Efficient Optimal Strategies for Universal Prediction

Peter Bartlett

Computer Science and Statistics
University of California at Berkeley

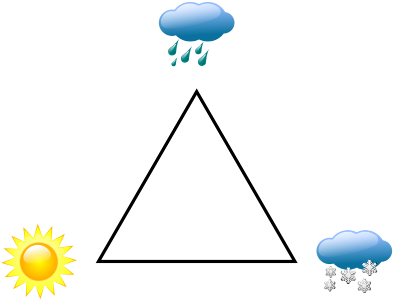
Mathematical Sciences
Queensland University of Technology

Joint work with Yasin Abbasi-Yadkori, Wouter Koolen, Alan Malek, Eiji Takimoto, Manfred Warmuth.

Online Prediction

A repeated game:

At round t:

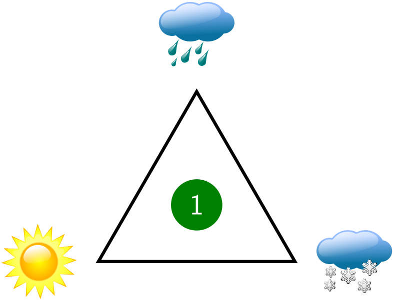


Online Prediction

A repeated game:

At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.

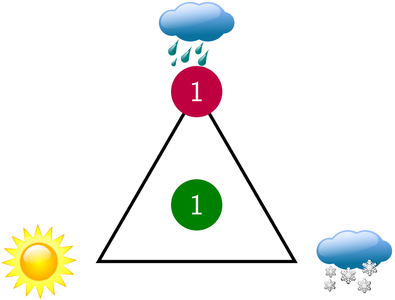


Online Prediction

A repeated game:

At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.
2. Adversary chooses outcome $y_t \in \mathcal{Y}$.

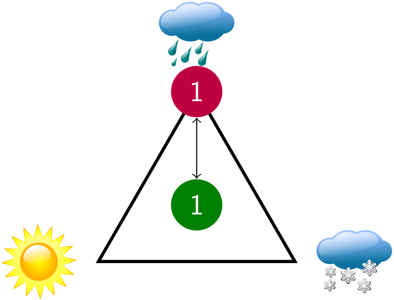


Online Prediction

A repeated game:

At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.
2. Adversary chooses outcome $y_t \in \mathcal{Y}$.
3. Player incurs loss $\ell(a_t, y_t)$.



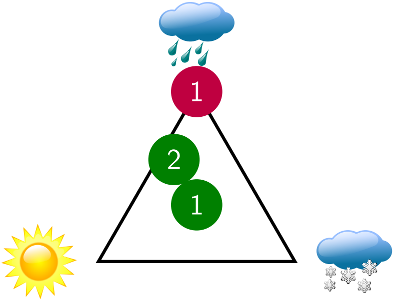
$$\ell(a_t, y_t) = \|a_t - y_t\|^2.$$

Online Prediction

A repeated game:

At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.
2. Adversary chooses outcome $y_t \in \mathcal{Y}$.
3. Player incurs loss $\ell(a_t, y_t)$.

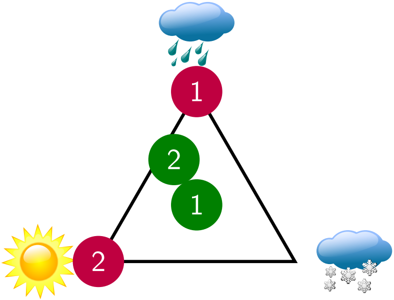


Online Prediction

A repeated game:

At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.
2. Adversary chooses outcome $y_t \in \mathcal{Y}$.
3. Player incurs loss $\ell(a_t, y_t)$.

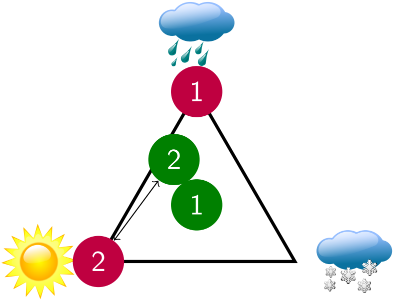


Online Prediction

A repeated game:

At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.
2. Adversary chooses outcome $y_t \in \mathcal{Y}$.
3. Player incurs loss $\ell(a_t, y_t)$.

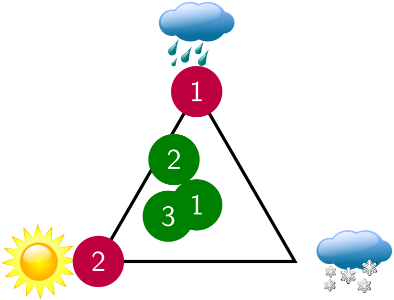


Online Prediction

A repeated game:

At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.
2. Adversary chooses outcome $y_t \in \mathcal{Y}$.
3. Player incurs loss $\ell(a_t, y_t)$.

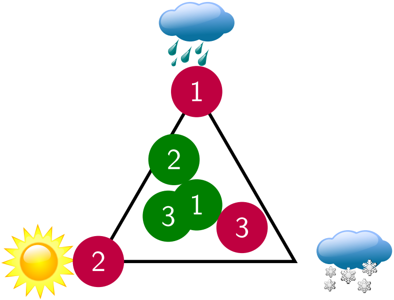


Online Prediction

A repeated game:

At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.
2. Adversary chooses outcome $y_t \in \mathcal{Y}$.
3. Player incurs loss $\ell(a_t, y_t)$.

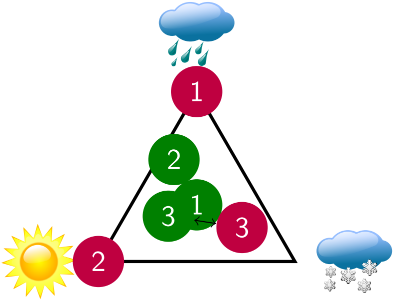


Online Prediction

A repeated game:

At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.
2. Adversary chooses outcome $y_t \in \mathcal{Y}$.
3. Player incurs loss $\ell(a_t, y_t)$.



Online Prediction

A repeated game:

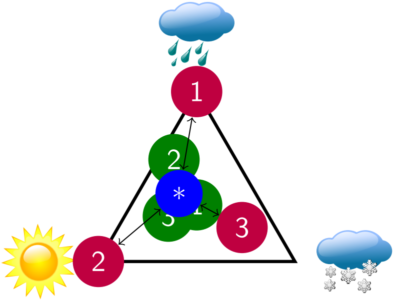
At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.
2. Adversary chooses outcome $y_t \in \mathcal{Y}$.
3. Player incurs loss $\ell(a_t, y_t)$.

Player's aim:

Minimize *regret:*

$$\sum_{t=1}^{T} \ell(a_t, y_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t).$$



Online Prediction

A repeated game:

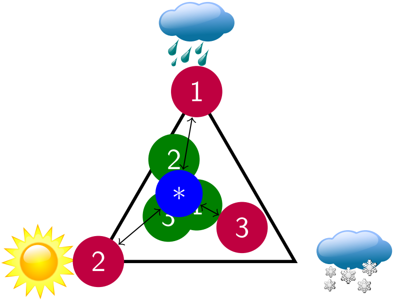
At round t:

1. Player chooses prediction $a_t \in \mathcal{A}$.
2. Adversary chooses outcome $y_t \in \mathcal{Y}$.
3. Player incurs loss $\ell(a_t, y_t)$.

Player's aim:

Minimize *regret* wrt comparison $\mathcal{C}$:

$$\sum_{t=1}^{T} \ell(a_t, y_t) - \inf_{\hat{a} \in \mathcal{C}} \sum_{t=1}^{T} \ell(\hat{a}_t, y_t).$$



Online Prediction Games: Why

- Universal prediction:
  very weak assumptions on process generating the data.

Online Prediction Games: Why

- Universal prediction:
  very weak assumptions on process generating the data.
- Deterministic heart of a decision problem.

Online Prediction Games: Why

- Universal prediction:
  very weak assumptions on process generating the data.
- Deterministic heart of a decision problem.
- Gives robust statistical methods.

Online Prediction Games: Why

- Universal prediction:
  very weak assumptions on process generating the data.
- Deterministic heart of a decision problem.
- Gives robust statistical methods.
- Typically streaming, so very scalable.

Online Prediction Games: Why

- Universal prediction:
  very weak assumptions on process generating the data.
- Deterministic heart of a decision problem.
- Gives robust statistical methods.
- Typically streaming, so very scalable.
- This talk: Minimax optimal strategies.

