the $i^{th}$ segment. When it exceeds its budget $b_i = b(\eta_i)$, a new segment is started.

**Algorithm 1** Doubling Trick AdaHedge

**Initialize:**
  $\eta_0 = 2$, $\Delta_0 = 0$, $\phi = 2$
**for** t = 1 to T **do**
  **if** $\Delta_{t-1} \geq b(\eta_{t-1}) := \frac{\ln 2}{\eta_{t-1}}$ **then**
    $\Delta_{t-1} \leftarrow 0$
    $\eta_t \leftarrow \frac{\eta_{t-1}}{\phi}$
  **end if**
  $w_{t,x} \leftarrow \frac{e^{-\eta_t L_{t-1,x}}}{e^{-\eta_t L_{t-1,0}} + e^{-\eta_t L_{t-1,1}}}$
  $\Delta_t \leftarrow \Delta_{t-1} + \delta_t(\eta_t)$
**end for**

Unlike the simple doubling trick, now whether $\eta$ will be halved depends on the mixability gap but not on the time itself. People may be confused about why we need to reset the cumulative mixability gap after a new learning rate is set up. This is because we only

want to focus on measuring how the algorithm performs under the new learning rate, so the previous cumulative mixability gap is irrelevant. But this doesn't mean all the information in history is all thrown away, $w_{t,x}$ still depends on $X_1, \ldots, X_{t-1}$.

1.3. **Discussion on $b(\eta)$.** Let's make a brief discussion on the choice of $b(\eta)$.
Note in this algorithm, the budget we set for the leaning rate $\eta_t$ is $b(\eta_t) = \frac{\ln 2}{\eta_t}$. So in this algorithm, as we **hedge** more, which means choosing a small $\eta$, the budget increases. This is the upper bound of approximation error for $\eta_t$ we have derived in the analysis of the exponential weight algorithm. The intuition behind this particular choice of $b$ is that we want to make the approximation error and the cumulative estimation error for $\eta_t$ to be equal before we change the learning rate, which may give us a good upper bound of the total error in the end. The concrete error analysis will be provided later.

1.4. **Analysis on the Regret Bound.** The performance of the Doubling Trick AdaHedge is analyzed in the following theorems.

**Theorem 1.1.** *Let $m(T)$ denote the number of learning rates decreasing before round $T$.*

$$R_{AdaHedge}(T) \leq 2 \ln 2 \left( \frac{\phi^{m(T)} - 1}{\phi - 1} \right) + m(T) \frac{1}{8} \tag{1.6}$$

*Proof.* First let's consider the case for just one segment, which means the agent runs Hedge with learning rate $\eta \in (0, 1]$, and that after $T$ rounds it has just used up the budget. Then from the definition and 1.3 we know that

$$b(\eta) \leq \Delta_T(\eta) < b(\eta) + \eta/8 \tag{1.7}$$

The cumulative loss of Hedge is bounded by

$$\sum_{t=1}^{T} \langle w_t, l_t \rangle = \Delta_T(\eta) - \frac{1}{\eta} \sum_{t=1}^{T} \ln \left( \langle w_t, e^{-\eta_t l_t} \rangle \right) \tag{1.8}$$

$$= \Delta_T(\eta) - \frac{1}{\eta} \ln \sum_{x=0}^{1} \frac{1}{2} e^{-\eta L_{T,x}} \tag{1.9}$$

$$\leq b(\eta) + \eta/8 + \frac{1}{\eta} \ln \frac{1}{2} e^{-\eta L_{T,x^*}} \tag{1.10}$$

$$\leq \frac{1}{8} + \frac{2}{\eta} \ln(2) + L_T^* \tag{1.11}$$

Therefore, the regret per segment is bounded by $\frac{1}{8} + \frac{2}{\eta} \ln(2)$. Summing over all $m(T)$ segments, and plugging in

$$\sum_{i=1}^{m(T)} \frac{1}{\eta_i} = \sum_{i=0}^{m(T)-1} \phi^i = \frac{\phi^{m(T)} - 1}{\phi - 1}$$

gives the required inequality. □

Since we have $\eta_T = \frac{2}{2^{m(T)}}$, $2^{m(T)} - 1$ is approximately $\frac{2}{\eta_T}$.

