

Sequential Learning Algorithms

D Manjunath & Jayakrishnan Nair

EE, IIT Bombay

January 13, 2022

Outline

1 Predicting an arbitrary binary sequence with experts:

- There is at least one perfect expert
- Algorithm: Weighted Majority Algorithm (WMA)
- Predicting an arbitrary real sequence with experts

Outline

1 Predicting an arbitrary binary sequence with experts:

- There is at least one perfect expert *Majority Algorithm*

■ All experts are imperfect, some more than others. *Weighted Majority Algorithm (WMA)*

2 Predicting an arbitrary real sequence with experts

Outline

1 Predicting an arbitrary binary sequence with experts:

- There is at least one perfect expert *Majority Algorithm*

■ All experts are imperfect, some more than others. *Weighted Majority Algorithm (WMA)*

2 Predicting an arbitrary real sequence with experts

Outline

- 1 Predicting an arbitrary binary sequence with experts:
 - There is at least one perfect expert *Majority Algorithm*
 - All experts are imperfect, some more than others. *Weighted Majority Algorithm (WMA)*
- 2 Predicting an arbitrary real sequence with experts

Outline

- 1 Predicting an arbitrary binary sequence with experts:
 - There is at least one perfect expert *Majority Algorithm*
 - All experts are imperfect, some more than others. *Weighted Majority Algorithm (WMA)*
- 2 Predicting an arbitrary real sequence with experts

Outline

- 1 Predicting an arbitrary binary sequence with experts:
 - There is at least one perfect expert *Majority Algorithm*
 - All experts are imperfect, some more than others. *Weighted Majority Algorithm* (WMA)
- 2 Predicting an arbitrary real sequence with experts
 - Deterministic Algorithm: Exponential WMA (EWMA)
 - Randomized WMA, aka *Hedge*

Outline

- 1 Predicting an arbitrary binary sequence with experts:
 - There is at least one perfect expert *Majority Algorithm*
 - All experts are imperfect, some more than others. *Weighted Majority Algorithm* (WMA)
- 2 Predicting an arbitrary real sequence with experts
 - Deterministic Algorithm: Exponential WMA (EWMA)
 - Randomized WMA, aka *Hedge*

Outline

- 1 Predicting an arbitrary binary sequence with experts:
 - There is at least one perfect expert *Majority Algorithm*
 - All experts are imperfect, some more than others. *Weighted Majority Algorithm* (WMA)
- 2 Predicting an arbitrary real sequence with experts
 - Deterministic Algorithm: Exponential WMA (EWMA)
 - Randomized WMA, aka *Hedge*

Predicting with experts: Preliminaries

Illustration

Predicting with experts: Preliminaries

Illustration

Predicting with experts: Preliminaries

Illustration

Predicting binary sequence with experts: Preliminaries

- As before, $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \dots$
- However, now we will not make any assumptions about the statistical nature of the data, i.e., the $\{X_t\}$ is an *arbitrary* sequence.
- We now have access to K experts with expert i giving the prediction $Y_{i,t}$.
- Using the history of the data sequence and the experts' predictions,

$$H_t = \{Y_{1,1}, \dots, Y_{K,1}, X_1, \dots, Y_{1,t-1}, \dots, Y_{K,t-1}, X_{t-1}\}$$

the algorithm will determine the prediction \hat{X}_t .

- The true value is revealed after the prediction is made.
- At least one of the K experts is *perfect* and does not make a mistake.
- The objective of the algorithm is to discover the perfect expert quickly without making too many mistakes.

Predicting binary sequence with experts: Preliminaries

- As before, $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \dots$
- However, now we will not make any assumptions about the statistical nature of the data, i.e., the $\{X_t\}$ is an *arbitrary* sequence.
- We now have access to K experts with expert i giving the prediction $Y_{i,t}$.
- Using the history of the data sequence and the experts' predictions,

$$H_t = \{Y_{1,1}, \dots, Y_{K,1}, X_1, \dots, Y_{1,t-1}, \dots, Y_{K,t-1}, X_{t-1}\}$$

the algorithm will determine the prediction \hat{X}_t .

- The true value is revealed after the prediction is made.
- At least one of the K experts is *perfect* and does not make a mistake.
- The algorithm can ask the experts for their predictions.
- The algorithm can ask the experts for their predictions.
- The algorithm can ask the experts for their predictions.
- The algorithm can ask the experts for their predictions.

Predicting binary sequence with experts: Preliminaries

- As before, $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \dots$
- However, now we will not make any assumptions about the statistical nature of the data, i.e., the $\{X_t\}$ is an *arbitrary* sequence.
- We now have access to K experts with expert i giving the prediction $Y_{i,t}$.
- Using the history of the data sequence and the experts' predictions,

$$H_t = \{Y_{1,1}, \dots, Y_{K,1}, X_1, \dots, Y_{1,t-1}, \dots, Y_{K,t-1}, X_{t-1}\}$$

the algorithm will determine the prediction \hat{X}_t .

- The true value is revealed after the prediction is made.
- At least one of the K experts is *perfect* and does not make a mistake.
- The algorithm must be able to learn from the experts' predictions.
- The algorithm must be able to learn from the experts' predictions.
- The algorithm must be able to learn from the experts' predictions.

Predicting binary sequence with experts: Preliminaries

- As before, $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \dots$
- However, now we will not make any assumptions about the statistical nature of the data, i.e., the $\{X_t\}$ is an *arbitrary* sequence.
- We now have access to K experts with expert i giving the prediction $Y_{i,t}$.
- Using the history of the data sequence and the experts' predictions,

$$H_t = \{Y_{1,1}, \dots, Y_{K,1}, X_1, \dots, Y_{1,t-1}, \dots, Y_{K,t-1}, X_{t-1}\}$$

the algorithm will determine the prediction \hat{X}_t .

- The true value is revealed after the prediction is made.
- At least one of the K experts is *perfect* and does not make a mistake.

Predicting binary sequence with experts: Preliminaries

- As before, $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \dots$
- However, now we will not make any assumptions about the statistical nature of the data, i.e., the $\{X_t\}$ is an *arbitrary* sequence.
- We now have access to K experts with expert i giving the prediction $Y_{i,t}$.
- Using the history of the data sequence and the experts' predictions,

$$H_t = \{Y_{1,1}, \dots, Y_{K,1}, X_1, \dots, Y_{1,t-1}, \dots, Y_{K,t-1}, X_{t-1}\}$$

the algorithm will determine the prediction \hat{X}_t .

- The true value is revealed after the prediction is made.
- At least one of the K experts is *perfect* and does not make a mistake.

Predicting binary sequence with experts: Preliminaries

- As before, $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \dots$
- However, now we will not make any assumptions about the statistical nature of the data, i.e., the $\{X_t\}$ is an *arbitrary* sequence.
- We now have access to K experts with expert i giving the prediction $Y_{i,t}$.
- Using the history of the data sequence and the experts' predictions,

$$H_t = \{Y_{1,1}, \dots, Y_{K,1}, X_1, \dots, Y_{1,t-1}, \dots, Y_{K,t-1}, X_{t-1}\}$$

the algorithm will determine the prediction \hat{X}_t .

- The true value is revealed after the prediction is made.
- At least one of the K experts is *perfect* and does not make a mistake.
but this expert is not known to the algorithm.

Predicting binary sequence with experts: Preliminaries

- As before, $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \dots$
- However, now we will not make any assumptions about the statistical nature of the data, i.e., the $\{X_t\}$ is an *arbitrary* sequence.
- We now have access to K experts with expert i giving the prediction $Y_{i,t}$.
- Using the history of the data sequence and the experts' predictions,

$$H_t = \{Y_{1,1}, \dots, Y_{K,1}, X_1, \dots, Y_{1,t-1}, \dots, Y_{K,t-1}, X_{t-1}\}$$

the algorithm will determine the prediction \hat{X}_t .

