Sequential Learning Algorithms

Samplers

D Manjunath & Jayakrishnan Nair

EE, IIT Bombay

January 8, 2022

Sampler 1: Predicting a Bernoulli sequence

- $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \ldots$.
- Let $\{X_t\}$ be i.i.d. Bernouilli(p) sequence.
- Let $\hat{X}_t$ be the prediction from the algorithm.
- Consider a randomized prediction, i.e., $\hat{X}_t$ is a Bernoulli(α) random variable.
- **Claim 1:** If p were known, the 'best' algorithm, the one that minimizes the expected number of mistakes upto to time T, would be, for $t = 1, \ldots, T$,

$$\hat{X}_t = I_{p \geq 0.5}$$

  i.e., a fixed prediction of 1 if $p \geq 0.5$ and 0 otherwise.
- Knowing the sequence $X_1, \ldots, X_{t-1}$ has no bearing on prediction for X_t.
- What if p is not known? And we want to predict for

Sampler 1: Predicting a Bernoulli sequence

- $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \ldots$.
- Let $\{X_t\}$ be i.i.d. Bernouilli(p) sequence.
- Let $\hat{X}_t$ be the prediction from the algorithm.
- Consider a randomized prediction, i.e., $\hat{X}_t$ is a Bernoulli(α) random variable.
- **Claim 1:** If p were known, the 'best' algorithm, the one that minimizes the expected number of mistakes upto to time T, would be, for $t = 1, \ldots, T$,

$$\hat{X}_t = \mathcal{I}_{p>0.5}$$

  i.e., a fixed prediction of 1 if $p \geq 0.5$ and 0 otherwise.
- Knowing the sequence $X_1, \ldots, X_{t-1}$ has no bearing on prediction for X_t
- What if p is not known? And we want to predict for

Sampler 1: Predicting a Bernoulli sequence

- $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \ldots$.
- Let $\{X_t\}$ be i.i.d. Bernouilli(p) sequence.
- Let $\hat{X}_t$ be the prediction from the algorithm.
- Consider a randomized prediction, i.e., $\hat{X}_t$ is a Bernoulli(α) random variable.
- **Claim 1:** If p were known, the 'best' algorithm, the one that minimizes the expected number of mistakes upto to time T, would be, for $t = 1, \ldots, T$,

$$\hat{X}_t = \mathcal{I}_{p \geq 0.5}$$

i.e., a fixed prediction of 1 if $p \geq 0.5$ and 0 otherwise.
- Knowing the sequence $X_1, \ldots, X_{t-1}$ has no bearing on prediction for X_t.
- What if p is not known? And we want to predict for

Sampler 1: Predicting a Bernoulli sequence

- $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \ldots$.
- Let $\{X_t\}$ be i.i.d. Bernouilli(p) sequence.
- Let $\hat{X}_t$ be the prediction from the algorithm.
- Consider a randomized prediction, i.e., $\hat{X}_t$ is a Bernoulli(α) random variable.
- **Claim 1:** If p were known, the 'best' algorithm, the one that minimizes the expected number of mistakes upto to time T, would be, for $t = 1, \ldots, T$,

$$\hat{X}_t = \mathbb{I}_{p > 0.5}$$

  i.e., a fixed prediction of 1 if $p > 0.5$ and 0 otherwise.
- Knowing the sequence $X_1, \ldots, X_{t-1}$ has no bearing on prediction for X_t.
- What if p is not known? And we want to predict for

Sampler 1: Predicting a Bernoulli sequence

- $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \ldots$.
- Let $\{X_t\}$ be i.i.d. Bernouilli(p) sequence.
- Let $\hat{X}_t$ be the prediction from the algorithm.
- Consider a randomized prediction, i.e., $\hat{X}_t$ is a Bernoulli(α) random variable.
- **Claim 1:** If p were known, the 'best' algorithm, the one that minimizes the expected number of mistakes upto to time T, would be, for $t = 1, \ldots, T$,

$$\hat{X}_t = \mathcal{I}_{p \geq 0.5}$$

  i.e., a fixed prediction of 1 if $p \geq 0.5$ and 0 otherwise.
- Knowing the sequence $X_1, \ldots, X_{t-1}$ has no bearing on prediction for X_t.
- What if p is not known? And we want to predict for

Sampler 1: Predicting a Bernoulli sequence

- $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \ldots$.
- Let $\{X_t\}$ be i.i.d. Bernouilli(p) sequence.
- Let $\hat{X}_t$ be the prediction from the algorithm.
- Consider a randomized prediction, i.e., $\hat{X}_t$ is a Bernoulli(α) random variable.
- **Claim 1:** If p were known, the 'best' algorithm, the one that minimizes the expected number of mistakes upto to time T, would be, for $t = 1, \ldots, T$,

$$\hat{X}_t = \mathcal{I}_{p \geq 0.5}$$

  i.e., a fixed prediction of 1 if $p \geq 0.5$ and 0 otherwise.
- Knowing the sequence $X_1, \ldots X_{t-1}$ has no bearing on prediction for X_t.
- What if p is not known? And we want to predict for

Sampler 1: Predicting a Bernoulli sequence

- $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \ldots$.
- Let $\{X_t\}$ be i.i.d. Bernouilli(p) sequence.
- Let $\hat{X}_t$ be the prediction from the algorithm.
- Consider a randomized prediction, i.e., $\hat{X}_t$ is a Bernoulli(α) random variable.
- **Claim 1:** If p were known, the 'best' algorithm, the one that minimizes the expected number of mistakes upto to time T, would be, for $t = 1, \ldots, T$,

$$\hat{X}_t = \mathcal{I}_{p \geq 0.5}$$

i.e., a fixed prediction of 1 if $p \geq 0.5$ and 0 otherwise.