1.5. **Discussion on $\phi$.** For the factor $\phi$, in the previous discussion, we just set it to be 2, but in fact it indeed have an impact on the performance of our prediction. It has been proved in [2] that:

**Theorem 1.2.** *Suppose the agent runs the Doubling Trick AdaHedge for $T$ rounds. Then its regret is bounded by*

$$R_{AdaHedge}(T) \leq \frac{\phi\sqrt{\phi^2-1}}{\phi-1}\sqrt{\frac{4}{e-1}L_T^* \ln(K)} + O(\ln(L_T^*+2)\ln(K)) \tag{1.12}$$

From this theorem we can see that for the best upper bound, we can set $\phi$ that minimizes the leading factor, which is in fact the golden ratio $\frac{1+\sqrt{5}}{2}$. In this case, $\frac{\phi\sqrt{\phi^2-1}}{\phi-1} \approx 3.33$. But in order to make life easier, we can simply choose $\phi = 2$, which leads to $\frac{\phi\sqrt{\phi^2-1}}{\phi-1} \approx 3.46$. In practice, the performance of $\phi = 2$ and $\phi = \frac{1+\sqrt{5}}{2}$ should be very similar. Numerical experiments showing simulation and examples are provided in Section 4.1.

## 2. Mixability and $\eta$

Instead of the upper bound for $\delta_t(\eta)$ we have mentioned in 1.3, in the next lemma we present a new upper bound.

**Lemma 2.1.**

$$\delta_t(\eta) \leq (e-2)\eta\left(1 - w_{t,x^*}(\eta)\right),$$

*for $\forall t$ and $\eta \in (0,1]$, where $w_{t,x^*}$ is the Hedge posterior probability of the best action, i.e., $w_{t,x^*} = \max_{1\leq k\leq K} w_{t,x_k}$.*

Some notes:

- One takeaway inferred by Lemma 2.1 is that when $w_{t,x^*}(\eta) \approx 1$ (i.e., when the weights concentrate), the mixability gap vanishes and therefore incurs only a small addition to the $\Delta_t$.
- In Section 4.2 we provide a heat map of the upper bound in Lemma 2.1 to further illustrate the relationship between $\delta_t$ and $(\eta, w)$ pair.

*Proof.* Let $h_t = \langle w_t, l_t\rangle = w_{t,1-x^*}\cdot 1 + w_{t,x^*}\cdot 0 = w_{t,1-x^*}$, therefore

$$\begin{aligned}\delta_t &= h_t + \frac{1}{\eta}\ln\left(w_{t,1-x^*}\cdot e^{-\eta\cdot 1} + w_{t,x^*}\cdot e^{-\mu\cdot 0}\right)\\ &= h_t + \frac{1}{\eta}\ln\left(h_t\cdot e^{-\eta} + (1-h_t)\right)\end{aligned}$$

if we do the Taylor expansion of the last equation both at $h_t = 0$ and $h_t = 1$, we have

$$\delta_t = 0 + \left(1 + \frac{1}{\eta}\cdot(e^{-\eta}-1)\right)h_t - \frac{1}{\eta}\cdot\frac{(e^{-\eta}-1)^2}{\left[\tilde{h}\,(e^{-\eta}-1)+1\right]^2}{h_t}^2, \tilde{h}\in(0,h_t)$$

$$\delta_t = 0 + \left(1 + \frac{1}{\eta}\cdot\frac{e^{-\eta}-1}{e^{-\eta}}\right)(h_t-1) - \frac{1}{\eta}\cdot\frac{(e^{-\eta}-1)^2}{\left[\hat{h}\,(e^{-\eta}-1)+1\right]^2}(h_t-1)^2, \hat{h}\in(h_t,1)$$

combined with the fact that $e^{-\eta} \le 1-\eta+\frac{\eta^2}{2}, \forall \eta$ and $e^{\eta} \le 1+\eta+(e-2)\eta^2, \eta \le 1$, the above inequalities give us

$$\delta_t \le \left(1+\frac{1}{\eta}\cdot(e^{-\eta}-1)\right) h_t \le \left[1+\frac{1}{\eta}\cdot\left(\frac{\eta^2}{2}-\eta\right)\right] h_t = \frac{1}{2}\eta \cdot h_t \tag{2.1}$$

$$\begin{aligned}\delta_t \le \left(1+\frac{1}{\eta}\cdot\frac{e^{-\eta}-1}{e^{-\eta}}\right)(h_t-1) &= \left[\frac{1}{\eta}\cdot(e^{\eta}-1)-1\right](1-h_t) \\ &\le \left[\frac{1}{\eta}\cdot\left(\eta+(e-2)\eta^2\right)-1\right](1-h_t) \\ &= (e-2)(1-h_t)\eta \end{aligned} \tag{2.2}$$

$$\implies \delta_t(\eta) \le \min\left\{\frac{1}{2}\eta\cdot h_t, (e-2)(1-h_t)\eta\right\} = (e-2)\eta\left(1-w_{t,x^*}(\eta)\right) \tag{2.3}$$