- The true value is revealed after the prediction is made.
- At least one of the K experts is *perfect* and does not make a mistake.
but this expert is not known to the algorithm.
- The objective of the algorithm would be to discover the perfect expert quickly without making too many mistakes.

Predicting binary sequence with experts: Preliminaries

- As before, $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \dots$
- However, now we will not make any assumptions about the statistical nature of the data, i.e., the $\{X_t\}$ is an *arbitrary* sequence.
- We now have access to K experts with expert i giving the prediction $Y_{i,t}$.
- Using the history of the data sequence and the experts' predictions,

$$H_t = \{Y_{1,1}, \dots, Y_{K,1}, X_1, \dots, Y_{1,t-1}, \dots, Y_{K,t-1}, X_{t-1}\}$$

the algorithm will determine the prediction \hat{X}_t .

- The true value is revealed after the prediction is made.
- At least one of the K experts is *perfect* and does not make a mistake. but this expert is not known to the algorithm.
- The objective of the algorithm would be to discover the perfect expert quickly without making too many mistakes.

Predicting binary sequence with experts: Preliminaries

- As before, $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \dots$
- However, now we will not make any assumptions about the statistical nature of the data, i.e., the $\{X_t\}$ is an *arbitrary* sequence.
- We now have access to K experts with expert i giving the prediction $Y_{i,t}$.
- Using the history of the data sequence and the experts' predictions,

$$H_t = \{Y_{1,1}, \dots, Y_{K,1}, X_1, \dots, Y_{1,t-1}, \dots, Y_{K,t-1}, X_{t-1}\}$$

the algorithm will determine the prediction \hat{X}_t .

- The true value is revealed after the prediction is made.
- At least one of the K experts is *perfect* and does not make a mistake. but this expert is not known to the algorithm.
- The objective of the algorithm would be to discover the perfect expert quickly without making too many mistakes.

Predicting binary sequence with experts: Majority Algorithm (MA)

- Consider the *majority* algorithm (MA)

- $w_{i,t}$ is the weight for expert i
- $w_{i,1} = 1$ for $i = 1, \dots, K$
- At time t ,

■ $w_{i,t} = \prod_{s=1}^{t-1} (1 - \eta \ell_{i,s})$ for $i = 1, \dots, K$ (where η is a constant)

■ MA chooses the expert i with the highest weight $w_{i,t}$ at time t .

■ MA makes a prediction \hat{y}_t based on the chosen expert's prediction.

■ The total loss of MA is $\sum_{t=1}^T \ell_{i_t,t}$ where i_t is the chosen expert at time t .

- **Claim:** MA will make at most $\log_2 K$ mistakes.

Predicting binary sequence with experts: Majority Algorithm (MA)

- Consider the *majority* algorithm (MA)

- $w_{i,t}$ is the weight for expert i

- $w_{i,1} = 1$ for $i = 1, \dots, K$

- At time t ,

- **Claim:** MA will make at most $\log_2 K$ mistakes.

Predicting binary sequence with experts: Majority Algorithm (MA)

- Consider the *majority* algorithm (MA)

- $w_{i,t}$ is the weight for expert i

- $w_{i,1} = 1$ for $i = 1, \dots, K$

- At time t ,

- Let $P_t = \{Y_{i,t} : w_{i,t} = 1\}$, i.e., the set of predictions from experts with $w_{i,t} = 1$

- MA predicts the majority element of P_t

- If P_t contains no majority element, MA chooses $Y_{i,t}$ arbitrarily

- If $P_t = \{1, \dots, K\}$, then set $w_{i,t+1} = 0$

- **Claim:** MA will make at most $\log_2 K$ mistakes.

Predicting binary sequence with experts: Majority Algorithm (MA)

- Consider the *majority* algorithm (MA)
 - $w_{i,t}$ is the weight for expert i
 - $w_{i,1} = 1$ for $i = 1, \dots, K$
 - At time t ,
 - Let $P_t = \{Y_{i,t} : w_{i,t} = 1\}$, i.e., the set of predictions from experts with $w_{i,t} = 1$
 - \hat{X}_t = Majority prediction from set P_t
 - Receive the true value X_t chosen by the environment
 - If $w_{i,t} = 1$, and $X_t \neq Y_{i,t}$ then set $w_{i,t+1} = 0$
- **Claim:** MA will make at most $\log_2 K$ mistakes.

Predicting binary sequence with experts: Majority Algorithm (MA)

- Consider the *majority* algorithm (MA)
 - $w_{i,t}$ is the weight for expert i
 - $w_{i,1} = 1$ for $i = 1, \dots, K$
 - At time t ,
 - Let $P_t = \{Y_{i,t} : w_{i,t} = 1\}$, i.e., the set of predictions from experts with $w_{i,t} = 1$
 - \hat{X}_t = Majority prediction from set P_t
 - Receive the true value X_t chosen by the environment
 - If $w_{i,t} = 1$, and $X_t \neq Y_{i,t}$ then set $w_{i,t+1} = 0$
- **Claim:** MA will make at most $\log_2 K$ mistakes.

Predicting binary sequence with experts: Majority Algorithm (MA)

- Consider the *majority* algorithm (MA)
 - $w_{i,t}$ is the weight for expert i
 - $w_{i,1} = 1$ for $i = 1, \dots, K$
 - At time t ,
 - Let $P_t = \{Y_{i,t} : w_{i,t} = 1\}$, i.e., the set of predictions from experts with $w_{i,t} = 1$
 - \hat{X}_t = Majority prediction from set P_t
 - Receive the true value X_t chosen by the environment
 - If $w_{i,t} = 1$, and $X_t \neq Y_{i,t}$ then set $w_{i,t+1} = 0$
- **Claim:** MA will make at most $\log_2 K$ mistakes.

Predicting binary sequence with experts: Majority Algorithm (MA)

- Consider the *majority* algorithm (MA)
 - $w_{i,t}$ is the weight for expert i
 - $w_{i,1} = 1$ for $i = 1, \dots, K$
 - At time t ,
 - Let $P_t = \{Y_{i,t} : w_{i,t} = 1\}$, i.e., the set of predictions from experts with $w_{i,t} = 1$
 - \hat{X}_t = Majority prediction from set P_t
 - Receive the true value X_t chosen by the environment
 - If $w_{i,t} = 1$, and $X_t \neq Y_{i,t}$ then set $w_{i,t+1} = 0$
- **Claim:** MA will make at most $\log_2 K$ mistakes.

Predicting binary sequence with experts: Majority Algorithm (MA)

- Consider the *majority* algorithm (MA)
 - $w_{i,t}$ is the weight for expert i
 - $w_{i,1} = 1$ for $i = 1, \dots, K$
 - At time t ,
 - Let $P_t = \{Y_{i,t} : w_{i,t} = 1\}$, i.e., the set of predictions from experts with $w_{i,t} = 1$
 - \hat{X}_t = Majority prediction from set P_t
 - Receive the true value X_t chosen by the environment
 - If $w_{i,t} = 1$, and $X_t \neq Y_{i,t}$ then set $w_{i,t+1} = 0$
- **Claim:** MA will make at most $\log_2 K$ mistakes.

Predicting binary sequence with experts: Majority Algorithm (MA)

- Consider the *majority* algorithm (MA)
 - $w_{i,t}$ is the weight for expert i
 - $w_{i,1} = 1$ for $i = 1, \dots, K$
 - At time t ,
 - Let $P_t = \{Y_{i,t} : w_{i,t} = 1\}$, i.e., the set of predictions from experts with $w_{i,t} = 1$
 - \hat{X}_t = Majority prediction from set P_t
 - Receive the true value X_t chosen by the environment
 - If $w_{i,t} = 1$, and $X_t \neq Y_{i,t}$ then set $w_{i,t+1} = 0$
- **Claim:** MA will make at most $\log_2 K$ mistakes.