- Knowing the sequence $X_1, \ldots X_{t-1}$ has no bearing on prediction for X_t.
- What if p is not known? And we want to predict for

Sampler 1: Predicting a Bernoulli sequence

- $X_t \in \{0,1\}$ is a binary sequence for $t = 1, 2, \ldots$.
- Let $\{X_t\}$ be i.i.d. Bernouilli(p) sequence.
- Let $\hat{X}_t$ be the prediction from the algorithm.
- Consider a randomized prediction, i.e., $\hat{X}_t$ is a Bernoulli(α) random variable.
- **Claim 1:** If p were known, the 'best' algorithm, the one that minimizes the expected number of mistakes upto to time T, would be, for $t = 1, \ldots, T$,

  $$\hat{X}_t = \mathcal{I}_{p \geq 0.5}$$

  i.e., a fixed prediction of 1 if $p \geq 0.5$ and 0 otherwise.
- Knowing the sequence $X_1, \ldots X_{t-1}$ has no bearing on prediction for X_t.
- What if p is not known? And we want to predict for

Sampler 1: Predicting a Bernoulli sequence

- $X_t \in \{0, 1\}$ is a binary sequence for $t = 1, 2, \ldots$.
- Let $\{X_t\}$ be i.i.d. Bernouilli(p) sequence.
- Let $\hat{X}_t$ be the prediction from the algorithm.
- Consider a randomized prediction, i.e., $\hat{X}_t$ is a Bernoulli(α) random variable.
- **Claim 1:** If p were known, the 'best' algorithm, the one that minimizes the expected number of mistakes upto to time T, would be, for $t = 1, \ldots, T$,

$$\hat{X}_t = \mathcal{I}_{p \geq 0.5}$$

  i.e., a fixed prediction of 1 if $p \geq 0.5$ and 0 otherwise.
- Knowing the sequence $X_1, \ldots X_{t-1}$ has no bearing on prediction for X_t.
- What if p is not known? And we want to predict for

Predicting a Bernoulli sequence

- **Claim 2:** More generally, let $\hat{X}_t$ be the prediction using the data $X_1, \ldots X_{t-1}$, from any algorithm. Then,

$$\mathsf{Pr}\left(\hat{X}_t \neq X_t \mid X_1, \ldots, X_{t-1}\right) \geq \min\{(p, (1-p)\}$$

i.e., a fixed prediction conditioned on the data has the minimum expected error.
Proof

Predicting a Berouilli sequence

- The last claim leads us to suggest the following *historical majority* algorithm (HMA) for prediction of a Benroulli sequence when p is unknown.
$$\hat{X}_t = \text{majority from } \{X_1, X_2, \ldots, X_{t-1}\}$$
- The expected loss for this algorithm upto time T is
$$\sum_{t=1}^{T} \Pr\left(\hat{X}_t \neq X_t\right) - \min\{p, (1-p)\}$$
- The best error that any algorithm could have had is with the knowledge of p. It is the difference that is the loss.
- The number of mistakes that the HM algorithm would make up to time T is the number of times that X_t was not equal to the majority of $\{X_1, X_2, \ldots, X_{t-1}\}$ for $t = 1, \ldots, T$.

Predicting a Berouilli sequence

- The last claim leads us to suggest the following *historical majority* algorithm (HMA) for prediction of a Benroulli sequence when p is unknown.

$$\hat{X}_t = \text{majority from } \{X_1, X_2, \ldots, X_{t-1}\}$$

- The expected loss for this algorithm upto time T is

$$\sum_{t=1}^{T} \mathsf{Pr}\left(\hat{X}_t \neq X_t\right) - \min\{p, (1-p)\}$$

- The best error that any algorithm could have had is with the knowledge of p. It is the difference that is the loss.
- The number of mistakes that the HM algorithm would make up to time T is the number of times that X_t was not equal to the majority of $\{X_1, X_2, \ldots, X_{t-1}\}$ for $t = 1, \ldots, T$.

Predicting a Berouilli sequence

- The last claim leads us to suggest the following *historical majority* algorithm (HMA) for prediction of a Benroulli sequence when p is unknown.
$$\hat{X}_t = \text{majority from } \{X_1, X_2, \ldots, X_{t-1}\}$$
- The expected loss for this algorithm upto time T is
$$\sum_{t=1}^{T} \mathsf{Pr}\left(\hat{X}_t \neq X_t\right) - \min\{p, (1-p)\}$$
- The best error that any algorithm could have had is with the knowledge of p. It is the difference that is the loss.
- The number of mistakes that the HM algorithm would make up to time T is the number of times that X_t was not equal to the majority of $\{X_1, X_2, \ldots, X_{t-1}\}$ for $t = 1, \ldots, T$.

Predicting a Berouilli sequence

- The last claim leads us to suggest the following *historical majority* algorithm (HMA) for prediction of a Benroulli sequence when p is unknown.
$$\hat{X}_t = \text{majority from } \{X_1, X_2, \ldots, X_{t-1}\}$$
- The expected loss for this algorithm upto time T is
$$\sum_{t=1}^{T} \mathsf{Pr}\left(\hat{X}_t \neq X_t\right) - \min\{p, (1-p)\}$$
- The best error that any algorithm could have had is with the knowledge of p. It is the difference that is the loss.
- The number of mistakes that the HM algorithm would make up to time T is the number of times that X_t was not equal to the majority of $\{X_1, X_2, \ldots, X_{t-1}\}$ for $t = 1, \ldots, T$.