Online Prediction Games

Regret

$$\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t)$$

Online Prediction Games

Minimax Regret

$$\left(\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right)$$

Online Prediction Games

Minimax Regret

$$\min_{a_1 \in \mathcal{A}} \qquad \left(\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right)$$

Online Prediction Games

Minimax Regret

$$\min_{a_1 \in \mathcal{A}} \max_{y_1 \in \mathcal{Y}} \left(\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right)$$

Online Prediction Games

Minimax Regret

$$\min_{a_1 \in \mathcal{A}} \max_{y_1 \in \mathcal{Y}} \cdots \min_{a_T \in \mathcal{A}} \left(\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right)$$

Online Prediction Games

Minimax Regret

$$\min_{a_1 \in \mathcal{A}} \max_{y_1 \in \mathcal{Y}} \cdots \min_{a_T \in \mathcal{A}} \max_{y_T \in \mathcal{Y}} \left(\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right)$$

Online Prediction Games

The value of the game: Minimax Regret

$$V_T(\mathcal{Y}, \mathcal{A}) = \min_{a_1 \in \mathcal{A}} \max_{y_1 \in \mathcal{Y}} \cdots \min_{a_T \in \mathcal{A}} \max_{y_T \in \mathcal{Y}} \left(\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right)$$

Online Prediction Games

The value of the game: Minimax Regret

$$V_T(\mathcal{Y},\mathcal{A}) = \min_{a_1\in\mathcal{A}} \max_{y_1\in\mathcal{Y}} \cdots \min_{a_T\in\mathcal{A}} \max_{y_T\in\mathcal{Y}} \left(\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a\in\mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right)$$

Strategy:

$$S : \bigcup_{t=0}^{T} \mathcal{Y}^t \to \mathcal{A}.$$

Online Prediction Games

The value of the game: Minimax Regret

$$V_T(\mathcal{Y},\mathcal{A}) = \min_{a_1\in\mathcal{A}} \max_{y_1\in\mathcal{Y}} \cdots \min_{a_T\in\mathcal{A}} \max_{y_T\in\mathcal{Y}} \left(\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a\in\mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right)$$

Strategy:

$$S : \bigcup_{t=0}^{T} \mathcal{Y}^t \to \mathcal{A}.$$

$$V_T(\mathcal{Y},\mathcal{A}) = \min_{S} \max_{y_1^T\in\mathcal{Y}^T} \left(\sum_{t=1}^{T} \ell\left(S\left(y_1^{t-1}\right), y_t\right) - \min_{a\in\mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right)$$

Online Prediction Games

The value of the game: Minimax Regret

$$V_T(\mathcal{Y}, \mathcal{A}) = \min_{a_1 \in \mathcal{A}} \max_{y_1 \in \mathcal{Y}} \cdots \min_{a_T \in \mathcal{A}} \max_{y_T \in \mathcal{Y}} \left(\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right)$$

Minimax Optimal Strategy:

$$S^* : \bigcup_{t=0}^{T} \mathcal{Y}^t \to \mathcal{A}.$$

$$\begin{aligned} V_T(\mathcal{Y}, \mathcal{A}) &= \min_{S} \max_{y_1^T \in \mathcal{Y}^T} \left(\sum_{t=1}^{T} \ell\left(S\left(y_1^{t-1}\right), y_t\right) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right) \\ &= \max_{y_1^T \in \mathcal{Y}^T} \left(\sum_{t=1}^{T} \ell\left(S^*\left(y_1^{t-1}\right), y_t\right) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right). \end{aligned}$$

Online Prediction Games

Questions

Online Prediction Games

Questions

- Minimax regret?

Online Prediction Games

Questions

- Minimax regret?
- Optimal player's strategy?

Online Prediction Games

Questions

- Minimax regret?
- Optimal player's strategy?
- Efficiently computable?

Online Prediction Games

Questions

- Minimax regret?
- Optimal player's strategy?
- Efficiently computable?
- Optimal adversary's strategy?

Online Prediction Games

Questions

- Minimax regret?
- Optimal player's strategy?
- Efficiently computable?
- Optimal adversary's strategy?
- How do they depend on ℓ?

Online Prediction Games

Questions

- Minimax regret?
- Optimal player's strategy?
- Efficiently computable?
- Optimal adversary's strategy?
- How do they depend on ℓ?

loss, $\ell(a, y)$:

1. $\|a - y\|_2^2$,

   $a, y \in \mathbb{R}^d$.

Online Prediction Games

Questions

- Minimax regret?
- Optimal player's strategy?
- Efficiently computable?
- Optimal adversary's strategy?
- How do they depend on ℓ?

loss, $\ell(a, y)$:

1. $\|a - y\|_2^2$,
   $a, y \in \mathbb{R}^d$.
2. $(x^\top a - y)^2$.

Online Prediction Games

Questions

- Minimax regret?
- Optimal player's strategy?
- Efficiently computable?
- Optimal adversary's strategy?
- How do they depend on ℓ?

loss, $\ell(a, y)$:

1. $\|a - y\|_2^2$,
   $a, y \in \mathbb{R}^d$.
2. $(x^\top a - y)^2$.
3. $-\log a(y)$,
   $a \in \{p_\theta : \theta \in \Theta\}$.

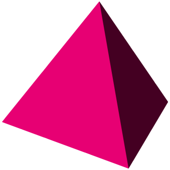
Online Prediction Games

Questions

- Minimax regret?
- Optimal player's strategy?
- Efficiently computable?
- Optimal adversary's strategy?
- How do they depend on ℓ, $\mathcal{Y}$, $\mathcal{A}$?

loss, $\ell(a, y)$:

1. $\|a - y\|_2^2$,
   $a, y \in \mathbb{R}^d$.
2. $(x^\top a - y)^2$.
3. $-\log a(y)$,
   $a \in \{p_\theta : \theta \in \Theta\}$.



wikimedia.org

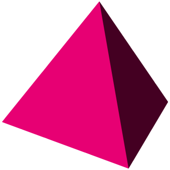
Online Prediction Games

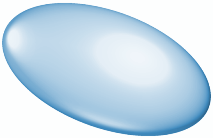
Questions

- Minimax regret?
- Optimal player's strategy?
- Efficiently computable?
- Optimal adversary's strategy?
- How do they depend on ℓ, $\mathcal{Y}$, $\mathcal{A}$?

loss, $\ell(a, y)$:

1. $\|a - y\|_2^2$,
   $a, y \in \mathbb{R}^d$.
2. $(x^\top a - y)^2$.
3. $-\log a(y)$,
   $a \in \{p_\theta : \theta \in \Theta\}$.





wikimedia.org
mfo.de

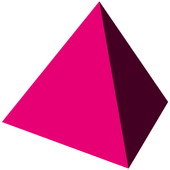
Online Prediction Games

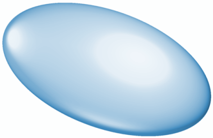
Questions

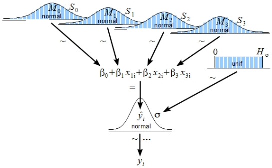
- Minimax regret?
- Optimal player's strategy?
- Efficiently computable?
- Optimal adversary's strategy?
- How do they depend on ℓ, $\mathcal{Y}$, $\mathcal{A}$?

loss, $\ell(a, y)$:

1. $\|a - y\|_2^2$, $\quad a, y \in \mathbb{R}^d$.
2. $(x^\top a - y)^2$.
3. $-\log a(y)$, $\quad a \in \{p_\theta : \theta \in \Theta\}$.







wikimedia.org
mfo.de
www.indiana.edu

Outline

Outline

- Computing minimax optimal strategies.

Outline

- Computing minimax optimal strategies.
- Part 1: Euclidean loss.

Outline

- Computing minimax optimal strategies.
- Part 1: Euclidean loss.
- Part 2: Linear regression.

Outline

- Computing minimax optimal strategies.
- Part 1: Euclidean loss.
- Part 2: Linear regression.
- Time series forecasting.

Outline

- **Computing minimax optimal strategies.**
- Part 1: Euclidean loss.
- Part 2: Linear regression.
- Time series forecasting.