Analysis of Majority Algorithm

- Let $W_t = \sum_{i=1}^K w_{i,t}$ be the number of experts that are contributing to \hat{X}_t at time t .
- At time t , if MA makes a mistake, at least half of the imperfect experts are eliminated; W_t decreases multiplicatively after every wrong prediction by MA.
- Let L_t be the number of mistakes upto time t
- If $\hat{X}_t \neq X_t$, $W_{t+1} \leq W_t/2$.
- K can be halved at most $\log_2 K$ times, hence there are at most $L_T \leq \min\{T, \log_2 K\}$ for all T , mistakes made by the algorithm.
- The *loss is a constant* and does not depend on time T .

Analysis of Majority Algorithm

- Let $W_t = \sum_{i=1}^K w_{i,t}$ be the number of experts that are contributing to \hat{X}_t at time t .
- At time t , if MA makes a mistake, at least half of the imperfect experts are eliminated; W_t decreases multiplicatively after every wrong prediction by MA.
- Let L_t be the number of mistakes upto time t
- If $\hat{X}_t \neq X_t$, $W_{t+1} \leq W_t/2$.
- K can be halved at most $\log_2 K$ times, hence there are at most $L_T \leq \min\{T, \log_2 K\}$ for all T , mistakes made by the algorithm.
- The *loss is a constant* and does not depend on time T .

Analysis of Majority Algorithm

- Let $W_t = \sum_{i=1}^K w_{i,t}$ be the number of experts that are contributing to \hat{X}_t at time t .
- At time t , if MA makes a mistake, at least half of the imperfect experts are eliminated; W_t decreases multiplicatively after every wrong prediction by MA.
- Let L_t be the number of mistakes upto time t
 - If $\hat{X}_t \neq X_t$, $W_{t+1} \leq W_t/2$.
 - K can be halved at most $\log_2 K$ times, hence there are at most $L_T \leq \min\{T, \log_2 K\}$ for all T , mistakes made by the algorithm.
 - The *loss is a constant* and does not depend on time T .

Analysis of Majority Algorithm

- Let $W_t = \sum_{i=1}^K w_{i,t}$ be the number of experts that are contributing to \hat{X}_t at time t .
- At time t , if MA makes a mistake, at least half of the imperfect experts are eliminated; W_t decreases multiplicatively after every wrong prediction by MA.
- Let L_t be the number of mistakes upto time t
- If $\hat{X}_t \neq X_t$, $W_{t+1} \leq W_t/2$.
- K can be halved at most $\log_2 K$ times, hence there are at most $L_T \leq \min\{T, \log_2 K\}$ for all T , mistakes made by the algorithm.
- The *loss is a constant* and does not depend on time T .

Analysis of Majority Algorithm

- Let $W_t = \sum_{i=1}^K w_{i,t}$ be the number of experts that are contributing to \hat{X}_t at time t .
- At time t , if MA makes a mistake, at least half of the imperfect experts are eliminated; W_t decreases multiplicatively after every wrong prediction by MA.
- Let L_t be the number of mistakes upto time t
- If $\hat{X}_t \neq X_t$, $W_{t+1} \leq W_t/2$.
- K can be halved at most $\log_2 K$ times, hence there are at most $L_T \leq \min\{T, \log_2 K\}$ for all T , mistakes made by the algorithm.
- The *loss is a constant* and does not depend on time T .

Analysis of Majority Algorithm

- Let $W_t = \sum_{i=1}^K w_{i,t}$ be the number of experts that are contributing to \hat{X}_t at time t .
- At time t , if MA makes a mistake, at least half of the imperfect experts are eliminated; W_t decreases multiplicatively after every wrong prediction by MA.
- Let L_t be the number of mistakes upto time t
- If $\hat{X}_t \neq X_t$, $W_{t+1} \leq W_t/2$.
- K can be halved at most $\log_2 K$ times, hence there are at most $L_T \leq \min\{T, \log_2 K\}$ for all T , mistakes made by the algorithm.
- The *loss is a constant* and does not depend on time T .

Predicting binary sequence: all experts are imperfect

- Now assume that there is no perfect expert; every expert makes errors.
 - Hence, elimination will not work.
- However, there is a ‘best’ expert who has made fewest errors upto T .
- Thus the best that the algorithm can do is to match the best expert.
- Consider the *weighted majority* algorithm (WMA)

At each time step t , we give the weight for expert i as w_i and we have a forecast of the prediction \hat{y}_t as follows:

$$\hat{y}_t = \frac{\sum_{i=1}^n w_i y_{i,t}}{\sum_{i=1}^n w_i}$$

At time t ,

Predicting binary sequence: all experts are imperfect

- Now assume that there is no perfect expert; every expert makes errors.
 - Hence, elimination will not work.
- However, there is a ‘best’ expert who has made fewest errors upto T .
- Thus the best that the algorithm can do is to match the best expert.
- Consider the *weighted majority* algorithm (WMA)

At each time step t , we give each expert i a weight w_i depending on the number of errors that expert has made upto time t . We then predict the majority class.

Predicting binary sequence: all experts are imperfect

- Now assume that there is no perfect expert; every expert makes errors.
 - Hence, elimination will not work.
- However, there is a ‘best’ expert who has made fewest errors upto T .
- Thus the best that the algorithm can do is to match the best expert.
- Consider the *weighted majority* algorithm (WMA)

At each time step t , the algorithm chooses a subset of experts to follow, and weights them according to their past performance. The weights are updated at each time step, and the algorithm chooses the subset of experts to follow based on the updated weights.

Predicting binary sequence: all experts are imperfect

- Now assume that there is no perfect expert; every expert makes errors.
 - Hence, elimination will not work.
- However, there is a ‘best’ expert who has made fewest errors upto T .
- Thus the best that the algorithm can do is to match the best expert.
- Consider the *weighted majority* algorithm (WMA)

■ $w_{i,t}$ is the weight for expert i at time t . $w_{i,t}$ is a measure of its credibility.
■ Initially $w_{i,0} = 1$ for all experts.
■ When expert i makes an error at time t , its weight is updated as follows:

Predicting binary sequence: all experts are imperfect

- Now assume that there is no perfect expert; every expert makes errors.
 - Hence, elimination will not work.
- However, there is a ‘best’ expert who has made fewest errors upto T .
- Thus the best that the algorithm can do is to match the best expert.
- Consider the *weighted majority* algorithm (WMA)
 - $w_{i,t}$ is the weight for expert i at time t . $w_{i,t}$ is a measure of its credibility.
 - Initialize $w_{i,1} = 1$ for $i = 1, \dots, K$
 - At time t ,

Predicting binary sequence: all experts are imperfect

- Now assume that there is no perfect expert; every expert makes errors.
 - Hence, elimination will not work.
- However, there is a ‘best’ expert who has made fewest errors upto T .
- Thus the best that the algorithm can do is to match the best expert.
- Consider the *weighted majority* algorithm (WMA)
 - $w_{i,t}$ is the weight for expert i at time t . $w_{i,t}$ is a measure of its credibility.
 - Initialize $w_{i,1} = 1$ for $i = 1, \dots, K$
 - At time t ,

Predicting binary sequence: all experts are imperfect

- Now assume that there is no perfect expert; every expert makes errors.
 - Hence, elimination will not work.
- However, there is a ‘best’ expert who has made fewest errors upto T .
- Thus the best that the algorithm can do is to match the best expert.
- Consider the *weighted majority* algorithm (WMA)
 - $w_{i,t}$ is the weight for expert i at time t . $w_{i,t}$ is a measure of its credibility.
 - Initialize $w_{i,1} = 1$ for $i = 1, \dots, K$
 - At time t ,

 ■ Obtain sum of the weights of experts predicting 0 and those predicting 1.

$$W_{0,t} = \sum_{i=1}^K w_{i,t} x_{i,t,0}$$

$$W_{1,t} = \sum_{i=1}^K w_{i,t} x_{i,t,1}$$

Predicting binary sequence: all experts are imperfect

- Now assume that there is no perfect expert; every expert makes errors.
 - Hence, elimination will not work.
- However, there is a ‘best’ expert who has made fewest errors upto T .
- Thus the best that the algorithm can do is to match the best expert.
- Consider the *weighted majority* algorithm (WMA)
 - $w_{i,t}$ is the weight for expert i at time t . $w_{i,t}$ is a measure of its credibility.
 - Initialize $w_{i,1} = 1$ for $i = 1, \dots, K$
 - At time t ,
 - Obtain sum of the weights of experts predicting 0 and those predicting 1.

$$W_{0,t} = \sum_{i=1}^K w_{i,t} \mathcal{I}_{Y_{i,t}=0}$$

$$W_{1,t} = \sum_{i=1}^K w_{i,t} \mathcal{I}_{Y_{i,t}=1}$$

- Predict $\hat{X}_t = \mathcal{I}_{W_{1,t} > W_{0,t}}$;
- Receive the true value X_t chosen by the environment
- Update the weights: If $X_t \neq Y_{i,t}$ then $w_{i,t+1} = w_{i,t} \times (1 - \beta)$ where $0 < \beta < 1$.
- β is called the learning parameter; β closer to 0, means weights change slowly and learning is ‘slower and steadier.’