Loss for the HMA

Worksheet

Loss for the HMA

- Without loss of generality, assume $p \geq 0.5$.
- Let L_t be the probability that the prediction for step t is different from the “full knowledge case”, i.e.,

$$L_t := \mathsf{Pr}\left(\hat{X}_t \neq 1\right)$$

- Clearly, L_t is the loss at step t.

$$\begin{aligned}
L_t &= \mathsf{Pr}\left(\sum_{k=1}^{t-1} X_k \leq \frac{t-1}{2}\right) \\
&= \mathsf{Pr}\left(\sum_{k=1}^{t-1} (X_k) - (t-1)p \leq \frac{t-1}{2} - (t-1)p\right) \\
&\leq \exp \frac{-2\left(\frac{t-1}{2} - (t-1)p\right)^2}{t-1} \qquad \text{Applying Hoeffding Inequality} \\
&= \exp -2(t-1)(0.5-p)^2
\end{aligned}$$

Loss for the HMA

- Without loss of generality, assume $p \geq 0.5$.
- Let L_t be the probability that the prediction for step t is different from the "full knowledge case", i.e.,

$$L_t := \Pr\left(\hat{X}_t \neq 1\right)$$

- Clearly, L_t is the loss at step t.

$$L_t = \Pr\left(\sum_{k=1}^{t-1} X_k \leq \frac{t-1}{2}\right)$$

Loss for the HMA

- Without loss of generality, assume $p \geq 0.5$.
- Let L_t be the probability that the prediction for step t is different from the "full knowledge case", i.e.,

$$L_t \;:=\; \Pr\left(\hat{X}_t \neq 1\right)$$

- Clearly, L_t is the loss at step t.

$$\begin{aligned}
L_t &= \Pr\left(\sum_{k=1}^{t-1} X_k \leq \frac{t-1}{2}\right) \\
&= \Pr\left(\sum_{k=1}^{t-1} (X_k) - (t-1)p \leq \frac{t-1}{2} - (t-1)p\right) \\
&\leq \exp \frac{-2\left(\frac{t-1}{2} - (t-1)p\right)^2}{t-1} \qquad \text{Applying Hoeffding Inequality} \\
&= \exp -2(t-1)(0.5-p)^2
\end{aligned}$$

Loss for the HMA

- Without loss of generality, assume $p \geq 0.5$.
- Let L_t be the probability that the prediction for step t is different from the "full knowledge case", i.e.,

$$L_t \;:=\; \mathsf{Pr}\left(\hat{X}_t \neq 1\right)$$

- Clearly, L_t is the loss at step t.

$$\begin{aligned}
L_t &= \mathsf{Pr}\left(\sum_{k=1}^{t-1} X_k \leq \frac{t-1}{2}\right) \\
&= \mathsf{Pr}\left(\sum_{k=1}^{t-1} (X_k) - (t-1)p \leq \frac{t-1}{2} - (t-1)p\right) \\
&\leq \exp \frac{-2\left(\frac{t-1}{2} - (t-1)p\right)^2}{t-1} \quad \text{Applying Hoeffding Inequality} \\
&= \exp -2(t-1)(0.5-p)^2
\end{aligned}$$

Loss for the HMA

- Without loss of generality, assume $p \geq 0.5$.
- Let L_t be the probability that the prediction for step t is different from the "full knowledge case", i.e.,

$$L_t \;:=\; \mathsf{Pr}\left(\hat{X}_t \neq 1\right)$$

- Clearly, L_t is the loss at step t.

$$\begin{aligned}
L_t &= \mathsf{Pr}\left(\sum_{k=1}^{t-1} X_k \leq \frac{t-1}{2}\right) \\
&= \mathsf{Pr}\left(\sum_{k=1}^{t-1} (X_k) - (t-1)p \leq \frac{t-1}{2} - (t-1)p\right) \\
&\leq \exp \frac{-2\left(\frac{t-1}{2} - (t-1)p\right)^2}{t-1} \qquad \text{Applying Hoeffding Inequality} \\
&= \exp -2(t-1)(0.5-p)^2
\end{aligned}$$

Loss for the HMA

- Without loss of generality, assume $p \geq 0.5$.
- Let L_t be the probability that the prediction for step t is different from the "full knowledge case", i.e.,

$$L_t := \mathsf{Pr}\left(\hat{X}_t \neq 1\right)$$

- Clearly, L_t is the loss at step t.

$$\begin{aligned}
L_t &= \mathsf{Pr}\left(\sum_{k=1}^{t-1} X_k \leq \frac{t-1}{2}\right) \\
&= \mathsf{Pr}\left(\sum_{k=1}^{t-1} (X_k) - (t-1)p \leq \frac{t-1}{2} - (t-1)p\right) \\
&\leq \exp \frac{-2\left(\frac{t-1}{2} - (t-1)p\right)^2}{t-1} \quad \text{Applying Hoeffding Inequality} \\
&= \exp -2(t-1)(0.5-p)^2
\end{aligned}$$

Loss for the HMA

- Without loss of generality, assume $p \geq 0.5$.
- Let L_t be the probability that the prediction for step t is different from the "full knowledge case", i.e.,

$$L_t \;:=\; \mathsf{Pr}\big(\hat{X}_t \neq 1\big)$$

- Clearly, L_t is the loss at step t.

$$\begin{aligned}
L_t &= \mathsf{Pr}\left(\sum_{k=1}^{t-1} X_k \leq \frac{t-1}{2}\right) \\
&= \mathsf{Pr}\left(\sum_{k=1}^{t-1} (X_k) - (t-1)p \leq \frac{t-1}{2} - (t-1)p\right) \\
&\leq \exp \frac{-2\left(\frac{t-1}{2} - (t-1)p\right)^2}{t-1} \quad \text{Applying Hoeffding Inequality} \\
&= \exp -2(t-1)(0.5-p)^2
\end{aligned}$$

Loss for the HMA

- Without loss of generality, assume $p \geq 0.5$.
- Let L_t be the probability that the prediction for step t is different from the "full knowledge case", i.e.,

$$L_t \;:=\; \mathsf{Pr}\big(\hat{X}_t \neq 1\big)$$

- Clearly, L_t is the loss at step t.

$$\begin{aligned}
L_t &= \mathsf{Pr}\left(\sum_{k=1}^{t-1} X_k \leq \frac{t-1}{2}\right) \\
&= \mathsf{Pr}\left(\sum_{k=1}^{t-1} (X_k) - (t-1)p \leq \frac{t-1}{2} - (t-1)p\right) \\
&\leq \exp \frac{-2\left(\frac{t-1}{2} - (t-1)p\right)^2}{t-1} \quad \text{Applying Hoeffding Inequality} \\
&= \exp -2(t-1)(0.5-p)^2
\end{aligned}$$

Loss for the HMA

- The cumulative regret in T steps is

$$\sum_{t=1}^{T} L_t \leq \sum_{t=1}^{T-1} \exp -2(t-1)(0.5-p)^2$$
$$\leq \frac{1}{1-\exp -2(0.5-p)^2}$$

- We thus see that the regret is upper bounded by a constant

Loss for the HMA

- The cumulative regret in T steps is

$$\sum_{t=1}^{T} L_t \leq \sum_{t=1}^{T-1} \exp -2(t-1)(0.5-p)^2$$
$$\leq \frac{1}{1-\exp -2(0.5-p)^2}$$

- We thus see that the regret is upper bounded by a constant.