Computing minimax optimal strategies

Computing minimax optimal strategies

The value of the game:

$$V_T(\mathcal{Y},\mathcal{A}) = \min_{a_1\in\mathcal{A}} \max_{y_1\in\mathcal{Y}} \cdots \min_{a_T\in\mathcal{A}} \max_{y_T\in\mathcal{Y}} \left(\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a\in\mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right).$$

Recursion for the value-to-go, given a history:

Computing minimax optimal strategies

The value of the game:

$$V_T(\mathcal{Y}, \mathcal{A}) = \min_{a_1 \in \mathcal{A}} \max_{y_1 \in \mathcal{Y}} \cdots \min_{a_T \in \mathcal{A}} \max_{y_T \in \mathcal{Y}} \left(\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right).$$

Recursion for the value-to-go, given a history:

$$V(y_1, \ldots, y_T) := -\min_{a} \sum_{t=1}^{T} \ell(a, y_t),$$

Computing minimax optimal strategies

The value of the game:

$$V_T(\mathcal{Y},\mathcal{A}) = \min_{a_1\in\mathcal{A}} \max_{y_1\in\mathcal{Y}} \cdots \min_{a_T\in\mathcal{A}} \max_{y_T\in\mathcal{Y}} \left(\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a\in\mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right).$$

Recursion for the value-to-go, given a history:

$$V(y_1,\ldots,y_T) := -\min_a \sum_{t=1}^{T} \ell(a, y_t),$$

$$V(y_1,\ldots,y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1,\ldots,y_t)\right),$$

Computing minimax optimal strategies

The value of the game:

$$V_T(\mathcal{Y}, \mathcal{A}) = \min_{a_1 \in \mathcal{A}} \max_{y_1 \in \mathcal{Y}} \cdots \min_{a_T \in \mathcal{A}} \max_{y_T \in \mathcal{Y}} \left(\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right).$$

Recursion for the value-to-go, given a history:

$$V(y_1, \ldots, y_T) := -\min_{a} \sum_{t=1}^{T} \ell(a, y_t),$$

$$V(y_1, \ldots, y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right),$$

$$V_T(\mathcal{Y}, \mathcal{A}) = V(),$$

Computing minimax optimal strategies

The value of the game:

$$V_T(\mathcal{Y}, \mathcal{A}) = \min_{a_1 \in \mathcal{A}} \max_{y_1 \in \mathcal{Y}} \cdots \min_{a_T \in \mathcal{A}} \max_{y_T \in \mathcal{Y}} \left(\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell(a, y_t) \right).$$

Recursion for the value-to-go, given a history:

$$\begin{aligned}
V(y_1, \ldots, y_T) &:= -\min_a \sum_{t=1}^{T} \ell(a, y_t), \\
V(y_1, \ldots, y_{t-1}) &:= \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right), \\
V_T(\mathcal{Y}, \mathcal{A}) &= V(), \\
S^*(y_1, \ldots, y_{t-1}) &= \arg\min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right).
\end{aligned}$$

Computing minimax optimal strategies

To play the minimax strategy: after seeing $y_1, \ldots, y_{t-1}$,

Computing minimax optimal strategies

To play the minimax strategy: after seeing $y_1, \ldots, y_{t-1}$,

1. Compute V,

Computing minimax optimal strategies

To play the minimax strategy: after seeing $y_1, \ldots, y_{t-1}$,

1. Compute V,
2. Choose a_t as the minimizer of

$$\max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right)$$

Computing minimax optimal strategies

To play the minimax strategy: after seeing $y_1, \ldots, y_{t-1}$,

1. Compute V,
2. Choose a_t as the minimizer of

$$\max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right)$$

Difficult!

Computing minimax optimal strategies

To play the minimax strategy: after seeing $y_1, \ldots, y_{t-1}$,

1. Compute V,
2. Choose a_t as the minimizer of

$$\max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right)$$

Difficult!

Efficient minimax optimal strategies

When is V a simple function of (statistics of) the history $y_1, \ldots, y_t$?

Outline

- Computing minimax optimal strategies.
- **Part 1: Euclidean loss.**
- Part 2: Linear regression.
- Time series forecasting.

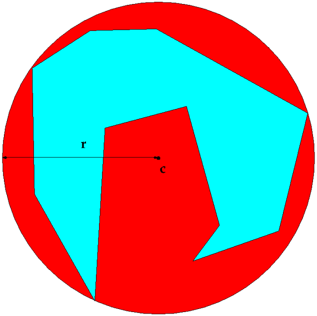
Online prediction with Euclidean loss (with Wouter Koolen, Alan Malek)

Online prediction with Euclidean loss (with Wouter Koolen, Alan Malek)

- Prediction in $\mathbb{R}^d$:
  $\mathcal{Y} \subseteq \mathbb{R}^d$, $\mathcal{A} = \mathbb{R}^d$, Euclidean loss: $\ell(\hat{y}, y) = \frac{1}{2} \|\hat{y} - y\|^2$.

Online prediction with Euclidean loss (with Wouter Koolen, Alan Malek)

- Prediction in $\mathbb{R}^d$:
  $\mathcal{Y} \subseteq \mathbb{R}^d$, $\mathcal{A} = \mathbb{R}^d$, Euclidean loss: $\ell(\hat{y}, y) = \frac{1}{2}\|\hat{y} - y\|^2$.
- Minimax strategy is empirical minimizer plus shrinkage towards center of smallest ball containing $\mathcal{Y}$: $a^*_{t+1} = t\alpha_{t+1}\bar{y}_t + (1 - t\alpha_{t+1})c$.



Online prediction with Euclidean loss (with Wouter Koolen, Alan Malek)

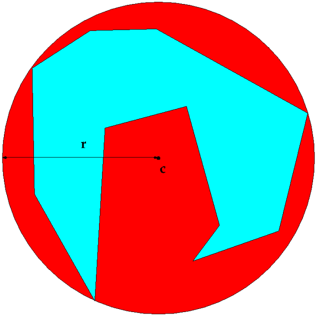
- Prediction in $\mathbb{R}^d$:
  $\mathcal{Y} \subseteq \mathbb{R}^d$, $\mathcal{A} = \mathbb{R}^d$, Euclidean loss: $\ell(\hat{y}, y) = \frac{1}{2}\|\hat{y} - y\|^2$.
- Minimax strategy is empirical minimizer plus shrinkage towards center of smallest ball containing $\mathcal{Y}$: $a^*_{t+1} = t\alpha_{t+1}\bar{y}_t + (1 - t\alpha_{t+1})c$.

- Regret:

$$\frac{r^2}{2}\sum_{t=1}^{T}\alpha_t,$$

where r is radius of smallest ball,

$$\alpha_T = \frac{1}{T}, \qquad \alpha_t = \alpha_{t+1}^2 + \alpha_{t+1}$$



Online prediction with quadratic loss

The simplex case

Suppose $\mathcal{Y}$ is a set of $d+1$ affinely independent points in $\mathbb{R}^d$, all lying on the surface of the smallest ball.

Online prediction with quadratic loss

The simplex case

Suppose $\mathcal{Y}$ is a set of $d+1$ affinely independent points in $\mathbb{R}^d$, all lying on the surface of the smallest ball.
Maintain statistics: $s_n = \sum_{t=1}^n (y_t - c)$, $\quad \sigma_n^2 = \sum_{t=1}^n \|y_t - c\|^2$.

Online prediction with quadratic loss

The simplex case

Suppose $\mathcal{Y}$ is a set of $d+1$ affinely independent points in $\mathbb{R}^d$, all lying on the surface of the smallest ball.
Maintain statistics: $s_n = \sum_{t=1}^{n}(y_t - c)$, $\qquad \sigma_n^2 = \sum_{t=1}^{n} \|y_t - c\|^2$.

Value-to-go: quadratic in state

$$\frac{1}{2}\left(\alpha_n \|s_n\|^2 - \sigma_n^2 + r^2 \sum_{t=n+1}^{T} \alpha_t\right).$$

Online prediction with quadratic loss

The simplex case

Suppose $\mathcal{Y}$ is a set of $d+1$ affinely independent points in $\mathbb{R}^d$, all lying on the surface of the smallest ball.
Maintain statistics: $s_n = \sum_{t=1}^n (y_t - c)$, $\sigma_n^2 = \sum_{t=1}^n \|y_t - c\|^2$.

Value-to-go: quadratic in state

$$\frac{1}{2}\left(\alpha_n \|s_n\|^2 - \sigma_n^2 + r^2 \sum_{t=n+1}^{T} \alpha_t\right).$$

$$\alpha_T = \frac{1}{T}, \qquad \alpha_t = \alpha_{t+1}^2 + \alpha_{t+1}$$

Online prediction with quadratic loss

The simplex case

Suppose $\mathcal{Y}$ is a set of $d+1$ affinely independent points in $\mathbb{R}^d$, all lying on the surface of the smallest ball.
Maintain statistics: $s_n = \sum_{t=1}^{n}(y_t - c)$, $\qquad \sigma_n^2 = \sum_{t=1}^{n} \|y_t - c\|^2$.

Value-to-go: quadratic in state

$$\frac{1}{2}\left(\alpha_n\|s_n\|^2 - \sigma_n^2 + r^2 \sum_{t=n+1}^{T} \alpha_t\right).$$

Minimax strategy: affine in state

$$a_{n+1}^* - c = n\alpha_{n+1}\frac{s_n}{n}.$$

$$\alpha_T = \frac{1}{T}, \qquad \alpha_t = \alpha_{t+1}^2 + \alpha_{t+1}$$

Online prediction with quadratic loss

The simplex case

Suppose $\mathcal{Y}$ is a set of $d+1$ affinely independent points in $\mathbb{R}^d$, all lying on the surface of the smallest ball.
Maintain statistics: $s_n = \sum_{t=1}^{n}(y_t - c)$, $\sigma_n^2 = \sum_{t=1}^{n} \|y_t - c\|^2$.