□

Now assume the separation condition is satisfied, then

$$1-w_{t,x^*}(\eta) = \frac{e^{-\eta L_{t-1,x}}}{e^{-\eta L_{t-1,x}}+e^{-\eta L_{t-1,x^*}}} = \frac{e^{-\eta(L_{t-1,x}-L_{t-1,x^*})}}{e^{-\eta(L_{t-1,x}-L_{t-1,x^*})}+1} \le \frac{e^{-\eta\alpha t}}{e^{-\eta\alpha t}+1} \le e^{-\eta\alpha t}. \tag{2.4}$$

Therefore $\delta_t \le (e-2)\eta\cdot e^{-\eta\alpha t}$ decays exponentially, and the cumulative mixability gap

$$\Delta_T = \sum_{t=t_{m^*}}^{T}\delta(t) \lesssim \sum_{t=t_{m^*}}^{T} e^{-\eta\alpha t} = \frac{e^{-\eta\alpha t_{m^*}}}{1-e^{-\eta\alpha}} \tag{2.5}$$

is fixed and doesn't accumulate, which indicates that we stop hedging at some point.

## 3. "Elegant" AdaHedge

In the previous classes we split the regret for Hedge into two parts: $R_T \le (M_T - L_T^*) + (H_T - M_T)$, obtained an upper bound for both and then equalize the two to get the optimal learning rate $\eta$. A less pessimistic and more "elegant" way to decrease the learning rate is that we don't consider an upper bound on $\Delta_T = H_T - M_T$, but instead we aim to equalize $\Delta$ and $\frac{\ln 2}{\eta}$ directly. The key observation here is that the mixability gap $\Delta_t$ is nondecreasing in $t$ and can be observed online, therefore we are able to update the learning rate on the fly.

The previous discussion inspires a new update rule as described below. At round $t$ we decrease the learning rate with time according to

$$\eta_t = \frac{\ln 2}{\Delta_{t-1}} \tag{3.1}$$

where $\Delta_0 = 0$ so that $\eta_1 = \infty$. One should note that the new learning rate doesn't involve the (potentially unknown) end time $T$; what's more, $\Delta_{t-1}$ is computable when we are at round $t$.

We will very soon see the regret analysis for the "elegant" AdaHedge but just as a preview, we state the regret bound as a theorem below as well as a sketch of the proof[1].

**Theorem 3.1.** *The "elegant" AdaHedge regret is*

$$R_T \leq 2\Delta_T \lesssim \sqrt{T} \tag{3.2}$$

*Proof.* To prove the first part, $R_T \leq 2\Delta_T$, notice that $R_T = M_T - L_T^* + \Delta_T$, therefore we aim to show

$$M_T - L_T^* \leq \Delta_T. \tag{3.3}$$

To see why this is true, we first need a Lemma saying that the mix loss is bounded by the mix loss we would have incurred if we would have used the final learning rate $\eta_T$ all along: *for a strategy dec that choose the learning rate such that $\eta_1 \geq \eta_2 \geq \ldots$ (which applies to our "elegant" AdaHedge case), the cumulative mix loss does not exceed the cumulative mix loss for the strategy that uses the last learning rate $\eta_T$ from the start, i.e., $M^{dec} \leq M^{(\eta_T)}$.* Therefore (*ah* stands for AdaHedge)

$$M^{ah} \leq M^{\eta_T^{ah}} \leq L^* + \frac{\ln 2}{\eta_T^{ah}} = L^* + \Delta_{T-1}^{ah} \leq L^* + \Delta_T^{ah}.$$

The remaining task is to establish a bound on $\Delta$. The detailed froof can be found in [1]. A sketch of the proof is as follows:

(1) We introduce the *variance of the losses*, $v_t = \mathrm{Var}_{k\sim\omega_t}[l_{t,k}] = \sum_k \omega_{t,k}(l_{t,k} - h_t)^2$. Intuitively, $v_t$ is small when all outcomes have approximately the same loss, or when the weights $w_t$ are concentrated on a single outcome.