Loss for the HMA

- The cumulative regret in T steps is

$$
\begin{aligned}
\sum_{t=1}^{T} L_t &\leq \sum_{t=1}^{T-1} \exp -2(t-1)(0.5-p)^2 \\
&\leq \frac{1}{1-\exp -2(0.5-p)^2}
\end{aligned}
$$

- We thus see that the regret is upper bounded by a constant.

Loss for the HMA

- The cumulative regret in T steps is

$$\begin{aligned}\sum_{t=1}^{T} L_t &\leq \sum_{t=1}^{T-1} \exp -2(t-1)(0.5-p)^2 \\ &\leq \frac{1}{1-\exp -2(0.5-p)^2}\end{aligned}$$

- We thus see that the regret is upper bounded by a constant.

Sampler 2: Predicting a "fixed point"

- Consider the following shooting game: You are shooting into a unit square and you have to "predict" the fixed point, say $\hat{X}_t$ for the t-th attempt.
- Let X_t be the actual location of the hit.
- For every $\hat{X}_t$ there is a cost $\|\hat{X}_t - X_t\|^2$.
- First consider the offline problem, or the one shot problem: the sequence X_t for $t = 1, \ldots, T$, is available to you and you have to choose the *best* fixed point

$$X_T^* := \arg \min_{x \in [0,1]^2} \sum_{t=1}^{T} \|X_t - x\|^2$$

- Claim: The explicit formula for X_T^* would be

$$X_T^* = \frac{1}{T} \sum_{t=1}^{T} X_t$$

- $\hat{X}_t = X_T^*$ minimizes the total cost.

Sampler 2: Predicting a "fixed point"

- Consider the following shooting game: You are shooting into a unit square and you have to "predict" the fixed point, say $\hat{X}_t$ for the t-th attempt.
- Let X_t be the actual location of the hit.
- For every $\hat{X}_t$ there is a cost $\|\hat{X}_t - X_t\|^2$.
- First consider the offline problem, or the one shot problem: the sequence X_t for $t = 1, \ldots, T$, is available to you and you have to choose the *best* fixed point

$$X_T^* := \arg \min_{x \in [0,1]^2} \sum_{t=1}^{T} \|X_t - x\|^2$$

Sampler 2: Predicting a “fixed point”

- Consider the following shooting game: You are shooting into a unit square and you have to “predict” the fixed point, say $\hat{X}_t$ for the t-th attempt.
- Let X_t be the actual location of the hit.
- For every $\hat{X}_t$ there is a cost $\|\hat{X}_t - X_t\|^2$.
- First consider the offline problem, or the one shot problem: the sequence X_t for $t = 1, \ldots, T$, is available to you and you have to choose the *best* fixed point

$$X_T^* := \arg \min_{x \in [0,1]^2} \sum_{t=1}^{T} \|X_t - x\|^2$$

Sampler 2: Predicting a “fixed point”

- Consider the following shooting game: You are shooting into a unit square and you have to “predict” the fixed point, say $\hat{X}_t$ for the t-th attempt.
- Let X_t be the actual location of the hit.
- For every $\hat{X}_t$ there is a cost $\|\hat{X}_t - X_t\|^2$.
- First consider the offline problem, or the one shot problem: the sequence X_t for $t = 1, \ldots, T$, is available to you and you have to choose the *best* fixed point

$$X_T^* := \arg \min_{x \in [0,1]^2} \sum_{t=1}^{T} \|X_t - x\|^2$$

- **Claim:** The explicit formula for X_T^* would be

$$X_T^* = \frac{1}{T} \sum_{t=1}^{T} X_t$$

- $\hat{X}_t = X_T^*$ minimizes the total cost

Sampler 2: Predicting a "fixed point"

- Consider the following shooting game: You are shooting into a unit square and you have to "predict" the fixed point, say $\hat{X}_t$ for the t-th attempt.
- Let X_t be the actual location of the hit.
- For every $\hat{X}_t$ there is a cost $\|\hat{X}_t - X_t\|^2$.
- First consider the offline problem, or the one shot problem: the sequence X_t for $t = 1, \ldots, T$, is available to you and you have to choose the *best* fixed point

$$X_T^* := \arg \min_{x \in [0,1]^2} \sum_{t=1}^{T} \|X_t - x\|^2$$

- **Claim:** The explicit formula for X_T^* would be

$$X_T^* = \frac{1}{T} \sum_{t=1}^{T} X_t$$

- $\hat{X}_t = X_T^*$ minimizes the total cost.

Sampler 2: Predicting a "fixed point"

- Consider the following shooting game: You are shooting into a unit square and you have to "predict" the fixed point, say $\hat{X}_t$ for the t-th attempt.
- Let X_t be the actual location of the hit.
- For every $\hat{X}_t$ there is a cost $\|\hat{X}_t - X_t\|^2$.
- First consider the offline problem, or the one shot problem: the sequence X_t for $t = 1, \ldots, T$, is available to you and you have to choose the *best* fixed point

$$X_T^* := \arg \min_{x \in [0,1]^2} \sum_{t=1}^{T} \|X_t - x\|^2$$

- **Claim:** The explicit formula for X_T^* would be

$$X_T^* = \frac{1}{T} \sum_{t=1}^{T} X_t$$

- $\hat{X}_t = X_T^*$ minimizes the total cost.