Value-to-go: quadratic in state

$$\frac{1}{2}\left(\alpha_n\|s_n\|^2 - \sigma_n^2 + r^2 \sum_{t=n+1}^{T} \alpha_t\right).$$

Minimax strategy: affine in state

$$a_{n+1}^* - c = n\alpha_{n+1}\frac{s_n}{n}.$$

$$a_{n+1}^* = n\alpha_{n+1}\bar{y}_n + (1 - n\alpha_{n+1})c$$

$$\alpha_T = \frac{1}{T}, \qquad \alpha_t = \alpha_{t+1}^2 + \alpha_{t+1}$$

Online prediction with quadratic loss

The simplex case

Suppose $\mathcal{Y}$ is a set of $d+1$ affinely independent points in $\mathbb{R}^d$, all lying on the surface of the smallest ball.
Maintain statistics: $s_n = \sum_{t=1}^n (y_t - c)$, $\sigma_n^2 = \sum_{t=1}^n \|y_t - c\|^2$.

Value-to-go: quadratic in state

$$\frac{1}{2}\left(\alpha_n \|s_n\|^2 - \sigma_n^2 + r^2 \sum_{t=n+1}^{T} \alpha_t\right).$$

Minimax strategy: affine in state

$$a_{n+1}^* - c = n\alpha_{n+1}\frac{s_n}{n}.$$

$$a_{n+1}^* = n\alpha_{n+1}\bar{y}_n + (1 - n\alpha_{n+1})c$$

$$\alpha_T = \frac{1}{T}, \qquad \alpha_t = \alpha_{t+1}^2 + \alpha_{t+1} \le \frac{1}{t}.$$

Online prediction with quadratic loss

The simplex case

Suppose $\mathcal{Y}$ is a set of $d+1$ affinely independent points in $\mathbb{R}^d$, all lying on the surface of the smallest ball.
Maintain statistics: $s_n = \sum_{t=1}^{n}(y_t - c)$, $\sigma_n^2 = \sum_{t=1}^{n} \|y_t - c\|^2$.

Value-to-go: quadratic in state

$$\frac{1}{2}\left(\alpha_n \|s_n\|^2 - \sigma_n^2 + r^2 \sum_{t=n+1}^{T} \alpha_t\right).$$

Minimax strategy: affine in state

$$a_{n+1}^* - c = n\alpha_{n+1}\frac{s_n}{n}.$$

$$a_{n+1}^* = n\alpha_{n+1}\bar{y}_n + (1 - n\alpha_{n+1})c$$

Maximin distribution: same mean.

$$\alpha_T = \frac{1}{T}, \qquad \alpha_t = \alpha_{t+1}^2 + \alpha_{t+1} \le \frac{1}{t}.$$

Online prediction with quadratic loss on the simplex

Proof idea

Online prediction with quadratic loss on the simplex

Proof idea

$$V(y_1, \ldots, y_T) := -\min_a \sum_{t=1}^{T} \ell(a, y_t),$$

$$V(y_1, \ldots, y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t)\right).$$

Online prediction with quadratic loss on the simplex

Proof idea

$$V(y_1, \ldots, y_T) := -\min_a \sum_{t=1}^{T} \ell(a, y_t),$$

$$V(y_1, \ldots, y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right).$$

The final $V(y_1, \ldots, y_T)$ is a (convex) quadratic in the state.

Online prediction with quadratic loss on the simplex

Proof idea

$$V(y_1, \ldots, y_T) := -\min_a \sum_{t=1}^{T} \ell(a, y_t),$$

$$V(y_1, \ldots, y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t)\right).$$

The final $V(y_1, \ldots, y_T)$ is a (convex) quadratic in the state.

$$V(y_1, \ldots, y_{t-1}) := \min_{a_t} \max_{p_t} \mathbb{E}_{y_t \sim p_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t)\right)$$

Online prediction with quadratic loss on the simplex

Proof idea

$$V(y_1, \ldots, y_T) := -\min_a \sum_{t=1}^{T} \ell(a, y_t),$$

$$V(y_1, \ldots, y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right).$$

The final $V(y_1, \ldots, y_T)$ is a (convex) quadratic in the state.

$$\begin{aligned} V(y_1, \ldots, y_{t-1}) &:= \min_{a_t} \max_{p_t} \mathbb{E}_{y_t \sim p_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right) \\ &= \max_{p_t} \min_{a_t} \mathbb{E}_{y_t \sim p_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right). \end{aligned}$$

Online prediction with quadratic loss on the simplex

Proof idea

$$V(y_1,\ldots,y_T) := -\min_a \sum_{t=1}^{T} \ell(a, y_t),$$

$$V(y_1,\ldots,y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1,\ldots,y_t)\right).$$

The final $V(y_1,\ldots,y_T)$ is a (convex) quadratic in the state.

$$\begin{aligned} V(y_1,\ldots,y_{t-1}) &:= \min_{a_t} \max_{p_t} \mathbb{E}_{y_t\sim p_t} \left(\ell(a_t, y_t) + V(y_1,\ldots,y_t)\right) \\ &= \max_{p_t} \min_{a_t} \mathbb{E}_{y_t\sim p_t} \left(\ell(a_t, y_t) + V(y_1,\ldots,y_t)\right). \end{aligned}$$

At each step, the unconstrained maximizer in $\{p \in \mathbb{R}^{d+1} : 1^\top p = 1\}$ keeps the value-to-go a quadratic function.

Online prediction with quadratic loss on the simplex

Proof idea

$$V(y_1, \ldots, y_T) := -\min_a \sum_{t=1}^{T} \ell(a, y_t),$$

$$V(y_1, \ldots, y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t)\right).$$

The final $V(y_1, \ldots, y_T)$ is a (convex) quadratic in the state.

$$\begin{aligned} V(y_1, \ldots, y_{t-1}) &:= \min_{a_t} \max_{p_t} \mathbb{E}_{y_t \sim p_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t)\right) \\ &= \max_{p_t} \min_{a_t} \mathbb{E}_{y_t \sim p_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t)\right). \end{aligned}$$

At each step, the unconstrained maximizer in $\{p \in \mathbb{R}^{d+1} : 1^\top p = 1\}$ keeps the value-to-go a quadratic function.
When the simplex points are on the surface of the smallest ball, the maximizer is a probability distribution.

Online prediction with quadratic loss on the ball

The ball case: $\mathcal{Y} = \{y : \|y - c\| \leq r\}$

Online prediction with quadratic loss on the ball

The ball case: $\mathcal{Y} = \{y : \|y - c\| \leq r\}$

Maintain statistics: $s_n = \sum_{t=1}^{n}(y_t - c)$, $\qquad \sigma_n^2 = \sum_{t=1}^{n} \|y_t - c\|^2$.

Online prediction with quadratic loss on the ball

The ball case: $\mathcal{Y} = \{y : \|y - c\| \leq r\}$

Maintain statistics: $s_n = \sum_{t=1}^n (y_t - c)$, $\sigma_n^2 = \sum_{t=1}^n \|y_t - c\|^2$.

Value-to-go: quadratic in state

$$\frac{1}{2}\left(\alpha_n \|s_n\|^2 - \sigma_n^2 + r^2 \sum_{t=n+1}^{T} \alpha_t\right).$$

Online prediction with quadratic loss on the ball

The ball case: $\mathcal{Y} = \{y : \|y - c\| \leq r\}$

Maintain statistics: $s_n = \sum_{t=1}^{n}(y_t - c)$, $\quad \sigma_n^2 = \sum_{t=1}^{n} \|y_t - c\|^2$.

Value-to-go: quadratic in state

$$\frac{1}{2}\left(\alpha_n\|s_n\|^2 - \sigma_n^2 + r^2 \sum_{t=n+1}^{T} \alpha_t\right).$$

Minimax strategy: affine in state

$$a_{n+1}^* - c = n\alpha_{n+1}\frac{s_n}{n}.$$

Online prediction with quadratic loss on the ball

The ball case: $\mathcal{Y} = \{y : \|y - c\| \le r\}$

Maintain statistics: $s_n = \sum_{t=1}^{n}(y_t - c)$, $\qquad \sigma_n^2 = \sum_{t=1}^{n} \|y_t - c\|^2$.