Predicting binary sequence: all experts are imperfect

- Now assume that there is no perfect expert; every expert makes errors.
 - Hence, elimination will not work.
- However, there is a ‘best’ expert who has made fewest errors upto T .
- Thus the best that the algorithm can do is to match the best expert.
- Consider the *weighted majority* algorithm (WMA)
 - $w_{i,t}$ is the weight for expert i at time t . $w_{i,t}$ is a measure of its credibility.
 - Initialize $w_{i,1} = 1$ for $i = 1, \dots, K$
 - At time t ,
 - Obtain sum of the weights of experts predicting 0 and those predicting 1.

$$W_{0,t} = \sum_{i=1}^K w_{i,t} \mathcal{I}_{Y_{i,t}=0}$$

$$W_{1,t} = \sum_{i=1}^K w_{i,t} \mathcal{I}_{Y_{i,t}=1}$$

- Predict $\hat{X}_t = \mathcal{I}_{W_{1,t} > W_{0,t}}$;
- Receive the true value X_t chosen by the environment
- Update the weights: If $X_t \neq Y_{i,t}$ then $w_{i,t+1} = w_{i,t} \times (1 - \beta)$ where $0 < \beta < 1$.
- β is called the learning parameter; β closer to 0, means weights change slowly and learning is ‘slower and steadier.’

Predicting binary sequence: all experts are imperfect

- Now assume that there is no perfect expert; every expert makes errors.
 - Hence, elimination will not work.
- However, there is a ‘best’ expert who has made fewest errors upto T .
- Thus the best that the algorithm can do is to match the best expert.
- Consider the *weighted majority* algorithm (WMA)
 - $w_{i,t}$ is the weight for expert i at time t . $w_{i,t}$ is a measure of its credibility.
 - Initialize $w_{i,1} = 1$ for $i = 1, \dots, K$
 - At time t ,
 - Obtain sum of the weights of experts predicting 0 and those predicting 1.

$$W_{0,t} = \sum_{i=1}^K w_{i,t} \mathcal{I}_{Y_{i,t}=0}$$

$$W_{1,t} = \sum_{i=1}^K w_{i,t} \mathcal{I}_{Y_{i,t}=1}$$

- Predict $\hat{X}_t = \mathcal{I}_{W_{1,t} > W_{0,t}}$;
- Receive the true value X_t chosen by the environment
- Update the weights: If $X_t \neq Y_{i,t}$ then $w_{i,t+1} = w_{i,t} \times (1 - \beta)$ where $0 < \beta < 1$.
- β is called the learning parameter; β closer to 0, means weights change slowly and learning is ‘slower and steadier.’

Predicting binary sequence: all experts are imperfect

- Now assume that there is no perfect expert; every expert makes errors.
 - Hence, elimination will not work.
- However, there is a ‘best’ expert who has made fewest errors upto T .
- Thus the best that the algorithm can do is to match the best expert.
- Consider the *weighted majority* algorithm (WMA)
 - $w_{i,t}$ is the weight for expert i at time t . $w_{i,t}$ is a measure of its credibility.
 - Initialize $w_{i,1} = 1$ for $i = 1, \dots, K$
 - At time t ,
 - Obtain sum of the weights of experts predicting 0 and those predicting 1.

$$W_{0,t} = \sum_{i=1}^K w_{i,t} \mathcal{I}_{Y_{i,t}=0}$$

$$W_{1,t} = \sum_{i=1}^K w_{i,t} \mathcal{I}_{Y_{i,t}=1}$$

- Predict $\hat{X}_t = \mathcal{I}_{W_{1,t} > W_{0,t}}$;
- Receive the true value X_t chosen by the environment
- Update the weights: If $X_t \neq Y_{i,t}$ then $w_{i,t+1} = w_{i,t} \times (1 - \beta)$ where $0 < \beta < 1$.
- β is called the learning parameter; β closer to 0, means weights change slowly and learning is ‘slower and steadier.’

Predicting binary sequence: all experts are imperfect

- Now assume that there is no perfect expert; every expert makes errors.
 - Hence, elimination will not work.
- However, there is a ‘best’ expert who has made fewest errors upto T .
- Thus the best that the algorithm can do is to match the best expert.
- Consider the *weighted majority* algorithm (WMA)
 - $w_{i,t}$ is the weight for expert i at time t . $w_{i,t}$ is a measure of its credibility.
 - Initialize $w_{i,1} = 1$ for $i = 1, \dots, K$
 - At time t ,
 - Obtain sum of the weights of experts predicting 0 and those predicting 1.

$$W_{0,t} = \sum_{i=1}^K w_{i,t} \mathcal{I}_{Y_{i,t}=0}$$

$$W_{1,t} = \sum_{i=1}^K w_{i,t} \mathcal{I}_{Y_{i,t}=1}$$

- Predict $\hat{X}_t = \mathcal{I}_{W_{1,t} > W_{0,t}}$;
- Receive the true value X_t chosen by the environment
- Update the weights: If $X_t \neq Y_{i,t}$ then $w_{i,t+1} = w_{i,t} \times (1 - \beta)$ where $0 < \beta < 1$.
- β is called the learning parameter; β closer to 0, means weights change slowly and learning is ‘slower and steadier.’

Predicting binary sequence: all experts are imperfect

- Now assume that there is no perfect expert; every expert makes errors.
 - Hence, elimination will not work.
- However, there is a ‘best’ expert who has made fewest errors upto T .
- Thus the best that the algorithm can do is to match the best expert.
- Consider the *weighted majority* algorithm (WMA)
 - $w_{i,t}$ is the weight for expert i at time t . $w_{i,t}$ is a measure of its credibility.
 - Initialize $w_{i,1} = 1$ for $i = 1, \dots, K$
 - At time t ,
 - Obtain sum of the weights of experts predicting 0 and those predicting 1.

$$W_{0,t} = \sum_{i=1}^K w_{i,t} \mathcal{I}_{Y_{i,t}=0} \qquad W_{1,t} = \sum_{i=1}^K w_{i,t} \mathcal{I}_{Y_{i,t}=1}$$

- Predict $\hat{X}_t = \mathcal{I}_{W_{1,t} > W_{0,t}}$;
- Receive the true value X_t chosen by the environment
- Update the weights: If $X_t \neq Y_{i,t}$ then $w_{i,t+1} = w_{i,t} \times (1 - \beta)$ where $0 < \beta < 1$.
- β is called the learning parameter; β closer to 0, means weights change slowly and learning is ‘slower and steadier.’