(2) To bound the $\Delta_T$, we start with a bound on mixability gap in a single round using Bernstein's bound[2]:

$$\delta_t \leq g(\eta_t)v_t$$

where $g(x) = \frac{e^x - x - 1}{x}$. Further, $v_t \leq (1 - h_t)h_t \leq 1/4$. Note that the Bernstein's bound is more sophisticated that Hoeffding's bound, because it expresses that the mixability gap $\delta_t$ is small not only when the learning rate $\eta_t$ is small, but also when the variance $v_t$ is small, which is the same as saying that the weights are concentrated.

(3) By telescoping and applying Bernstein's inequality, we obtain a bound on $\Delta_T$[3]:

$$(\Delta_T)^2 \leq V_T \ln 2 + \left(\frac{2}{3}\ln 2 + 1\right) T\Delta_T$$

where $V_T$ is the cumulative variance of losses $V_T = \sum v_t$.

(4) Combination of the results yields the following regret bound[4]:

$$R \leq 2\sqrt{V \ln 2} + T\left(\frac{4}{3}\ln 2 + 2\right)$$

[1]For those especially interested in a more rigorous statement of the proof, we provide more technical details in the Appendix.

[2]See Appendix, Proposition 5.1.

[3]See Appendix, Remark 5.1.

[4]See Appendix, Remark 5.2.

(5) The first regret bound above is difficult to interpret, because the cumulative loss $V$ depends on the actions of the "elegant" AdaHedge strategy itself (through the weights $w_t$). A regret bound for the "elegant" AdaHedge that depends only on the data can be further derived by observing the following inequality on $V$[5]:

$$V \leq (T - L^*)L^* + 2T\Delta$$

and therefore we obtain the AdaHedge worst-case regret bound[6]

$$R \leq 2\sqrt{(T - L^*)L^* \ln 2} + T\left(\frac{16}{3}\ln 2 + 2\right)$$

which is our main result.

□

## 4. Numerical Results

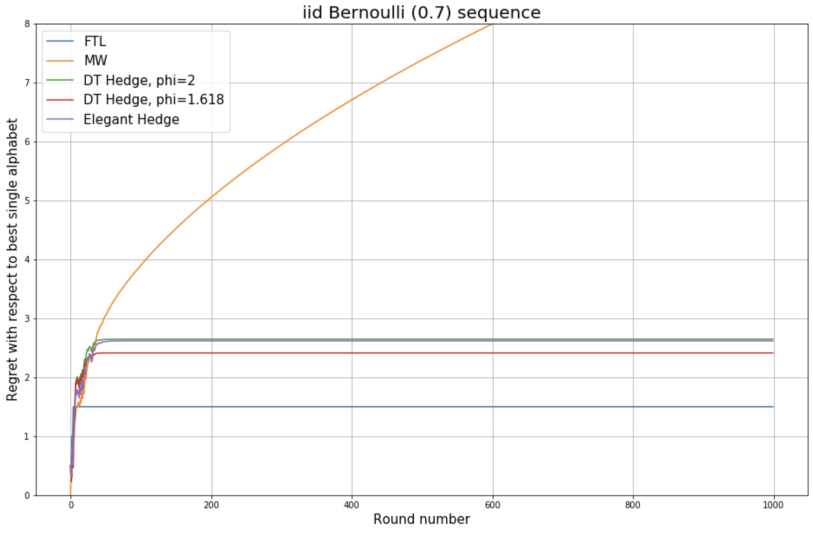
4.1. **Regrets and Learning Rates with Multiple Learning Algorithms.** In this section we show some numerical results learning (a) i.i.d. Ber(0.7) and (b) 1-periodic sequence (i.e., the '0101...' sequence) using

- FTL (Follow-the-Leader),
- Multiplicative Weight,
- DT AdaHedge, and
- Elegant AdaHedge.

Source code available upon request.

First, we focus on the i.i.d. Ber(0.7) sequence, where $T = 1000$. The patterns of regret and learning rate $\eta_t$ with different learning algorithms are compared in Figure 1 and 2.