Value-to-go: quadratic in state

$$\frac{1}{2}\left(\alpha_n \|s_n\|^2 - \sigma_n^2 + r^2 \sum_{t=n+1}^{T} \alpha_t\right).$$

Minimax strategy: affine in state

$$a_{n+1}^* - c = n\alpha_{n+1}\frac{s_n}{n}.$$

$$a_{n+1}^* = n\alpha_{n+1}\bar{y}_n + (1 - n\alpha_{n+1})c$$

Online prediction with quadratic loss on the ball

The ball case: $\mathcal{Y} = \{y : \|y - c\| \le r\}$

Maintain statistics: $s_n = \sum_{t=1}^{n}(y_t - c)$, $\sigma_n^2 = \sum_{t=1}^{n} \|y_t - c\|^2$.

Value-to-go: quadratic in state

$$\frac{1}{2}\left(\alpha_n \|s_n\|^2 - \sigma_n^2 + r^2 \sum_{t=n+1}^{T} \alpha_t\right).$$

Minimax strategy: affine in state

$$a_{n+1}^* - c = n\alpha_{n+1}\frac{s_n}{n}.$$

$$a_{n+1}^* = n\alpha_{n+1}\bar{y}_n + (1 - n\alpha_{n+1})c$$

Maximin distribution: same mean.

Online prediction with quadratic loss on the ball

The ball case: $\mathcal{Y} = \{y : \|y - c\| \le r\}$

Maintain statistics: $s_n = \sum_{t=1}^{n}(y_t - c)$, $\sigma_n^2 = \sum_{t=1}^{n} \|y_t - c\|^2$.

Value-to-go: quadratic in state

$$\frac{1}{2}\left(\alpha_n \|s_n\|^2 - \sigma_n^2 + r^2 \sum_{t=n+1}^{T} \alpha_t\right).$$

Minimax strategy: affine in state

$$a_{n+1}^* - c = n\alpha_{n+1}\frac{s_n}{n}.$$

$$a_{n+1}^* = n\alpha_{n+1}\bar{y}_n + (1 - n\alpha_{n+1})c$$

Maximin distribution: same mean.

Minimax regret for ball

$$V(\mathcal{Y}) = \frac{r^2}{2}\sum_{t=1}^{T}\alpha_t.$$

Online prediction with quadratic loss on the ball

Proof idea

Online prediction with quadratic loss on the ball

Proof idea

$$V(y_1, \ldots, y_T) := -\min_a \sum_{t=1}^{T} \ell(a, y_t),$$

$$V(y_1, \ldots, y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right).$$

Online prediction with quadratic loss on the ball

Proof idea

$$V(y_1, \ldots, y_T) := -\min_a \sum_{t=1}^{T} \ell(a, y_t),$$

$$V(y_1, \ldots, y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right).$$

The final $V(y_1, \ldots, y_T)$ is a (convex) quadratic in the state.

Online prediction with quadratic loss on the ball

Proof idea

$$V(y_1, \ldots, y_T) := -\min_a \sum_{t=1}^{T} \ell(a, y_t),$$

$$V(y_1, \ldots, y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t)\right).$$

The final $V(y_1, \ldots, y_T)$ is a (convex) quadratic in the state.

$$V(y_1, \ldots, y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t)\right).$$

Online prediction with quadratic loss on the ball

Proof idea

$$V(y_1, \dots, y_T) := -\min_a \sum_{t=1}^{T} \ell(a, y_t),$$

$$V(y_1, \dots, y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \dots, y_t) \right).$$

The final $V(y_1, \dots, y_T)$ is a (convex) quadratic in the state.

$$V(y_1, \dots, y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \dots, y_t) \right).$$

At each step, the inner maximum is of a (convex) quadratic criterion with a single quadratic constraint. This is a rare example of a nonconvex problem where strong duality holds.

Online prediction with quadratic loss on the ball

Proof idea

$$V(y_1, \ldots, y_T) := -\min_a \sum_{t=1}^{T} \ell(a, y_t),$$

$$V(y_1, \ldots, y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right).$$

The final $V(y_1, \ldots, y_T)$ is a (convex) quadratic in the state.

$$V(y_1, \ldots, y_{t-1}) := \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right).$$

At each step, the inner maximum is of a (convex) quadratic criterion with a single quadratic constraint. This is a rare example of a nonconvex problem where strong duality holds. Evaluating the dual gives the recurrence for the value-to-go.

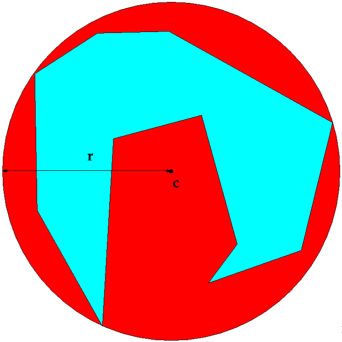
Online prediction with quadratic loss

The general case: closed, bounded $\mathcal{Y} \subset \mathbb{R}^d$

Online prediction with quadratic loss

The general case: closed, bounded $\mathcal{Y} \subset \mathbb{R}^d$

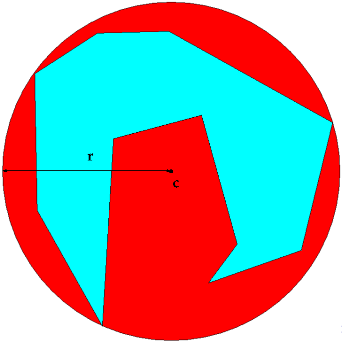
Recall: the smallest ball containing $\mathcal{Y}$ is $B_{\mathcal{Y}} = \{x \in \mathbb{R}^d : \|x - c\| \leq r\}$.



Online prediction with quadratic loss

The general case: closed, bounded $\mathcal{Y} \subset \mathbb{R}^d$

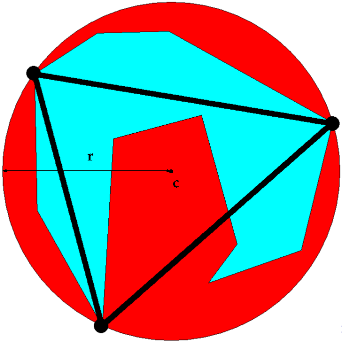
Recall: the smallest ball containing $\mathcal{Y}$ is $B_{\mathcal{Y}} = \{x \in \mathbb{R}^d : \|x - c\| \leq r\}$.
A Lagrange dual argument shows that the optimal center is in the convex hull of a set of *contact points* of $\mathcal{Y}$ at radius r.



Online prediction with quadratic loss

The general case: closed, bounded $\mathcal{Y} \subset \mathbb{R}^d$

Recall: the smallest ball containing $\mathcal{Y}$ is $B_{\mathcal{Y}} = \{x \in \mathbb{R}^d : \|x - c\| \leq r\}$.
A Lagrange dual argument shows that the optimal center is in the convex hull of a set of *contact points* of $\mathcal{Y}$ at radius r.
From Carathéodory's Theorem, there is an affinely independent subset S of these contact points, with $|S| \leq d + 1$.



Online prediction with quadratic loss

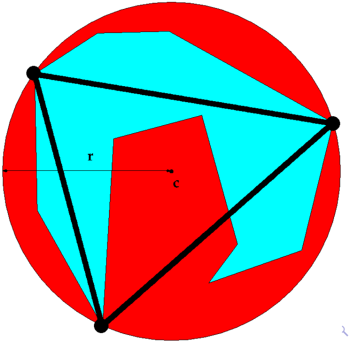
The general case: closed, bounded $\mathcal{Y} \subset \mathbb{R}^d$

Recall: the smallest ball containing $\mathcal{Y}$ is $B_{\mathcal{Y}} = \{x \in \mathbb{R}^d : \|x - c\| \leq r\}$.
A Lagrange dual argument shows that the optimal center is in the convex hull of a set of *contact points* of $\mathcal{Y}$ at radius r.
From Carathéodory's Theorem, there is an affinely independent subset S of these contact points, with $|S| \leq d + 1$.

From below

$\mathcal{Y} \supseteq S$, so

$$V(\mathcal{Y}) \geq V(S) = \frac{r^2}{2} \sum_{i=1}^{T} \alpha_i.$$



Online prediction with quadratic loss

The general case: closed, bounded $\mathcal{Y} \subset \mathbb{R}^d$

Recall: the smallest ball containing $\mathcal{Y}$ is $B_{\mathcal{Y}} = \{x \in \mathbb{R}^d : \|x - c\| \leq r\}$.
A Lagrange dual argument shows that the optimal center is in the convex hull of a set of *contact points* of $\mathcal{Y}$ at radius r.
From Carathéodory's Theorem, there is an affinely independent subset S of these contact points, with $|S| \leq d + 1$.