Sampler 2: Predicting a "fixed point"

- Consider the following shooting game: You are shooting into a unit square and you have to "predict" the fixed point, say $\hat{X}_t$ for the t-th attempt.
- Let X_t be the actual location of the hit.
- For every $\hat{X}_t$ there is a cost $\|\hat{X}_t - X_t\|^2$.
- First consider the offline problem, or the one shot problem: the sequence X_t for $t = 1, \ldots, T$, is available to you and you have to choose the *best* fixed point

$$X_T^* := \arg \min_{x \in [0,1]^2} \sum_{t=1}^{T} \|X_t - x\|^2$$

- **Claim:** The explicit formula for X_T^* would be

$$X_T^* = \frac{1}{T} \sum_{t=1}^{T} X_t$$

- $\hat{X}_t = X_T^*$ minimizes the total cost.

Sampler 2: Predicting a “fixed point”

- Now convert this to an online question: Predict the point after every shot such that the cumulative error is minimized.
- Specifically,

$$\min \sum_{t=1} \|\hat{X}_t - X_t\|^2$$

with $\hat{X}_t$ to be determined before the t-th shot and can be done using the information $X_1, \ldots, X_{t-1}$.
- Consider the following algorithm

$$\hat{X}_t = X^*_{t-1} = \arg\min_{x \in [0,1]^2} \sum_{s=1}^{t-1} \|X_s - x\|^2 = \frac{1}{t-1} \sum_{s=1}^{t-1} X_s$$

- How does this compare with the “offline” algorithm above. Specifically, what is the cumulative *loss* in the online algorithm compared to the full information offline case, i.e.,

Sampler 2: Predicting a “fixed point”

- Now convert this to an online question: Predict the point after every shot such that the cumulative error is minimized.
- Specifically,

$$\min \sum_{t=1} \|\hat{X}_t - X_t\|^2$$

with $\hat{X}_t$ to be determined before the t-th shot and can be done using the information $X_1, \ldots, X_{t-1}$.

- Consider the following algorithm

$$\hat{X}_t = X^*_{t-1}$$

Sampler 2: Predicting a "fixed point"

- Now convert this to an online question: Predict the point after every shot such that the cumulative error is minimized.
- Specifically,

$$\min \sum_{t=1} \|\hat{X}_t - X_t\|^2$$

with $\hat{X}_t$ to be determined before the t-th shot and can be done using the information $X_1, \ldots, X_{t-1}$.

- Consider the following algorithm

$$\hat{X}_t = X^*_{t-1} = \arg \min_{X \in [0,1]^2} \sum_{s=1}^{t-1} \|X_s - x\|^2 = \frac{1}{t-1} \sum_{s=1}^{t-1} X_s$$

Sampler 2: Predicting a "fixed point"

- Now convert this to an online question: Predict the point after every shot such that the cumulative error is minimized.
- Specifically,

$$\min \sum_{t=1} \|\hat{X}_t - X_t\|^2$$

with $\hat{X}_t$ to be determined before the t-th shot and can be done using the information $X_1, \ldots, X_{t-1}$.
- Consider the following algorithm

$$\hat{X}_t = X^*_{t-1} = \arg \min_{X \in [0,1]^2} \sum_{s=1}^{t-1} \|X_s - x\|^2 = \frac{1}{t-1} \sum_{s=1}^{t-1} X_s$$

- How does this compare with the "offline" algorithm above. Specifically, what is the cumulative *loss* in the online algorithm compared to the full information offline case, i.e.,

Sampler 2: Predicting a “fixed point”

- Now convert this to an online question: Predict the point after every shot such that the cumulative error is minimized.
- Specifically,

$$\min \sum_{t=1} \|\hat{X}_t - X_t\|^2$$

  with $\hat{X}_t$ to be determined before the t-th shot and can be done using the information $X_1, \ldots, X_{t-1}$.
- Consider the following algorithm

$$\hat{X}_t = X^*_{t-1} = \arg \min_{X \in [0,1]^2} \sum_{s=1}^{t-1} \|X_s - x\|^2 = \frac{1}{t-1} \sum_{s=1}^{t-1} X_s$$

- How does this compare with the “offline” algorithm above. Specifically, what is the cumulative *loss* in the online algorithm compared to the full information offline case, i.e.,

Sampler 2: Predicting a “fixed point”

- Now convert this to an online question: Predict the point after every shot such that the cumulative error is minimized.
- Specifically,

$$\min \sum_{t=1} \|\hat{X}_t - X_t\|^2$$

with $\hat{X}_t$ to be determined before the t-th shot and can be done using the information $X_1, \ldots, X_{t-1}$.

- Consider the following algorithm

$$\hat{X}_t = X^*_{t-1} = \arg \min_{X \in [0,1]^2} \sum_{s=1}^{t-1} \|X_s - x\|^2 = \frac{1}{t-1} \sum_{s=1}^{t-1} X_s$$

- How does this compare with the “offline” algorithm above. Specifically, what is the cumulative *loss* in the online algorithm compared to the full information offline case, i.e.,

$$\sum^{T} \|\hat{X}_t - X_t\|^2 - \sum^{T} \|\hat{X}_T - X_t\|^2 = \sum^{T} \|X^*_{t-1} - X_t\|^2 - \sum^{T} \|X^*_T - X_t\|^2$$

Sampler 2: Predicting a "fixed point"

- Now convert this to an online question: Predict the point after every shot such that the cumulative error is minimized.
- Specifically,

$$\min \sum_{t=1} \|\hat{X}_t - X_t\|^2$$

with $\hat{X}_t$ to be determined before the t-th shot and can be done using the information $X_1, \ldots, X_{t-1}$.

- Consider the following algorithm

$$\hat{X}_t = X^*_{t-1} = \arg \min_{X \in [0,1]^2} \sum_{s=1}^{t-1} \|X_s - x\|^2 = \frac{1}{t-1} \sum_{s=1}^{t-1} X_s$$

- How does this compare with the "offline" algorithm above. Specifically, what is the cumulative *loss* in the online algorithm compared to the full information offline case, i.e.,

$$\sum^{T} \|\hat{X}_t - X_t\|^2 - \sum^{T} \|\hat{X}_T - X_t\|^2 = \sum^{T} \|X^*_{t-1} - X_t\|^2 - \sum^{T} \|X^*_T - X_t\|^2$$

Sampler 2: Predicting a “fixed point”

- Now convert this to an online question: Predict the point after every shot such that the cumulative error is minimized.
- Specifically,

$$\min \sum_{t=1} \|\hat{X}_t - X_t\|^2$$

with $\hat{X}_t$ to be determined before the t-th shot and can be done using the information $X_1, \ldots, X_{t-1}$.