Analysis of Weighted Majority Algorithm (WMA)

Analysis of the *weighted majority* algorithm

- Define a potential function $W_t := \sum_{i=1}^K w_{i,t}$
- Let $L_{i,t}$ be the number of errors made by expert i upto time t and L_t be the number of errors made by WMA upto time t .
- Let $GW_t = \sum_i w_{i,t} \mathcal{I}_{X_t=Y_{i,t}}$ be the sum of the weights of experts predicting correctly in t and $BW_t = \sum_i w_{i,t} \mathcal{I}_{X_t \neq Y_{i,t}}$ be that of those that predicted wrongly at time t .
- Clearly, $W_{t+1} = GW_t + (1 - \beta)BW_t$. Also, since $0 < \beta < 1$, $W_{t+1} \leq W_t$.
- If $\hat{X}_t \neq X_t$ (prediction is wrong) then

$$GW_t \leq \frac{1}{2} W_t \quad (\text{because correct predictions are in minority})$$

$$W_{t+1} = GW_t + (1 - \beta)BW_t = GW_t + (1 - \beta)(W_t - GW_t)$$

$$= (1 - \beta)W_t + \beta GW_t \leq (1 - \beta)W_t + \beta W_t$$

$$= (1 - \beta + \beta)W_t = W_t$$

Analysis of Weighted Majority Algorithm (WMA)

Analysis of the *weighted majority* algorithm

- Define a potential function $W_t := \sum_{i=1}^K w_{i,t}$
- Let $L_{i,t}$ be the number of errors made by expert i upto time t and L_t be the number of errors made by WMA upto time t .
- Let $GW_t = \sum_i w_{i,t} \mathcal{I}_{X_t=Y_{i,t}}$ be the sum of the weights of experts predicting correctly in t and $BW_t = \sum_i w_{i,t} \mathcal{I}_{X_t \neq Y_{i,t}}$ be that of those that predicted wrongly at time t .
- Clearly, $W_{t+1} = GW_t + (1 - \beta)BW_t$. Also, since $0 < \beta < 1$, $W_{t+1} \leq W_t$.
- If $\hat{X}_t \neq X_t$ (prediction is wrong) then

$$GW_t \leq \frac{1}{2} W_t \quad (\text{because correct predictions are in minority})$$

Analysis of Weighted Majority Algorithm (WMA)

Analysis of the *weighted majority* algorithm

- Define a potential function $W_t := \sum_{i=1}^K w_{i,t}$
- Let $L_{i,t}$ be the number of errors made by expert i upto time t and L_t be the number of errors made by WMA upto time t .
- Let $GW_t = \sum_i w_{i,t} \mathcal{I}_{X_t=Y_{i,t}}$ be the sum of the weights of experts predicting correctly in t and $BW_t = \sum_i w_{i,t} \mathcal{I}_{X_t \neq Y_{i,t}}$ be that of those that predicted wrongly at time t .
- Clearly, $W_{t+1} = GW_t + (1 - \beta)BW_t$. Also, since $0 < \beta < 1$, $W_{t+1} \leq W_t$.
- If $\hat{X}_t \neq X_t$ (prediction is wrong) then

$$GW_t \leq \frac{1}{2} W_t \quad (\text{because correct predictions are in minority})$$

Analysis of Weighted Majority Algorithm (WMA)

Analysis of the *weighted majority* algorithm

- Define a potential function $W_t := \sum_{i=1}^K w_{i,t}$
- Let $L_{i,t}$ be the number of errors made by expert i upto time t and L_t be the number of errors made by WMA upto time t .
- Let $GW_t = \sum_i w_{i,t} \mathcal{I}_{X_t=Y_{i,t}}$ be the sum of the weights of experts predicting correctly in t and $BW_t = \sum_i w_{i,t} \mathcal{I}_{X_t \neq Y_{i,t}}$ be that of those that predicted wrongly at time t .
- Clearly, $W_{t+1} = GW_t + (1 - \beta)BW_t$. Also, since $0 < \beta < 1$, $W_{t+1} \leq W_t$.
- If $\hat{X}_t \neq X_t$ (prediction is wrong) then

$$GW_t \leq \frac{1}{2} W_t \quad (\text{because correct predictions are in minority})$$

Analysis of Weighted Majority Algorithm (WMA)

Analysis of the *weighted majority* algorithm

- Define a potential function $W_t := \sum_{i=1}^K w_{i,t}$
- Let $L_{i,t}$ be the number of errors made by expert i upto time t and L_t be the number of errors made by WMA upto time t .
- Let $GW_t = \sum_i w_{i,t} \mathcal{I}_{X_t=Y_{i,t}}$ be the sum of the weights of experts predicting correctly in t and $BW_t = \sum_i w_{i,t} \mathcal{I}_{X_t \neq Y_{i,t}}$ be that of those that predicted wrongly at time t .
- Clearly, $W_{t+1} = GW_t + (1 - \beta)BW_t$. Also, since $0 < \beta < 1$, $W_{t+1} \leq W_t$.
- If $\hat{X}_t \neq X_t$ (prediction is wrong) then

$$GW_t \leq \frac{1}{2} W_t \quad (\text{because correct predictions are in minority})$$

$$W_{t+1} = GW_t + (1 - \beta)BW_t = GW_t + (1 - \beta)(W_t - GW_t)$$

$$= (1 - \beta)W_t + \beta GW_t \leq (1 - \beta)W_t + \frac{1}{2}\beta W_t$$

$$= \left(1 - \frac{\beta}{2}\right) W_t \leq \left(1 - \frac{\beta}{2}\right)^{L_t} W_1 = \left(1 - \frac{\beta}{2}\right)^{L_t} K$$

Analysis of Weighted Majority Algorithm (WMA)

Analysis of the *weighted majority* algorithm

- Define a potential function $W_t := \sum_{i=1}^K w_{i,t}$
- Let $L_{i,t}$ be the number of errors made by expert i upto time t and L_t be the number of errors made by WMA upto time t .
- Let $GW_t = \sum_i w_{i,t} \mathcal{I}_{X_t=Y_{i,t}}$ be the sum of the weights of experts predicting correctly in t and $BW_t = \sum_i w_{i,t} \mathcal{I}_{X_t \neq Y_{i,t}}$ be that of those that predicted wrongly at time t .
- Clearly, $W_{t+1} = GW_t + (1 - \beta)BW_t$. Also, since $0 < \beta < 1$, $W_{t+1} \leq W_t$.
- If $\hat{X}_t \neq X_t$ (prediction is wrong) then

$$GW_t \leq \frac{1}{2} W_t \quad (\text{because correct predictions are in minority})$$

$$W_{t+1} = GW_t + (1 - \beta)BW_t = GW_t + (1 - \beta)(W_t - GW_t)$$

$$= (1 - \beta)W_t + \beta GW_t \leq (1 - \beta)W_t + \frac{1}{2}\beta W_t$$

$$= \left(1 - \frac{\beta}{2}\right) W_t \leq \left(1 - \frac{\beta}{2}\right)^{L_t} W_1 \leq \left(1 - \frac{\beta}{2}\right)^{L_t} K$$

Analysis of Weighted Majority Algorithm (WMA)

Analysis of the *weighted majority* algorithm

- Define a potential function $W_t := \sum_{i=1}^K w_{i,t}$
- Let $L_{i,t}$ be the number of errors made by expert i upto time t and L_t be the number of errors made by WMA upto time t .
- Let $GW_t = \sum_i w_{i,t} \mathcal{I}_{X_t=Y_{i,t}}$ be the sum of the weights of experts predicting correctly in t and $BW_t = \sum_i w_{i,t} \mathcal{I}_{X_t \neq Y_{i,t}}$ be that of those that predicted wrongly at time t .
- Clearly, $W_{t+1} = GW_t + (1 - \beta)BW_t$. Also, since $0 < \beta < 1$, $W_{t+1} \leq W_t$.
- If $\hat{X}_t \neq X_t$ (prediction is wrong) then

$$GW_t \leq \frac{1}{2} W_t \quad (\text{because correct predictions are in minority})$$

$$W_{t+1} = GW_t + (1 - \beta)BW_t = GW_t + (1 - \beta)(W_t - GW_t)$$

$$= (1 - \beta)W_t + \beta GW_t \leq (1 - \beta)W_t + \frac{1}{2}\beta W_t$$

$$= \left(1 - \frac{\beta}{2}\right) W_t \leq \left(1 - \frac{\beta}{2}\right)^{L_t} W_1 = \left(1 - \frac{\beta}{2}\right)^{L_t} K$$