From below

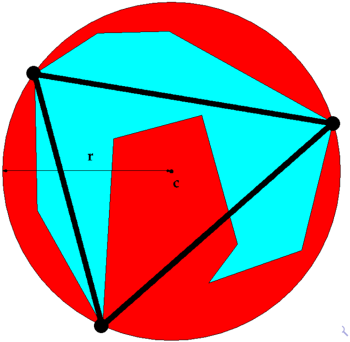
$\mathcal{Y} \supseteq S$, so

$$V(\mathcal{Y}) \geq V(S) = \frac{r^2}{2} \sum_{i=1}^{T} \alpha_i.$$

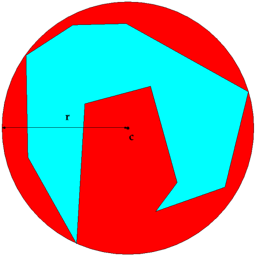
From above

$\mathcal{Y} \subseteq B_{\mathcal{Y}}$, so

$$V(\mathcal{Y}) \leq V(B_{\mathcal{Y}}) = \frac{r^2}{2} \sum_{i=1}^{T} \alpha_i.$$



Main result: the role of the smallest ball



The smallest ball: $B_{\mathcal{Y}}$

The smallest ball containing $\mathcal{Y}$ is
$B_{\mathcal{Y}} = \{y \in \mathbb{R}^d : \|y - c\| \le r\}$, with
$c = \arg\min_c \max_{y\in\mathcal{Y}} \|y - c\|$,
$r = \min_c \max_{y\in\mathcal{Y}} \|y - c\|$.

Main Theorem

For closed, bounded $\mathcal{Y} \subset \mathbb{R}^d$:

Minimax strategy is $a^*_{n+1} = n\alpha_{n+1} \frac{1}{n} \sum_{t=1}^{n} y_t + (1 - n\alpha_{n+1})c$.

Optimal regret is $V(\mathcal{Y}) = \frac{r^2}{2} \sum_{n=1}^{T} \alpha_n$.

Online prediction with quadratic loss

Minimax regret

$$V(\mathcal{Y}) = \frac{r^2}{2} \sum_{t=1}^{T} \alpha_t$$

Online prediction with quadratic loss

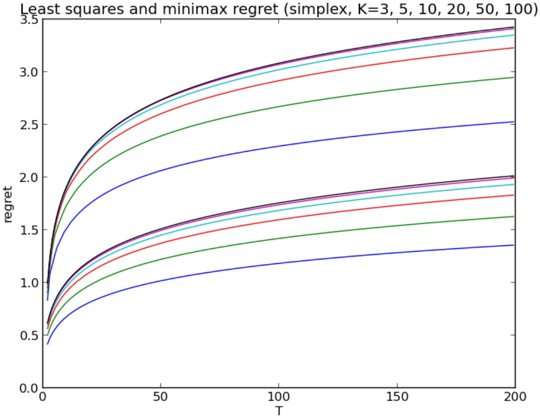
Minimax regret

$$V(\mathcal{Y}) = \frac{r^2}{2}\sum_{t=1}^{T}\alpha_t = \frac{r^2}{2}\left(\log T - \log\log T + O\left(\frac{\log\log T}{\log T}\right)\right).$$

Online prediction with quadratic loss

Minimax regret

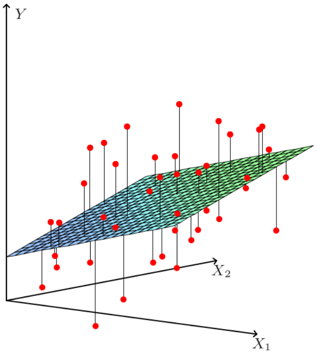
$$V(\mathcal{Y}) = \frac{r^2}{2}\sum_{t=1}^{T}\alpha_t = \frac{r^2}{2}\left(\log T - \log\log T + O\left(\frac{\log\log T}{\log T}\right)\right).$$



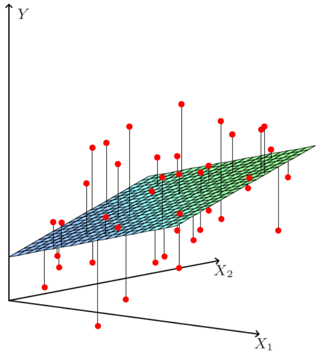
Outline

- Computing minimax optimal strategies.
- Part 1: Euclidean loss.
- **Part 2: Linear regression.**
  - Fixed design.
  - Minimax strategy is regularized least squares.
  - Box and ellipsoid constraints.
  - Adversarial covariates.
- Time series forecasting.

Online fixed design linear regression (with Koolen, Malek, Takimoto, Warmuth)

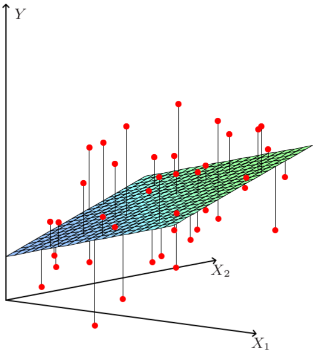


Online fixed design linear regression (with Koolen, Malek, Takimoto, Warmuth)



Protocol

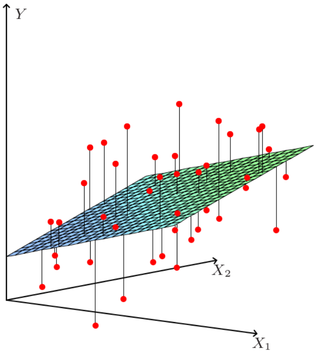
Online fixed design linear regression (with Koolen, Malek, Takimoto, Warmuth)



Protocol

Given: T;

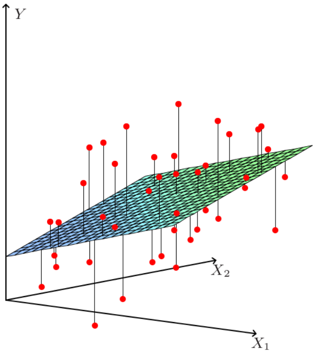
Online fixed design linear regression (with Koolen, Malek, Takimoto, Warmuth)



Protocol

Given: T; $x_1, \ldots, x_T \in \mathbb{R}^p$;

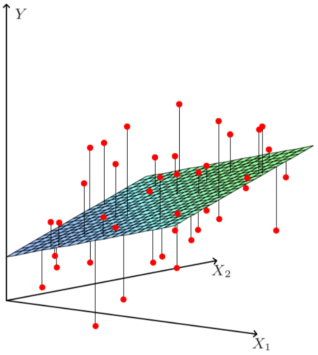
Online fixed design linear regression (with Koolen, Malek, Takimoto, Warmuth)



Protocol

Given: T; $x_1, \ldots, x_T \in \mathbb{R}^p$; $\mathcal{Y}^T \subset \mathbb{R}^T$.

Online fixed design linear regression (with Koolen, Malek, Takimoto, Warmuth)

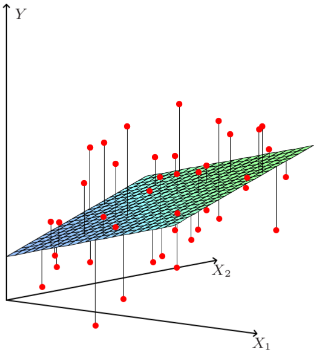


Protocol

Given: T; $x_1, \ldots, x_T \in \mathbb{R}^p$; $\mathcal{Y}^T \subset \mathbb{R}^T$.

For $t = 1, 2, \ldots, T$:

Online fixed design linear regression (with Koolen, Malek, Takimoto, Warmuth)

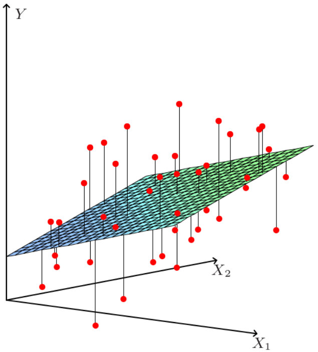


Protocol

Given: T; $x_1, \ldots, x_T \in \mathbb{R}^p$; $\mathcal{Y}^T \subset \mathbb{R}^T$.
For $t = 1, 2, \ldots, T$:

- Learner predicts $\hat{y}_t \in \mathbb{R}$

Online fixed design linear regression (with Koolen, Malek, Takimoto, Warmuth)

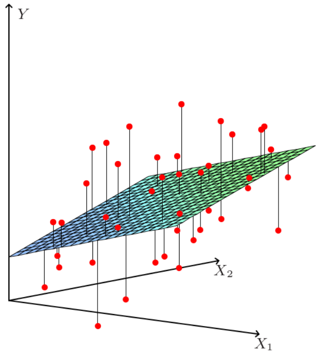


Protocol

Given: T; $x_1, \ldots, x_T \in \mathbb{R}^p$; $\mathcal{Y}^T \subset \mathbb{R}^T$.
For $t = 1, 2, \ldots, T$:

- Learner predicts $\hat{y}_t \in \mathbb{R}$
- Adversary reveals $y_t \in \mathbb{R}$

Online fixed design linear regression (with Koolen, Malek, Takimoto, Warmuth)