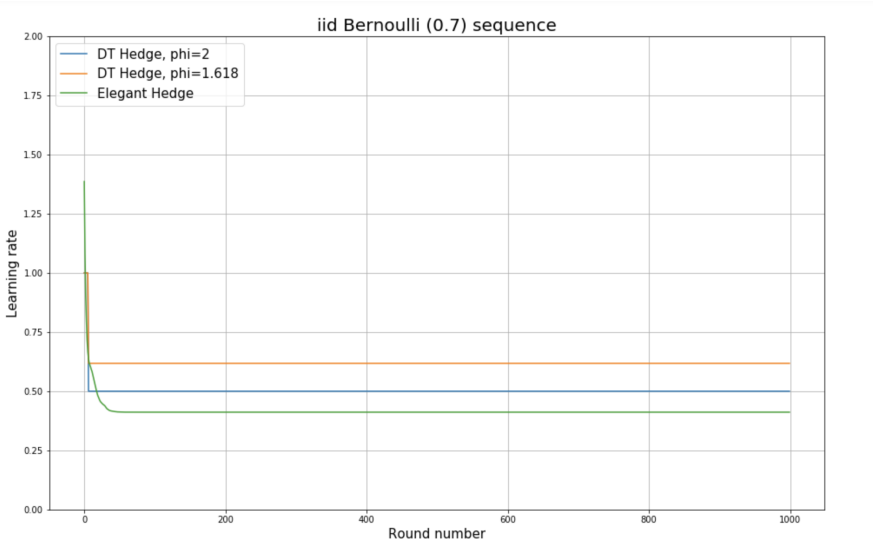
Figure 1. Regret with i.i.d. Ber(0.7) sequence.




[5]See Appendix, Remark 5.3.

[6]Same trick as in Remark 5.2.

FIGURE 2. Learning rate with i.i.d. Ber(0.7) sequence.




1. Regret: we see that with FTL, DT AdaHedge and Elegant AdaHedge, the regrets stop growing in less than just 100 rounds.
2. Learning rate: the learning rates $\eta_t$ stay constant after 50 rounds, with the magnitude well apart away from zero (at least 0.3 in our experiment).

This is a decent illustration of the case where we get a sequence from nature and it satisfies the separation condition. No need to hedge too much in this case – the learning rate could be kept relatively large and we identify the best arm rapidly.

We then consider the 1-periodic sequence, again with $T = 1000$. The regret results are displayed in Figure 3 and learning rate in Figure 4.

1. Regret: the behaviors of regret get different from what we saw for the i.i.d. Ber(0.7) case: now FTL suffers a linear growth regret, while with Multiplicative Weight, DT AdaHedge and Elegant AdaHedge, instead of staying constant after several rounds, we observe sub-linear regret growth (which is actually of order $\mathcal{O}(\sqrt{T})$).
2. Learning rate: the algorithms keep shrinking the learning rates and become much more conservative than we were when we had a sequence satisfying the separation condition.

FIGURE 3. Regret with 1-periodic sequence, and its zoom in.