Analysis of Weighted Majority Algorithm (WMA)

Analysis of the *weighted majority* algorithm

- Define a potential function $W_t := \sum_{i=1}^K w_{i,t}$
- Let $L_{i,t}$ be the number of errors made by expert i upto time t and L_t be the number of errors made by WMA upto time t .
- Let $GW_t = \sum_i w_{i,t} \mathcal{I}_{X_t=Y_{i,t}}$ be the sum of the weights of experts predicting correctly in t and $BW_t = \sum_i w_{i,t} \mathcal{I}_{X_t \neq Y_{i,t}}$ be that of those that predicted wrongly at time t .
- Clearly, $W_{t+1} = GW_t + (1 - \beta)BW_t$. Also, since $0 < \beta < 1$, $W_{t+1} \leq W_t$.
- If $\hat{X}_t \neq X_t$ (prediction is wrong) then

$$GW_t \leq \frac{1}{2} W_t \quad (\text{because correct predictions are in minority})$$

$$W_{t+1} = GW_t + (1 - \beta)BW_t = GW_t + (1 - \beta)(W_t - GW_t)$$

$$= (1 - \beta)W_t + \beta GW_t \leq (1 - \beta)W_t + \frac{1}{2}\beta W_t$$

$$= \left(1 - \frac{\beta}{2}\right) W_t \leq \left(1 - \frac{\beta}{2}\right)^{L_t} W_1 = \left(1 - \frac{\beta}{2}\right)^{L_t} K$$

Analysis of Weighted Majority Algorithm (WMA)

Analysis of the *weighted majority* algorithm

- Define a potential function $W_t := \sum_{i=1}^K w_{i,t}$
- Let $L_{i,t}$ be the number of errors made by expert i upto time t and L_t be the number of errors made by WMA upto time t .
- Let $GW_t = \sum_i w_{i,t} \mathcal{I}_{X_t=Y_{i,t}}$ be the sum of the weights of experts predicting correctly in t and $BW_t = \sum_i w_{i,t} \mathcal{I}_{X_t \neq Y_{i,t}}$ be that of those that predicted wrongly at time t .
- Clearly, $W_{t+1} = GW_t + (1 - \beta)BW_t$. Also, since $0 < \beta < 1$, $W_{t+1} \leq W_t$.
- If $\hat{X}_t \neq X_t$ (prediction is wrong) then

$$GW_t \leq \frac{1}{2} W_t \quad (\text{because correct predictions are in minority})$$

$$W_{t+1} = GW_t + (1 - \beta)BW_t = GW_t + (1 - \beta)(W_t - GW_t)$$

$$= (1 - \beta)W_t + \beta GW_t \leq (1 - \beta)W_t + \frac{1}{2}\beta W_t$$

$$= \left(1 - \frac{\beta}{2}\right) W_t \leq \left(1 - \frac{\beta}{2}\right)^{L_t} W_1 = \left(1 - \frac{\beta}{2}\right)^{L_t} K$$

Analysis of Weighted Majority Algorithm (WMA)

Analysis of the *weighted majority* algorithm

- Define a potential function $W_t := \sum_{i=1}^K w_{i,t}$
- Let $L_{i,t}$ be the number of errors made by expert i upto time t and L_t be the number of errors made by WMA upto time t .
- Let $GW_t = \sum_i w_{i,t} \mathcal{I}_{X_t=Y_{i,t}}$ be the sum of the weights of experts predicting correctly in t and $BW_t = \sum_i w_{i,t} \mathcal{I}_{X_t \neq Y_{i,t}}$ be that of those that predicted wrongly at time t .
- Clearly, $W_{t+1} = GW_t + (1 - \beta)BW_t$. Also, since $0 < \beta < 1$, $W_{t+1} \leq W_t$.
- If $\hat{X}_t \neq X_t$ (prediction is wrong) then

$$GW_t \leq \frac{1}{2} W_t \quad (\text{because correct predictions are in minority})$$

$$\begin{aligned} W_{t+1} &= GW_t + (1 - \beta)BW_t = GW_t + (1 - \beta)(W_t - GW_t) \\ &= (1 - \beta)W_t + \beta GW_t \leq (1 - \beta)W_t + \frac{1}{2}\beta W_t \end{aligned}$$

$$= \left(1 - \frac{\beta}{2}\right) W_t \leq \left(1 - \frac{\beta}{2}\right)^{L_t} W_1 = \left(1 - \frac{\beta}{2}\right)^{L_t} K$$

Analysis of Weighted Majority Algorithm (WMA)

Analysis of the *weighted majority* algorithm

- Define a potential function $W_t := \sum_{i=1}^K w_{i,t}$
- Let $L_{i,t}$ be the number of errors made by expert i upto time t and L_t be the number of errors made by WMA upto time t .
- Let $GW_t = \sum_i w_{i,t} \mathcal{I}_{X_t=Y_{i,t}}$ be the sum of the weights of experts predicting correctly in t and $BW_t = \sum_i w_{i,t} \mathcal{I}_{X_t \neq Y_{i,t}}$ be that of those that predicted wrongly at time t .
- Clearly, $W_{t+1} = GW_t + (1 - \beta)BW_t$. Also, since $0 < \beta < 1$, $W_{t+1} \leq W_t$.
- If $\hat{X}_t \neq X_t$ (prediction is wrong) then

$$GW_t \leq \frac{1}{2} W_t \quad (\text{because correct predictions are in minority})$$

$$W_{t+1} = GW_t + (1 - \beta)BW_t = GW_t + (1 - \beta)(W_t - GW_t)$$

$$= (1 - \beta)W_t + \beta GW_t \leq (1 - \beta)W_t + \frac{1}{2}\beta W_t$$

$$= \left(1 - \frac{\beta}{2}\right) W_t \leq \left(1 - \frac{\beta}{2}\right)^{L_t} W_1 = \left(1 - \frac{\beta}{2}\right)^{L_t} K$$

Analysis of Weighted Majority Algorithm (WMA)

Analysis of the *weighted majority* algorithm

- Define a potential function $W_t := \sum_{i=1}^K w_{i,t}$
- Let $L_{i,t}$ be the number of errors made by expert i upto time t and L_t be the number of errors made by WMA upto time t .
- Let $GW_t = \sum_i w_{i,t} \mathcal{I}_{X_t=Y_{i,t}}$ be the sum of the weights of experts predicting correctly in t and $BW_t = \sum_i w_{i,t} \mathcal{I}_{X_t \neq Y_{i,t}}$ be that of those that predicted wrongly at time t .
- Clearly, $W_{t+1} = GW_t + (1 - \beta)BW_t$. Also, since $0 < \beta < 1$, $W_{t+1} \leq W_t$.
- If $\hat{X}_t \neq X_t$ (prediction is wrong) then

$$GW_t \leq \frac{1}{2} W_t \quad (\text{because correct predictions are in minority})$$

$$W_{t+1} = GW_t + (1 - \beta)BW_t = GW_t + (1 - \beta)(W_t - GW_t)$$

$$= (1 - \beta)W_t + \beta GW_t \leq (1 - \beta)W_t + \frac{1}{2}\beta W_t$$

$$= \left(1 - \frac{\beta}{2}\right) W_t \leq \left(1 - \frac{\beta}{2}\right)^{L_t} W_1 = \left(1 - \frac{\beta}{2}\right)^{L_t} K$$

Analysis of Weighted Majority Algorithm (WMA)