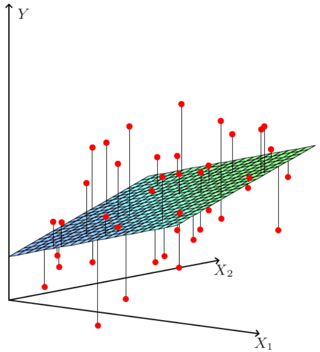


Protocol

Given: T; $x_1, \ldots, x_T \in \mathbb{R}^p$; $\mathcal{Y}^T \subset \mathbb{R}^T$.
For $t = 1, 2, \ldots, T$:

- Learner predicts $\hat{y}_t \in \mathbb{R}$
- Adversary reveals $y_t \in \mathbb{R}$ $\quad (y_1^T \in \mathcal{Y}^T)$

Online fixed design linear regression (with Koolen, Malek, Takimoto, Warmuth)

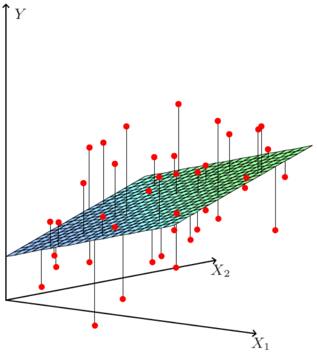


Protocol

Given: T; $x_1, \ldots, x_T \in \mathbb{R}^p$; $\mathcal{Y}^T \subset \mathbb{R}^T$.
For $t = 1, 2, \ldots, T$:

- Learner predicts $\hat{y}_t \in \mathbb{R}$
- Adversary reveals $y_t \in \mathbb{R}$ $\quad (y_1^T \in \mathcal{Y}^T)$
- Learner incurs loss $(\hat{y}_t - y_t)^2$.

Online fixed design linear regression (with Koolen, Malek, Takimoto, Warmuth)



Protocol

Given: T; $x_1, \ldots, x_T \in \mathbb{R}^p$; $\mathcal{Y}^T \subset \mathbb{R}^T$.
For $t = 1, 2, \ldots, T$:

- Learner predicts $\hat{y}_t \in \mathbb{R}$
- Adversary reveals $y_t \in \mathbb{R}$ $\quad (y_1^T \in \mathcal{Y}^T)$
- Learner incurs loss $(\hat{y}_t - y_t)^2$.

$$\text{Regret} = \sum_{t=1}^{T} (\hat{y}_t - y_t)^2 - \min_{\beta \in \mathbb{R}^p} \sum_{t=1}^{T} \left(\beta^\top x_t - y_t\right)^2.$$

Online fixed design linear regression

Online linear regression: previous work

- (Foster, 1991): ℓ_2-regularized least squares.
- (Cesa-Bianchi et al, 1996): ℓ_2-constrained least squares.
- (Kivinen and Warmuth, 1997): exponentiated gradient (relative entropy regularization).
- (Vovk, 1998): aggregating algorithm.
- (Forster, 1999; Azoury and Warmuth, 2001): aggregating algorithm is last-step minimax.

Linear regression in a probabilistic setting

Ordinary least squares (linear model, uncorrelated errors)

Given $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$,

Linear regression in a probabilistic setting

Ordinary least squares (linear model, uncorrelated errors)

Given $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$, choose

$$\hat{\beta} = \left(\sum_{t=1}^{n} x_t x_t^\top \right)^{-1} \sum_{t=1}^{n} x_t y_t ,$$

Linear regression in a probabilistic setting

Ordinary least squares (linear model, uncorrelated errors)

Given $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$, choose

$$\hat{\beta} = \left(\sum_{t=1}^{n} x_t x_t^\top \right)^{-1} \sum_{t=1}^{n} x_t y_t,$$

and for a subsequent $x \in \mathbb{R}^p$, predict

$$\hat{y} = x^\top \hat{\beta} = x^\top \left(\sum_{t=1}^{n} x_t x_t^\top \right)^{-1} \sum_{t=1}^{n} x_t y_t,$$

Linear regression in a probabilistic setting

Ordinary least squares (linear model, uncorrelated errors)

Given $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$, choose

$$\hat{\beta} = \left(\sum_{t=1}^{n} x_t x_t^\top\right)^{-1} \sum_{t=1}^{n} x_t y_t,$$

and for a subsequent $x \in \mathbb{R}^p$, predict

$$\hat{y} = x^\top \hat{\beta} = x^\top \left(\sum_{t=1}^{n} x_t x_t^\top\right)^{-1} \sum_{t=1}^{n} x_t y_t,$$

A sequential version of OLS

$$\hat{y}_{n+1} := x_{n+1}^\top \left(\sum_{t=1}^{n} x_t x_t^\top\right)^{-1} \sum_{t=1}^{n} x_t y_t.$$

Linear regression in a probabilistic setting

Ordinary least squares (linear model, uncorrelated errors)

Given $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$, choose

$$\hat{\beta} = \left(\sum_{t=1}^{n} x_t x_t^\top \right)^{-1} \sum_{t=1}^{n} x_t y_t,$$

and for a subsequent $x \in \mathbb{R}^p$, predict

$$\hat{y} = x^\top \hat{\beta} = x^\top \left(\sum_{t=1}^{n} x_t x_t^\top \right)^{-1} \sum_{t=1}^{n} x_t y_t,$$

A sequential version of ridge regression

$$\hat{y}_{n+1} := x_{n+1}^\top \left(\sum_{t=1}^{n} x_t x_t^\top + \lambda I \right)^{-1} \sum_{t=1}^{n} x_t y_t.$$

Online fixed design linear regression

Fix $x_1, \ldots, x_T \in \mathbb{R}^p$.

Online fixed design linear regression

Fix $x_1, \ldots, x_T \in \mathbb{R}^p$. $\mathcal{Y}^T = \{(y_1, \ldots, y_T) : |y_t| \leq B_t\}$.

Online fixed design linear regression

Sufficient statistics

Fix $x_1, \ldots, x_T \in \mathbb{R}^p$. $\mathcal{Y}^T = \{(y_1, \ldots, y_T) : |y_t| \le B_t\}$.
Maintain statistics: $s_n = \sum_{t=1}^{n} y_t x_t$

Online fixed design linear regression

Sufficient statistics

Fix $x_1, \ldots, x_T \in \mathbb{R}^p$.

$\mathcal{Y}^T = \{(y_1, \ldots, y_T) : |y_t| \leq B_t\}$.

Maintain statistics: $s_n = \sum_{t=1}^n y_t x_t$

Minimax* strategy: linear

$$\hat{y}^*_{n+1} = x_{n+1}^\top P_{n+1} s_n.$$

Online fixed design linear regression

Sufficient statistics

Fix $x_1, \ldots, x_T \in \mathbb{R}^p$. $\mathcal{Y}^T = \{(y_1, \ldots, y_T) : |y_t| \le B_t\}$.
Maintain statistics: $s_n = \sum_{t=1}^n y_t x_t$

Minimax* strategy: linear

$$\hat{y}^*_{n+1} = x_{n+1}^\top P_{n+1} s_n.$$

$$P_n^{-1} = \sum_{t=1}^{n} x_t x_t^\top + \qquad .$$

Online fixed design linear regression

Sufficient statistics

Fix $x_1, \ldots, x_T \in \mathbb{R}^p$. $\mathcal{Y}^T = \{(y_1, \ldots, y_T) : |y_t| \le B_t\}$.
Maintain statistics: $s_n = \sum_{t=1}^n y_t x_t$

Minimax* strategy: linear

$$\hat{y}^*_{n+1} = x_{n+1}^\top P_{n+1} s_n.$$

$$P_n^{-1} = \sum_{t=1}^n x_t x_t^\top + \sum_{t=n+1}^T \frac{x_t^\top P_t x_t}{1 + x_t^\top P_t x_t} x_t x_t^\top.$$

Online fixed design linear regression

Sufficient statistics

Fix $x_1, \ldots, x_T \in \mathbb{R}^p$. $\mathcal{Y}^T = \{(y_1, \ldots, y_T) : |y_t| \leq B_t\}$.

Maintain statistics: $s_n = \sum_{t=1}^n y_t x_t$

Minimax* strategy: linear

$$\hat{y}^*_{n+1} = x_{n+1}^\top P_{n+1} s_n.$$

Maximin distribution:

$$\Pr(\pm B_{n+1}) = \frac{1}{2} \pm \frac{x_{n+1}^\top P_{n+1} s_n}{2B_{n+1}}$$

$$P_n^{-1} = \sum_{t=1}^n x_t x_t^\top + \sum_{t=n+1}^T \frac{x_t^\top P_t x_t}{1 + x_t^\top P_t x_t} x_t x_t^\top.$$

Online fixed design linear regression

Sufficient statistics

Fix $x_1, \ldots, x_T \in \mathbb{R}^p$.