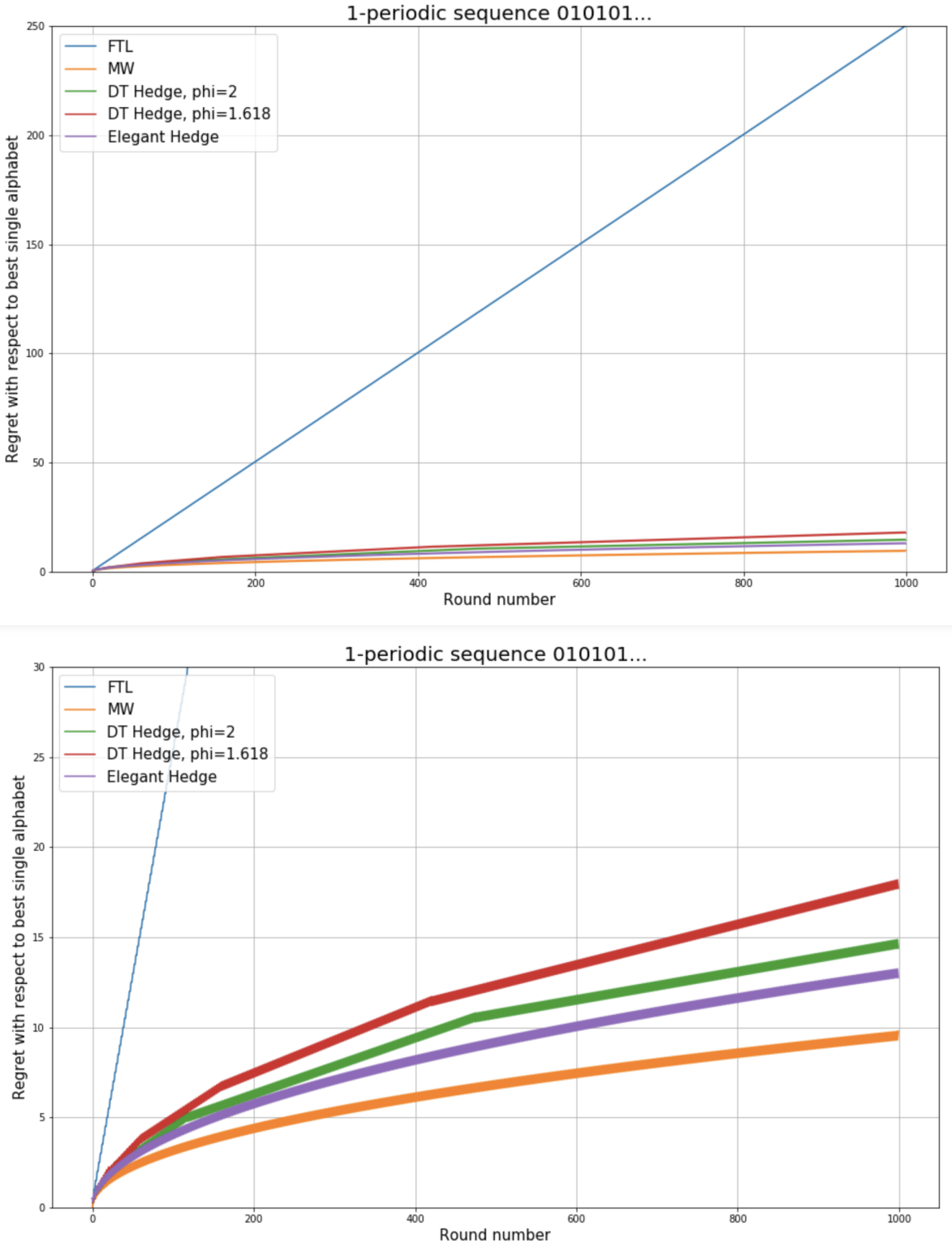
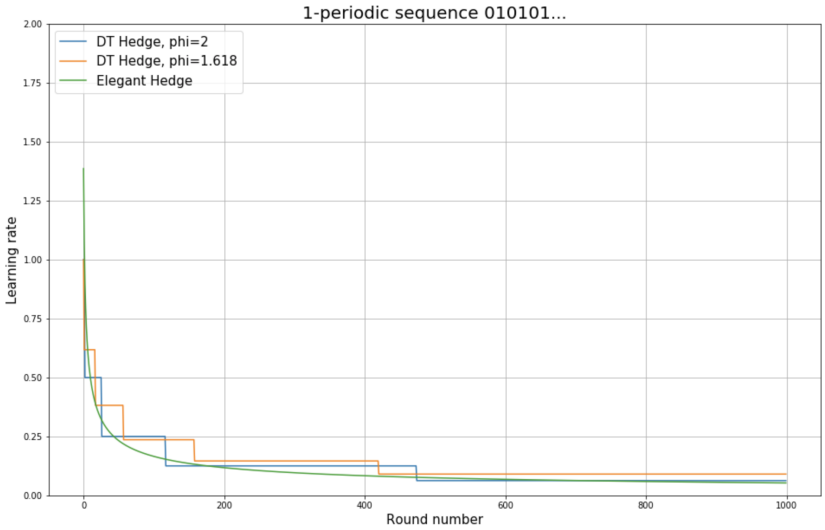


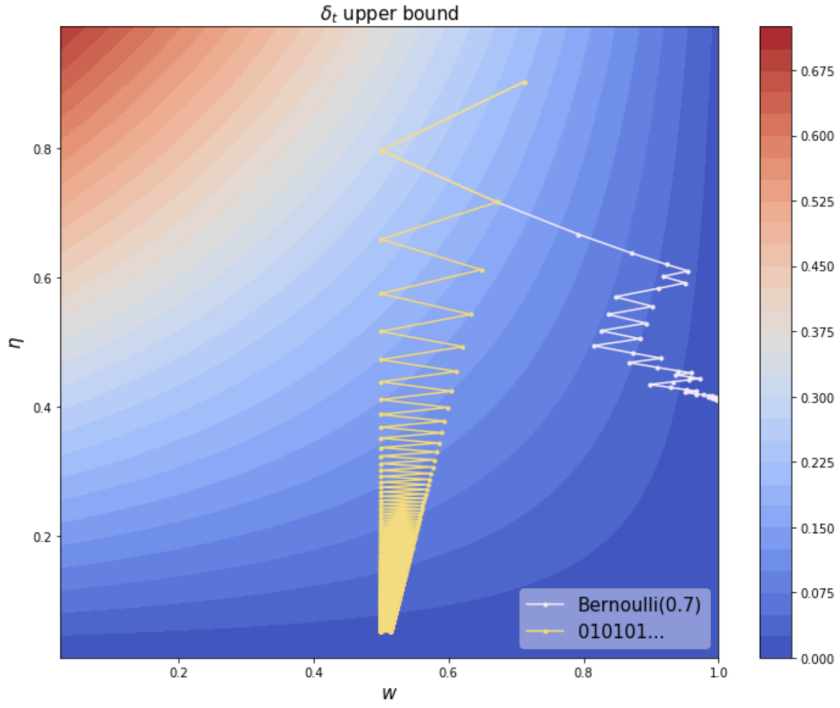
FIGURE 4. Learning rate with 1-periodic sequence.