- Consider the following algorithm

$$\hat{X}_t = X^*_{t-1} = \arg \min_{X \in [0,1]^2} \sum_{s=1}^{t-1} \|X_s - x\|^2 = \frac{1}{t-1} \sum_{s=1}^{t-1} X_s$$

- How does this compare with the “offline” algorithm above. Specifically, what is the cumulative *loss* in the online algorithm compared to the full information offline case, i.e.,

$$\sum^{T} \|\hat{X}_t - X_t\|^2 - \sum^{T} \|\hat{X}_T - X_t\|^2 = \sum^{T} \|X^*_{t-1} - X_t\|^2 - \sum^{T} \|X^*_T - X_t\|^2$$

Analyzing Follow the Leader (FTL)

Lemma 1

$$\sum_{t=1}^{T} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T} \|X_T^* - X_t\|^2$$

Analyzing Follow the Leader (FTL)

Lemma 1

$$\sum_{t=1}^{T} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T} \|X_T^* - X_t\|^2$$

Proof

- Proof is by induction;
- Trivially true for $T = 1$ with equality.
- Induction hypothesis: For $T > 1$
- We will now prove

Analyzing Follow the Leader (FTL)

Lemma 1

$$\sum_{t=1}^{T} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T} \|X_T^* - X_t\|^2$$

Proof

- Proof is by induction;
- Trivially true for $T = 1$ with equality.
- Induction hypothesis: For $T > 1$,

$$\sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T-1} \|X_{T-1}^* - X_t\|^2$$

- We will now prove

$$\sum_{t=1}^{T} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T} \|X_T^* - X_t\|^2$$

Analyzing Follow the Leader (FTL)

Lemma 1

$$\sum_{t=1}^{T} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T} \|X_T^* - X_t\|^2$$

Proof

- Proof is by induction;
- Trivially true for $T = 1$ with equality.
- Induction hypothesis: For $T > 1$,

$$\sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T-1} \|X_{T-1}^* - X_t\|^2$$

- We will now prove

$$\sum_{t=1}^{T} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T} \|X_T^* - X_t\|^2$$

Analyzing Follow the Leader (FTL)

Lemma 1

$$\sum_{t=1}^{T} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T} \|X_T^* - X_t\|^2$$

Proof

- Proof is by induction;
- Trivially true for $T = 1$ with equality.
- Induction hypothesis: For $T > 1$,

$$\sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T-1} \|X_{T-1}^* - X_t\|^2$$

- We will now prove

$$\sum_{t=1}^{T} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T} \|X_T^* - X_t\|^2$$

Analyzing Follow the Leader (FTL)

Lemma 1

$$\sum_{t=1}^{T} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T} \|X_T^* - X_t\|^2$$

Proof

- Proof is by induction;
- Trivially true for $T = 1$ with equality.
- Induction hypothesis: For $T > 1$,

$$\sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T-1} \|X_{T-1}^* - X_t\|^2$$

- We will now prove

$$\sum_{t=1}^{T} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T} \|X_T^* - X_t\|^2$$

Analyzing Follow the Leader (FTL)

Lemma 1

$$\sum_{t=1}^{T} \|X_t^* - X_t\|^2 \quad \leq \quad \sum_{t=1}^{T} \|X_T^* - X_t\|^2$$

$$\Rightarrow \quad \sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 \quad \leq \quad \sum_{t=1}^{T-1} \|X_T^* - X_t\|^2 \text{ knocking of term for } t = T$$

We now prove the preceding inequality.

$$\begin{aligned} \sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 \quad &\leq \quad \sum_{t=1}^{T-1} \|X_{T-1}^* - X_t\|^2 \quad && \text{from induction hypothesis} \\ &\leq \quad \sum_{t=1}^{T-1} \|X_T^* - X_t\|^2 \quad && \text{because } X_T^* \text{ is not the minimizer} \quad \square \end{aligned}$$

Analyzing Follow the Leader (FTL)

Lemma 1

$$
\begin{aligned}
\sum_{t=1}^{T} \|X_t^* - X_t\|^2 &\leq \sum_{t=1}^{T} \|X_T^* - X_t\|^2 \\
\Rightarrow \sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 &\leq \sum_{t=1}^{T-1} \|X_T^* - X_t\|^2 \text{ knocking of term for } t = T
\end{aligned}
$$

We now prove the preceding inequality.

$$
\begin{aligned}
\sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 &\leq \sum_{t=1}^{T-1} \|X_{T-1}^* - X_t\|^2 \quad \text{from induction hypothesis} \\
&\leq \sum_{t=1}^{T-1} \|X_T^* - X_t\|^2 \quad \text{because } X_T^* \text{ is not the minimizer} \quad \square
\end{aligned}
$$

Analyzing Follow the Leader (FTL)

Lemma 1

$$\sum_{t=1}^{T} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T} \|X_T^* - X_t\|^2$$

$$\Rightarrow \sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T-1} \|X_T^* - X_t\|^2 \text{ knocking of term for } t = T$$

We now prove the preceding inequality.

$$\sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T-1} \|X_{T-1}^* - X_t\|^2 \quad \text{from induction hypothesis}$$

$$\leq \sum_{t=1}^{T-1} \|X_T^* - X_t\|^2 \quad \text{because } X_T^* \text{ is not the minimizer} \quad \square$$

Analyzing Follow the Leader (FTL)

Lemma 1

$$
\begin{aligned}
\sum_{t=1}^{T} \|X_t^* - X_t\|^2 &\leq \sum_{t=1}^{T} \|X_T^* - X_t\|^2 \\
\Rightarrow \sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 &\leq \sum_{t=1}^{T-1} \|X_T^* - X_t\|^2 \text{ knocking of term for } t = T
\end{aligned}
$$