Analysis of the *weighted majority* algorithm

- Define a potential function $W_t := \sum_{i=1}^K w_{i,t}$
- Let $L_{i,t}$ be the number of errors made by expert i upto time t and L_t be the number of errors made by WMA upto time t .
- Let $GW_t = \sum_i w_{i,t} \mathcal{I}_{X_t=Y_{i,t}}$ be the sum of the weights of experts predicting correctly in t and $BW_t = \sum_i w_{i,t} \mathcal{I}_{X_t \neq Y_{i,t}}$ be that of those that predicted wrongly at time t .
- Clearly, $W_{t+1} = GW_t + (1 - \beta)BW_t$. Also, since $0 < \beta < 1$, $W_{t+1} \leq W_t$.
- If $\hat{X}_t \neq X_t$ (prediction is wrong) then

$$GW_t \leq \frac{1}{2} W_t \quad (\text{because correct predictions are in minority})$$

$$W_{t+1} = GW_t + (1 - \beta)BW_t = GW_t + (1 - \beta)(W_t - GW_t)$$

$$= (1 - \beta)W_t + \beta GW_t \leq (1 - \beta)W_t + \frac{1}{2}\beta W_t$$

$$= \left(1 - \frac{\beta}{2}\right) W_t \leq \left(1 - \frac{\beta}{2}\right)^{L_t} W_1 = \left(1 - \frac{\beta}{2}\right)^{L_t} K$$

Analysis of Weighted Majority Algorithm (WMA)

- Just obtained an upper bound on W_{T+1} ; Following lower bound is easy.
For $i \in \{1, \dots, K\}$,

$$W_{T+1} \geq w_{i,T+1} = (1 - \beta)^{L_{i,T}}$$

- **Useful Inequality:** For $x > 0$, $\ln x \leq x - 1$.

Analysis of Weighted Majority Algorithm (WMA)

- Just obtained an upper bound on W_{T+1} ; Following lower bound is easy.
For $i \in \{1, \dots, K\}$,

$$W_{T+1} \geq w_{i,T+1} = (1 - \beta)^{L_{i,T}}$$

- **Useful Inequality:** For $x > 0$, $\ln x \leq x - 1$. This also implies $\ln(1 - x) < -x$ for $0 < x < 1$.
- Thus, for any T ,

$$\left(1 - \frac{\beta}{2}\right)^K \geq W_{T+1} \geq (1 - \beta)^{L_{i,T}}$$

$$\begin{aligned} & \rightarrow \ln\left(1 - \frac{\beta}{2}\right) \geq \ln W_{T+1} \geq \ln(1 - \beta) L_{i,T} \\ & \rightarrow L_{i,T} \leq \frac{\ln\left(1 - \frac{\beta}{2}\right)}{\ln(1 - \beta)} \leq \frac{\ln\left(1 - \frac{\beta}{2}\right)}{-\beta/2} \leq \frac{2}{\beta} \ln\left(\frac{2}{1 - \beta}\right) \end{aligned}$$

Analysis of Weighted Majority Algorithm (WMA)

- Just obtained an upper bound on W_{T+1} ; Following lower bound is easy.
For $i \in \{1, \dots, K\}$,

$$W_{T+1} \geq w_{i,T+1} = (1 - \beta)^{L_{i,T}}$$

- **Useful Inequality:** For $x > 0$, $\ln x \leq x - 1$. This also implies $\ln(1 - x) < -x$ for $0 < x < 1$.
- Thus, for any T ,

$$\left(1 - \frac{\beta}{2}\right)^K \geq W_{T+1} \geq (1 - \beta)^{L_{i,T}}$$

$$\begin{aligned} & \Rightarrow \ln\left(1 - \frac{\beta}{2}\right) \geq \ln W_{T+1} \geq \ln(1 - \beta) L_{i,T} \\ & \Rightarrow L_{i,T} \leq \frac{\ln\left(1 - \frac{\beta}{2}\right)}{\ln(1 - \beta)} \leq \frac{\ln\left(1 - \frac{\beta}{2}\right)}{-\beta/2} \leq \frac{2}{\beta} \ln\left(\frac{2}{1 - \beta}\right) \end{aligned}$$

Analysis of Weighted Majority Algorithm (WMA)

- Just obtained an upper bound on W_{T+1} ; Following lower bound is easy.
For $i \in \{1, \dots, K\}$,

$$W_{T+1} \geq w_{i,T+1} = (1 - \beta)^{L_{i,T}}$$

- **Useful Inequality:** For $x > 0$, $\ln x \leq x - 1$. This also implies $\ln(1 - x) < -x$ for $0 < x < 1$.
- Thus, for any T ,

$$\left(1 - \frac{\beta}{2}\right)^{L_T} K \geq W_{T+1} \geq (1 - \beta)^{L_{i,T}}$$

$$\implies L_{i,T} \ln(1 - \beta) - \ln K \leq L_T \ln \left(1 - \frac{\beta}{2}\right) \leq L_T (-\beta/2)$$

$$\implies L_T \leq -\frac{2}{\beta} L_{i,T} \ln \left(\frac{1}{1 - \beta}\right) + (2/\beta) \ln K$$

$$\implies L_T \leq \frac{2}{1 - \beta} L_{i,T} + \frac{2}{\beta} \ln K$$

Analysis of Weighted Majority Algorithm (WMA)

- Just obtained an upper bound on W_{T+1} ; Following lower bound is easy.
For $i \in \{1, \dots, K\}$,

$$W_{T+1} \geq w_{i,T+1} = (1 - \beta)^{L_{i,T}}$$

- **Useful Inequality:** For $x > 0$, $\ln x \leq x - 1$. This also implies $\ln(1 - x) < -x$ for $0 < x < 1$.
- Thus, for any T ,

$$\left(1 - \frac{\beta}{2}\right)^{L_T} K \geq W_{T+1} \geq (1 - \beta)^{L_{i,T}}$$

$$\implies L_{i,T} \ln(1 - \beta) - \ln K \leq L_T \ln \left(1 - \frac{\beta}{2}\right) \leq L_T (-\beta/2)$$

$$\implies L_T \leq -\frac{2}{\beta} L_{i,T} \ln \left(\frac{1}{1 - \beta}\right) + (2/\beta) \ln K$$

$$\implies L_T \leq \frac{2}{1 - \beta} L_{i,T} + \frac{2}{\beta} \ln K$$

Analysis of Weighted Majority Algorithm (WMA)

- Just obtained an upper bound on W_{T+1} ; Following lower bound is easy.
For $i \in \{1, \dots, K\}$,

$$W_{T+1} \geq w_{i,T+1} = (1 - \beta)^{L_{i,T}}$$

- **Useful Inequality:** For $x > 0$, $\ln x \leq x - 1$. This also implies $\ln(1 - x) < -x$ for $0 < x < 1$.
- Thus, for any T ,

$$\left(1 - \frac{\beta}{2}\right)^{L_T} K \geq W_{T+1} \geq (1 - \beta)^{L_{i,T}}$$

$$\implies L_{i,T} \ln(1 - \beta) - \ln K \leq L_T \ln \left(1 - \frac{\beta}{2}\right) \leq L_T (-\beta/2)$$

$$\implies L_T \leq -\frac{2}{\beta} L_{i,T} \ln \left(\frac{1}{1 - \beta}\right) + (2/\beta) \ln K$$

$$\implies L_T \leq \frac{2}{1 - \beta} L_{i,T} + \frac{2}{\beta} \ln K$$

Analysis of Weighted Majority Algorithm (WMA)

- Just obtained an upper bound on W_{T+1} ; Following lower bound is easy.
For $i \in \{1, \dots, K\}$,

$$W_{T+1} \geq w_{i,T+1} = (1 - \beta)^{L_{i,T}}$$

- **Useful Inequality:** For $x > 0$, $\ln x \leq x - 1$. This also implies $\ln(1 - x) < -x$ for $0 < x < 1$.
- Thus, for any T ,

$$\left(1 - \frac{\beta}{2}\right)^{L_T} K \geq W_{T+1} \geq (1 - \beta)^{L_{i,T}}$$

$$\implies L_{i,T} \ln(1 - \beta) - \ln K \leq L_T \ln \left(1 - \frac{\beta}{2}\right) \leq L_T (-\beta/2)$$

$$\implies L_T \leq -\frac{2}{\beta} L_{i,T} \ln \left(\frac{1}{1 - \beta}\right) + (2/\beta) \ln K$$

$$\implies L_T \leq \frac{2}{1 - \beta} L_{i,T} + \frac{2}{\beta} \ln K$$

Analysis of Weighted Majority Algorithm (WMA)