4.2. **Heat Map of $\delta_t$ Upper Bound versus $(\eta, w)$.** Recall that in Lemma 2.1 we derived a new upper bound of $\delta_t$, which depends both on the learning rate $\eta$ and the weight $w$. To further see this relationship, we display a heat map in Figure 5, where the value of the heat map is simply the right hand side of the inequality in Lemma 2.1, i.e.,

$$U(\eta, w) = (e-2)\eta\,(1-w)\,. \tag{4.1}$$

FIGURE 5. Learning rate with 1-periodic sequence.



Some observations we want to point out:

1. From the expression of $U(\eta, w)$, note that for a fixed weight $w_0$, $U(\eta, w_0)$ is an increasing linear function of $\eta$ with slope $(e-2)(1-w_0) \geq 0$. This means that if one draws a vertical line corresponding to $w = w_0$ and goes upward along that line, she would first see deep blue (same as saying $\delta_t$ very close to 0) and warm up as she climbs (larger $\delta_t$), and one finds herself warm up more rapidly with smaller $w_0$.
2. Similarly, for a fixed learning rate $\eta_0$, $U(\eta_0, w)$ is a decreasing linear function of $w$ with slope $-(e-2)\eta_0 \leq 0$. Therefore one cools down when she goes from left to right along a horizontal line.

Regarding the two trajectories:

1. The trajectories are the numerical realizations of $(\eta_t, w_{t,x^*})$, with i.i.d. Ber(0.7) (the white trajectory) and 1-periodic sequence (the yellow one) using Elegant AdaHedge.
2. The trajectories go downward as the round $t$ increases.
3. It is clear to see how Elegant AdaHedge performs differently with the sequences: for the i.i.d. Ber(0.7) sequence in our experiment, the trajectory $(\eta_t, w_{t,x^*})$ converges to $(\hat{\eta}, 1)$ where $\hat{\eta} \approx 0.4$, while for the 1-periodic sequence, the trajectory converges to $(0, \frac{1}{2})$.

## 5. Appendix

**Proposition 5.1.** *The Bernstein's Bound says that: let $X$ be a random variable taking values in $[0,1]$. Let $\sigma = \sqrt{\mathbb{E}X^2 - (\mathbb{E}X)^2}$. Then for any $\eta > 0$,*

$$\ln \mathbb{E}\left[e^{-\eta(X-\mathbb{E}X)}\right] \leq \sigma^2(e^\eta - 1 - \eta) \leq \mathbb{E}X(1-\mathbb{E}X)(e^\eta - 1 - \eta). \tag{5.1}$$

*Proof.* First note that $Y = -X + \mathbb{E}X$ is a zero-mean random variable taking values in $[-1, 1]$ with variance $\sigma^2$. Observe that the function $(e^x - x - 1)/x^2$ is nondecreasing for all $x \in \mathbb{R}$. Since $-Y \leq 1, \eta > 0$,

$$e^{-\eta Y} + \eta Y - 1 \leq Y^2(e^\eta - \eta - 1)$$

Taking expected values on both sides, taking logarithms, and using $\ln(1+x) \leq x$, we obtain the following inequality

$$\ln \mathbb{E}[e^{-\eta Y}] \leq \sigma^2(e^\eta - 1 - \eta).$$

Also note that since $X \in [0,1]$,

$$\sigma^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2 \leq \mathbb{E}X - (\mathbb{E}X)^2 = \mathbb{E}X(1-\mathbb{E}X).$$

□

*In the regret bound proof we take $X = l_{t,k}$, therefore $\mathbb{E}X = h_t, \sigma^2 = v_t, \ln \mathbb{E}\left[e^{-\eta(X-\mathbb{E}X)}\right] = \eta_t \cdot h_t + \eta_t(\delta_t - h_t) = \eta_t \cdot \delta_t$, hence $\delta_t \leq g(\eta_t)v_t$.*