We now prove the preceding inequality.

$$
\begin{aligned}
\sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 &\leq \sum_{t=1}^{T-1} \|X_{T-1}^* - X_t\|^2 \quad \text{from induction hypothesis} \\
&\leq \sum_{t=1}^{T-1} \|X_T^* - X_t\|^2 \quad \text{because } X_T^* \text{ is not the minimizer} \quad \square
\end{aligned}
$$

Analyzing Follow the Leader (FTL)

Lemma 1

$$\sum_{t=1}^{T} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T} \|X_T^* - X_t\|^2$$

$$\Rightarrow \sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T-1} \|X_T^* - X_t\|^2 \text{ knocking of term for } t = T$$

We now prove the preceding inequality.

$$\sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 \leq \sum_{t=1}^{T-1} \|X_{T-1}^* - X_t\|^2 \quad \text{from induction hypothesis}$$

$$\leq \sum_{t=1}^{T-1} \|X_T^* - X_t\|^2 \quad \text{because } X_T^* \text{ is not the minimizer} \quad \square$$

Analyzing Follow the Leader (FTL)

Lemma 1

$$
\begin{aligned}
\sum_{t=1}^{T} \|X_t^* - X_t\|^2 &\leq \sum_{t=1}^{T} \|X_T^* - X_t\|^2 \\
\Rightarrow \sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 &\leq \sum_{t=1}^{T-1} \|X_T^* - X_t\|^2 \text{ knocking of term for } t = T
\end{aligned}
$$

We now prove the preceding inequality.

$$
\begin{aligned}
\sum_{t=1}^{T-1} \|X_t^* - X_t\|^2 &\leq \sum_{t=1}^{T-1} \|X_{T-1}^* - X_t\|^2 \quad \text{from induction hypothesis} \\
&\leq \sum_{t=1}^{T-1} \|X_T^* - X_t\|^2 \quad \text{because } X_T^* \text{ is not the minimizer} \quad \square
\end{aligned}
$$

Analyzing Follow the Leader (FTL)

$$
\begin{aligned}
L_T \;&:=\; \sum_{t=1}^{T} \|\hat{X}_t - X_t\|^2 \;-\; \sum_{t=1}^{T} \|\hat{X}_T^* - X_t\|^2 \\
&=\; \sum_{t=1}^{T} \|X_{t-1}^* - X_t\|^2 \;-\; \sum_{t=1}^{T} \|\hat{X}_T^* - X_t\|^2 \\
&\leq\; \sum_{t=1}^{T} \|X_{t-1}^* - X_t\|^2 \;-\; \sum_{t=1}^{T} \|\hat{X}_t^* - X_t\|^2 \quad \text{by Lemma 1 above} \\
&=\; \sum_{t=1}^{T} \langle X_{t-1}^* - X_t^*,\; X_{t-1}^* + X_t^* - 2X_t \rangle \quad \text{using } \|x\|^2 - \|y\|^2 = \langle x+y, x-y \rangle
\end{aligned}
$$

Analyzing Follow the Leader (FTL)

$$
\begin{aligned}
L_T &:= \sum_{t=1}^{T} \|\hat{X}_t - X_t\|^2 - \sum_{t=1}^{T} \|\hat{X}_T^* - X_t\|^2 \\
&= \sum_{t=1}^{T} \|X_{t-1}^* - X_t\|^2 - \sum_{t=1}^{T} \|\hat{X}_T^* - X_t\|^2 \\
&\leq \sum_{t=1}^{T} \|X_{t-1}^* - X_t\|^2 - \sum_{t=1}^{T} \|\hat{X}_t^* - X_t\|^2 \quad \text{by Lemma 1 above} \\
&= \sum_{t=1}^{T} \langle X_{t-1}^* - X_t^*,\, X_{t-1}^* + X_t^* - 2X_t \rangle \quad \text{using } \|x\|^2 - \|y\|^2 = \langle x+y, x-y \rangle
\end{aligned}
$$

Analyzing Follow the Leader (FTL)

$$
\begin{aligned}
L_T &:= \sum_{t=1}^{T} \|\hat{X}_t - X_t\|^2 - \sum_{t=1}^{T} \|\hat{X}_T^* - X_t\|^2 \\
&= \sum_{t=1}^{T} \|X_{t-1}^* - X_t\|^2 - \sum_{t=1}^{T} \|\hat{X}_T^* - X_t\|^2 \\
&\leq \sum_{t=1}^{T} \|X_{t-1}^* - X_t\|^2 - \sum_{t=1}^{T} \|\hat{X}_t^* - X_t\|^2 \quad \text{by Lemma 1 above} \\
&= \sum_{t=1}^{T} \langle X_{t-1}^* - X_t^*,\, X_{t-1}^* + X_t^* - 2X_t \rangle \quad \text{using } \|x\|^2 - \|y\|^2 = \langle x+y, x-y \rangle
\end{aligned}
$$

Analyzing Follow the Leader (FTL)

$$
\begin{aligned}
L_T &:= \sum_{t=1}^{T} \|\hat{X}_t - X_t\|^2 - \sum_{t=1}^{T} \|\hat{X}_T^* - X_t\|^2 \\
&= \sum_{t=1}^{T} \|X_{t-1}^* - X_t\|^2 - \sum_{t=1}^{T} \|\hat{X}_T^* - X_t\|^2 \\
&\leq \sum_{t=1}^{T} \|X_{t-1}^* - X_t\|^2 - \sum_{t=1}^{T} \|\hat{X}_t^* - X_t\|^2 \quad \text{by Lemma 1 above} \\
&= \sum_{t=1}^{T} \langle X_{t-1}^* - X_t^*,\ X_{t-1}^* + X_t^* - 2X_t \rangle \quad \text{using } \|x\|^2 - \|y\|^2 = \langle x+y, x-y \rangle
\end{aligned}
$$