- Just obtained an upper bound on W_{T+1} ; Following lower bound is easy.
For $i \in \{1, \dots, K\}$,

$$W_{T+1} \geq w_{i,T+1} = (1 - \beta)^{L_{i,T}}$$

- **Useful Inequality:** For $x > 0$, $\ln x \leq x - 1$. This also implies $\ln(1 - x) < -x$ for $0 < x < 1$.
- Thus, for any T ,

$$\left(1 - \frac{\beta}{2}\right)^{L_T} K \geq W_{T+1} \geq (1 - \beta)^{L_{i,T}}$$

$$\implies L_{i,T} \ln(1 - \beta) - \ln K \leq L_T \ln \left(1 - \frac{\beta}{2}\right) \leq L_T (-\beta/2)$$

$$\implies L_T \leq -\frac{2}{\beta} L_{i,T} \ln \left(\frac{1}{1 - \beta}\right) + (2/\beta) \ln K$$

$$\implies L_T \leq \frac{2}{1 - \beta} L_{i,T} + \frac{2}{\beta} \ln K$$

Analysis of Weighted Majority Algorithm (WMA)

- Just obtained an upper bound on W_{T+1} ; Following lower bound is easy.
For $i \in \{1, \dots, K\}$,

$$W_{T+1} \geq w_{i,T+1} = (1 - \beta)^{L_{i,T}}$$

- **Useful Inequality:** For $x > 0$, $\ln x \leq x - 1$. This also implies $\ln(1 - x) < -x$ for $0 < x < 1$.
- Thus, for any T ,

$$\left(1 - \frac{\beta}{2}\right)^{L_T} K \geq W_{T+1} \geq (1 - \beta)^{L_{i,T}}$$

$$\implies L_{i,T} \ln(1 - \beta) - \ln K \leq L_T \ln \left(1 - \frac{\beta}{2}\right) \leq L_T (-\beta/2)$$

$$\implies L_T \leq -\frac{2}{\beta} L_{i,T} \ln \left(\frac{1}{1 - \beta}\right) + (2/\beta) \ln K$$

$$\implies L_T \leq \frac{2}{1 - \beta} L_{i,T} + \frac{2}{\beta} \ln K$$

Analysis of Weighted Majority Algorithm (WMA)

- The last bound is true for all $i = 1, \dots, K$. In particular, it is true for best expert upto time T , i.e., one with least mistakes. Thus

$$L_T \leq \frac{2}{1-\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

If $\beta \leq 0.5$, then

$$L_T \leq \frac{2}{\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

Observe that the perfect expert case ($\min_i L_{i,t} = 0$ for all t) is recovered from the inequality.

- This implies that regret also depends on the quality of the best expert.
- Thus having imperfect experts can take you from constant loss to linear loss if the loss of the best expert is linear.
- If you know it, can you choose a good β as a function of T to make sure regret is linear in T ?

Analysis of Weighted Majority Algorithm (WMA)

- The last bound is true for all $i = 1, \dots, K$. In particular, it is true for best expert upto time T , i.e., one with least mistakes. Thus

$$L_T \leq \frac{2}{1-\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

If $\beta \leq 0.5$, then

$$L_T \leq \frac{2}{\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

Observe that the perfect expert case ($\min_i L_{i,t} = 0$ for all t) is recovered from the inequality.

- This implies that regret also depends on the quality of the best expert.
- Thus having imperfect experts can take you from constant loss to linear loss if the loss of the best expert is linear.
- If you know it, can you choose a good β as a function of T to make sure regret is $O(\sqrt{T})$?

Analysis of Weighted Majority Algorithm (WMA)

- The last bound is true for all $i = 1, \dots, K$. In particular, it is true for best expert upto time T , i.e., one with least mistakes. Thus

$$L_T \leq \frac{2}{1-\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

If $\beta \leq 0.5$, then

$$L_T \leq \frac{2}{\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

Observe that the perfect expert case ($\min_i L_{i,t} = 0$ for all t) is recovered from the inequality.

- This result is interesting because it depends on the quality of the best experts.
- Thus having imperfect experts can also not harm exponential loss to lower bound.
- If the bound of the best expert is known, we can choose β to minimize the bound.
- If the bound of the best expert is unknown, we can choose β to minimize the worst case bound.

Analysis of Weighted Majority Algorithm (WMA)

- The last bound is true for all $i = 1, \dots, K$. In particular, it is true for best expert upto time T , i.e., one with least mistakes. Thus

$$L_T \leq \frac{2}{1-\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

If $\beta \leq 0.5$, then

$$L_T \leq \frac{2}{\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

Observe that the perfect expert case ($\min_i L_{i,t} = 0$ for all t) is recovered from the inequality.

- This implies that regret loss depends on the quality of the best expert.
- Thus having improved upper bound, we now need to know how good is the quality of the best expert in terms of regret loss.
- Regret loss depends on the quality of the algorithm of J to make predictions.
- This is known as

Analysis of Weighted Majority Algorithm (WMA)

- The last bound is true for all $i = 1, \dots, K$. In particular, it is true for best expert upto time T , i.e., one with least mistakes. Thus

$$L_T \leq \frac{2}{1-\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

If $\beta \leq 0.5$, then

$$L_T \leq \frac{2}{\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

Observe that the perfect expert case ($\min_i L_{i,t} = 0$ for all t) is recovered from the inequality.

- This implies that regret/loss depends on the quality of the best expert.
- Thus having imperfect experts can take you from constant loss to linear loss if the loss of the best expert is linear.

- Regret/loss is a function of β to minimize the loss.

Analysis of Weighted Majority Algorithm (WMA)

- The last bound is true for all $i = 1, \dots, K$. In particular, it is true for best expert upto time T , i.e., one with least mistakes. Thus

$$L_T \leq \frac{2}{1-\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

If $\beta \leq 0.5$, then

$$L_T \leq \frac{2}{\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

Observe that the perfect expert case ($\min_i L_{i,t} = 0$ for all t) is recovered from the inequality.

- This implies that regretloss depends on the quality of the best expert.
- Thus having imperfect experts can take you from constant loss to linear loss if the loss of the best expert is linear.
- If you know T , can you choose a good β as a function of T to make loss sub linear in T . ?

Analysis of Weighted Majority Algorithm (WMA)

- The last bound is true for all $i = 1, \dots, K$. In particular, it is true for best expert upto time T , i.e., one with least mistakes. Thus

$$L_T \leq \frac{2}{1-\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

If $\beta \leq 0.5$, then

$$L_T \leq \frac{2}{\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

Observe that the perfect expert case ($\min_i L_{i,t} = 0$ for all t) is recovered from the inequality.

- This implies that regretloss depends on the quality of the best expert.
- Thus having imperfect experts can take you from constant loss to linear loss if the loss of the best expert is linear.
- If you know T , can you choose a good β as a function of T to make loss sub linear in T . ?

Analysis of Weighted Majority Algorithm (WMA)

- The last bound is true for all $i = 1, \dots, K$. In particular, it is true for best expert upto time T , i.e., one with least mistakes. Thus

$$L_T \leq \frac{2}{1-\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

If $\beta \leq 0.5$, then

$$L_T \leq \frac{2}{\beta} \left(\min_{1 \leq K} L_{i,T} \right) + \frac{2}{\beta} \ln K$$

Observe that the perfect expert case ($\min_i L_{i,t} = 0$ for all t) is recovered from the inequality.

- This implies that regretloss depends on the quality of the best expert.
- Thus having imperfect experts can take you from constant loss to linear loss if the loss of the best expert is linear.
- If you know T , can you choose a good β as a function of T to make loss sub linear in T . ?