**Remark 5.1.** First we do the telescoping

$$\Delta_2 = \sum_{t=1}^{T} \left(\Delta_t^2 - \Delta_{t-1}^2\right) = \sum_t \left((\Delta_{t-1} + \delta_t)^2 - \Delta_{t-1}^2\right) = \sum_t \left(2\delta_t \Delta_{t-1} + \delta_t^2\right) \tag{5.2}$$

$$= \sum_t \left(2\delta_t \frac{\ln 2}{\eta_t} + \delta_t^2\right) \leq \sum_t \left(2\delta_t \frac{\ln 2}{\eta_t} + \delta_t\right) \leq 2\ln 2 \sum_t \frac{\delta_t}{\eta_t} + T\Delta \tag{5.3}$$

where we use the fact that $\delta_t \leq \max_{1\leq k\leq K}\{l_{t,k}\} - \min_{1\leq k\leq K}\{l_{t,k}\} = 1 - 0 = 1$. We will now show

$$\frac{\delta_t}{\eta_t} \leq \frac{1}{2} v_t + \frac{1}{3}\delta_t. \tag{5.4}$$

Note that we can rewrite Bernstein's bound as follows:

$$\frac{1}{2} v_t \geq \delta_t \cdot \frac{1}{2g(\eta_t)} = \frac{\delta_t}{\eta_t} - f(\eta_t)\delta_t \tag{5.5}$$

where $f(x) = (e^x - \frac{1}{2}x^2 - x - 1)/(xe^x - x^2 - x)$. Remains to show that $f(x) \leq 1/3$ for all $x \geq 0$. After rearranging we find this to be the case if

$$(3 - x)e^x \leq \frac{1}{2}x^2 - 2x + 3 \tag{5.6}$$

Taylor expansion of the left-hand side around zero reveals that $(3-x)e^x = \frac{1}{2}x^2 + 2x + 3 - \frac{1}{6}x^3 u e^u$ for some $0 \leq u \leq x$, from which the result follows.

**Remark 5.2.** Remark 5.1 is of the form

$$\Delta^2 \leq a + b\Delta, \tag{5.7}$$

with $a$ and $b$ nonnegative numbers. Solving for $\Delta$ then gives

$$\Delta \leq \frac{1}{2}b + \frac{1}{2}\sqrt{b^2 + 4a} \leq \frac{1}{2}b + \frac{1}{2}\left(\sqrt{b^2} + \sqrt{4a}\right) = \sqrt{a} + b \tag{5.8}$$

which implies

$$R \leq 2\sqrt{a} + 2b. \tag{5.9}$$

Plugging in the values $a = V$ and $b = T\left(\frac{2}{3}\ln 2 + 1\right)$ completes the proof.

**Remark 5.3.** We have shown that $v_t \leq (1 - h_t)h_t$. Now

$$V \leq \sum_{t=1}^{T} v_t \leq \sum_t (1 - h_t)h_t \leq T \sum_t \frac{(1-h_t)h_t}{1} = T^2 \sum_t \frac{1}{T}\frac{(1-h_t)h_t}{(1-h_t) + h_t} \leq T\frac{(T-H)H}{T} \tag{5.10}$$

where the last inequality is an instance of Jensen's inequality applied to the function $B$ defined on the domain $x, y \geq 0$ by $B(x, y) = \frac{xy}{x+y}$ and $B(x, y) = 0$ for $xy = 0$ to ensure continuity. We can verify that $B(x, y)$ is jointly concave by computing the Hessian and observe that the Hessian is negative semi-definite. Subsequently using $H \geq L^*$ and $H \leq L^* + 2\Delta$ yields

$$\frac{(T-H)H}{T} \leq \frac{(T-L^*)(L^*+2\Delta)}{T} \leq \frac{(T-L^*)L^*}{T} + 2\Delta \tag{5.11}$$

as desired.

## References

1. Steven De Rooij, Tim Van Erven, Peter D Grünwald, and Wouter M Koolen, *Follow the leader if you can, hedge if you must*, The Journal of Machine Learning Research **15** (2014), no. 1, 1281–1316.
2. Tim V Erven, Wouter M Koolen, Steven D Rooij, and Peter Grünwald, *Adaptive hedge*, Advances in Neural Information Processing Systems, 2011, pp. 1656–1664.