Analyzing Follow the Leader (FTL)

$$
\begin{aligned}
L_T &\leq \sum_{t=1}^{T} \|X_{t-1}^* - X_t^*\| \cdot \ \|X_{t-1}^* + X_t^* - 2X_t\| \\
&\quad \text{using Cauchy-Schwartz inequality } |\langle x, y\rangle| \leq \|x\| \cdot \|y\|. \\
&\leq \sum_{t=1}^{T} \|X_{t-1}^* - X_t^*\| \cdot \left(\|X_{t-1}^*\| + \|X_t^*\| + \|2X_t\|\right) \quad \text{using triangle inequality} \\
&\leq \sum_{t=1}^{T} 4 \, \|X_{t-1}^* - X_t^*\| \qquad \text{using the upper bound for } \|\cdot\| \text{ in 2nd term}
\end{aligned}
$$

Use the formula for x_t^*, triangle inequality and upper bound for $\|\cdot\|$

$$
\begin{aligned}
\|X_{t-1}^* - X_t^*\| &= \|X_{t-1}^* - \frac{(t-1)X_{t-1}^* + X_t}{t}\| = \frac{1}{t}\|X_{t-1}^* - X_t\| \\
&= \frac{1}{t}\left(\|X_{t-1}^*\| + \|X_t\|\right) \leq \frac{2}{t}
\end{aligned}
$$

Analyzing Follow the Leader (FTL)

$$
\begin{aligned}
L_T &\leq \sum_{t=1}^{T} \|X_{t-1}^* - X_t^*\| \cdot \ \|X_{t-1}^* + X_t^* - 2X_t\| \\
&\quad \text{using Cauchy-Schwartz inequality } |\langle x, y\rangle| \leq \|x\| \cdot \|y\|. \\
&\leq \sum_{t=1}^{T} \|X_{t-1}^* - X_t^*\| \ \cdot \left(\|X_{t-1}^*\| + \|X_t^*\| + \|2X_t\|\right) \quad \text{using triangle inequality} \\
&\leq \sum_{t=1}^{T} 4 \, \|X_{t-1}^* - X_t^*\| \qquad \text{using the upper bound for } \|\cdot\| \text{ in 2nd term}
\end{aligned}
$$

Use the formula for x_t^*, triangle inequality and upper bound for $\|\cdot\|$

$$
\begin{aligned}
\|X_{t-1}^* - X_t^*\| &= \|X_{t-1}^* - \frac{(t-1)X_{t-1}^* + X_t}{t}\| = \frac{1}{t}\|X_{t-1}^* - X_t\| \\
&= \frac{1}{t}\left(\|X_{t-1}^*\| + \|X_t\|\right) \leq \frac{2}{t}
\end{aligned}
$$

Analyzing Follow the Leader (FTL)

$$
\begin{aligned}
L_T &\leq \sum_{t=1}^{T} \|X^*_{t-1} - X^*_t\| \cdot \ \|X^*_{t-1} + X^*_t - 2X_t\| \\
&\quad \text{using Cauchy-Schwartz inequality } |\langle x, y\rangle| \leq \|x\| \cdot \|y\|. \\
&\leq \sum_{t=1}^{T} \|X^*_{t-1} - X^*_t\| \ \cdot \left(\|X^*_{t-1}\| + \|X^*_t\| + \|2X_t\|\right) \ \text{ using triangle inequality} \\
&\leq \sum_{t=1}^{T} 4\, \|X^*_{t-1} - X^*_t\| \quad \text{ using the upper bound for } \|\cdot\| \text{ in 2nd term}
\end{aligned}
$$

Use the formula for x^*_t, triangle inequality and upper bound for $\|\cdot\|$

$$
\begin{aligned}
\|X^*_{t-1} - X^*_t\| &= \|X^*_{t-1} - \frac{(t-1)X^*_{t-1} + X_t}{t}\| = \frac{1}{t}\|X^*_{t-1} - X_t\| \\
&= \frac{1}{t}\left(\|X^*_{t-1}\| + \|X_t\|\right) \leq \frac{2}{t}
\end{aligned}
$$

Analyzing Follow the Leader (FTL)

$$
\begin{aligned}
L_T &\leq \sum_{t=1}^{T} \|X^*_{t-1} - X^*_t\| \cdot \ \|X^*_{t-1} + X^*_t - 2X_t\| \\
&\quad \text{using Cauchy-Schwartz inequality } |\langle x, y\rangle| \leq \|x\| \cdot \|y\|. \\
&\leq \sum_{t=1}^{T} \|X^*_{t-1} - X^*_t\| \ \cdot \left(\|X^*_{t-1}\| + \|X^*_t\| + \|2X_t\|\right) \ \ \text{using triangle inequality} \\
&\leq \sum_{t=1}^{T} 4 \, \|X^*_{t-1} - X^*_t\| \quad \text{using the upper bound for } \|\cdot\| \text{ in 2nd term}
\end{aligned}
$$

Use the formula for x^*_t, triangle inequality and upper bound for $\|\cdot\|$

$$
\begin{aligned}
\|X^*_{t-1} - X^*_t\| &= \|X^*_{t-1} - \frac{(t-1)X^*_{t-1} + X_t}{t}\| = \frac{1}{t}\|X^*_{t-1} - X_t\| \\
&= \frac{1}{t}\left(\|X^*_{t-1}\| + \|X_t\|\right) \ \leq \ \frac{2}{t}
\end{aligned}
$$

Analyzing Follow the Leader (FTL)

$$R_T \leq 4 \sum_{t=1}^{T} \frac{8}{T} \leq 8(1 + \ln(T))$$

Theorem For any sequence $X_1, \ldots X_T$, in the unit circle the FTL algorithm has a loss of at most $(1 + 8\ln(T))$.

Analyzing Follow the Leader (FTL)

$$R_T \leq 4 \sum_{t=1}^{T} \frac{8}{T} \leq 8(1 + \ln(T))$$

Theorem For any sequence $X_1, \ldots X_T$, in the unit circle the FTL algorithm has a loss of at most $(1 + 8\ln(T))$.