$\mathcal{Y}^T = \{(y_1, \ldots, y_T) : |y_t| \leq B_t\}$.

Maintain statistics: $s_n = \sum_{t=1}^n y_t x_t$

Value-to-go: quadratic

$$s_n^\top P_n s_n - \sigma_n^2 + \sum_{t=n+1}^{T} B_t^2 x_t^\top P_t x_t.$$

Minimax* strategy: linear

$$\hat{y}_{n+1}^* = x_{n+1}^\top P_{n+1} s_n.$$

Maximin distribution:

$$\Pr\left(\pm B_{n+1}\right) = \frac{1}{2} \pm \frac{x_{n+1}^\top P_{n+1} s_n}{2B_{n+1}}$$

$$P_n^{-1} = \sum_{t=1}^{n} x_t x_t^\top + \sum_{t=n+1}^{T} \frac{x_t^\top P_t x_t}{1 + x_t^\top P_t x_t} x_t x_t^\top.$$

Online fixed design linear regression

Sufficient statistics

Fix $x_1, \ldots, x_T \in \mathbb{R}^p$. $\mathcal{Y}^T = \{(y_1, \ldots, y_T) : |y_t| \leq B_t\}$.
Maintain statistics: $s_n = \sum_{t=1}^n y_t x_t$, $\sigma_n^2 = \sum_{t=1}^n y_t^2$.

Value-to-go: quadratic

$$s_n^\top P_n s_n - \sigma_n^2 + \sum_{t=n+1}^{T} B_t^2 x_t^\top P_t x_t.$$

Minimax* strategy: linear

$$\hat{y}_{n+1}^* = x_{n+1}^\top P_{n+1} s_n.$$

Maximin distribution:

$$\Pr(\pm B_{n+1}) = \frac{1}{2} \pm \frac{x_{n+1}^\top P_{n+1} s_n}{2 B_{n+1}}$$

$$P_n^{-1} = \sum_{t=1}^{n} x_t x_t^\top + \sum_{t=n+1}^{T} \frac{x_t^\top P_t x_t}{1 + x_t^\top P_t x_t} x_t x_t^\top.$$

Online fixed design linear regression

Sufficient statistics

Fix $x_1, \ldots, x_T \in \mathbb{R}^p$. $\mathcal{Y}^T = \{(y_1, \ldots, y_T) : |y_t| \le B_t\}$.
Maintain statistics: $s_n = \sum_{t=1}^n y_t x_t$, $\sigma_n^2 = \sum_{t=1}^n y_t^2$.

Value-to-go: quadratic

$$s_n^\top P_n s_n - \sigma_n^2 + \sum_{t=n+1}^{T} B_t^2 x_t^\top P_t x_t.$$

$*$ provided: $$B_n \ge \sum_{t=1}^{n-1} \left| x_n^\top P_n x_t \right| B_t.$$

Minimax* strategy: linear

$$\hat{y}_{n+1}^* = x_{n+1}^\top P_{n+1} s_n.$$

Maximin distribution:

$$\Pr\left(\pm B_{n+1}\right) = \frac{1}{2} \pm \frac{x_{n+1}^\top P_{n+1} s_n}{2 B_{n+1}}$$

$$P_n^{-1} = \sum_{t=1}^{n} x_t x_t^\top + \sum_{t=n+1}^{T} \frac{x_t^\top P_t x_t}{1 + x_t^\top P_t x_t} x_t x_t^\top.$$

Linear regression

Box constraints

$$\mathcal{Y}^T = \{(y_1, \ldots, y_T) : |y_n| \leq B_n\} \qquad B_n \geq \sum_{t=1}^{n-1} \left| x_n^\top P_n x_t \right| B_t.$$

Linear regression

Box constraints

$$\mathcal{Y}^T = \{(y_1, \dots, y_T) : |y_n| \le B_n\} \qquad B_n \ge \sum_{t=1}^{n-1} \left| x_n^\top P_n x_t \right| B_t .$$

Minimax strategy: linear

$$\hat{y}_n^* = x_n^\top P_n s_{n-1} .$$

Linear regression

Box constraints

$$\mathcal{Y}^T = \{(y_1, \ldots, y_T) : |y_n| \le B_n\} \qquad B_n \ge \sum_{t=1}^{n-1} \left| x_n^\top P_n x_t \right| B_t.$$

Minimax strategy: linear

$$\hat{y}_n^* = x_n^\top P_n s_{n-1}.$$

$$P_n^{-1} = \sum_{t=1}^{n} x_t x_t^\top + \sum_{t=n+1}^{T} \frac{x_t^\top P_t x_t}{1 + x_t^\top P_t x_t} x_t x_t^\top.$$

Linear regression

Box constraints

$$\mathcal{Y}^T = \{(y_1, \ldots, y_T) : |y_n| \le B_n\} \qquad B_n \ge \sum_{t=1}^{n-1} \left| x_n^\top P_n x_t \right| B_t.$$

$$\text{Regret} = \sum_{t=1}^{T} B_t^2 x_t^\top P_t x_t.$$

Minimax strategy: linear

$$\hat{y}_n^* = x_n^\top P_n s_{n-1}.$$

$$P_n^{-1} = \sum_{t=1}^{n} x_t x_t^\top + \sum_{t=n+1}^{T} \frac{x_t^\top P_t x_t}{1 + x_t^\top P_t x_t} x_t x_t^\top.$$

Linear regression

Box constraints

$$\mathcal{Y}^T = \{(y_1, \ldots, y_T) : |y_n| \le B_n\} \qquad B_n \ge \sum_{t=1}^{n-1} \left| x_n^\top P_n x_t \right| B_t.$$

$$\text{Regret} = \sum_{t=1}^{T} B_t^2 x_t^\top P_t x_t.$$

Minimax strategy: linear

$$\hat{y}_n^* = x_n^\top P_n s_{n-1}.$$

$$P_n^{-1} = \sum_{t=1}^{n} x_t x_t^\top + \sum_{t=n+1}^{T} \frac{x_t^\top P_t x_t}{1 + x_t^\top P_t x_t} x_t x_t^\top.$$

c.f. ridge regression: $\displaystyle\sum_{t=1}^{n} x_t x_t^\top + \lambda I.$

Linear regression

Box constraints

$$\mathcal{Y}^T = \{(y_1, \ldots, y_T) : |y_n| \leq B_n\} \qquad B_n \geq \sum_{t=1}^{n-1} \left| x_n^\top P_n x_t \right| B_t.$$

$$\text{Regret} = \sum_{t=1}^{T} B_t^2 x_t^\top P_t x_t.$$

Minimax strategy: linear

$$\hat{y}_n^* = x_n^\top P_n s_{n-1}.$$

Optimal shrinkage

$$P_n^{-1} = \sum_{t=1}^{n} x_t x_t^\top + \sum_{t=n+1}^{T} \frac{x_t^\top P_t x_t}{1 + x_t^\top P_t x_t} x_t x_t^\top.$$

c.f. ridge regression:

$$\sum_{t=1}^{n} x_t x_t^\top + \lambda I.$$

Linear regression

Box constraints

$$\mathcal{Y}^T = \{(y_1, \ldots, y_T) : |y_n| \leq B_n\} \qquad B_n \geq \sum_{t=1}^{n-1} \left| x_n^\top P_n x_t \right| B_t.$$

$$\text{Regret} = \sum_{t=1}^{T} B_t^2 x_t^\top P_t x_t.$$

Minimax strategy: linear

$$\hat{y}_n^* = x_n^\top P_n s_{n-1}.$$

Optimal shrinkage

$$P_n^{-1} = \sum_{t=1}^{n} x_t x_t^\top + \sum_{t=n+1}^{T} \frac{x_t^\top P_t x_t}{1 + x_t^\top P_t x_t} x_t x_t^\top.$$

c.f. ridge regression:

$$\sum_{t=1}^{n} x_t x_t^\top + \lambda I.$$

Linear regression: Regret

Theorem

$$\max_{x_1,\dots x_T} \sum_{t=1}^{T} x_t^\top P_t x_t \leq p\left(1 + 2\ln\left(1 + \frac{T}{2}\right)\right).$$

Linear regression

Ellipsoid constraints

$$\mathcal{Y}_R^T = \left\{ (y_1, \ldots, y_T) : \sum_{t=1}^{T} y_t^2 x_t^\top P_t x_t \leq R \right\}.$$

Linear regression

Ellipsoid constraints

$$\mathcal{Y}_R^T = \left\{ (y_1, \ldots, y_T) : \sum_{t=1}^{T} y_t^2 x_t^\top P_t x_t \leq R \right\}.$$

Minimax strategy: linear

$$\hat{y}_n^* = x_n^\top P_n s_{n-1}.$$

Linear regression

Ellipsoid constraints

$$\mathcal{Y}_R^T = \left\{ (y_1, \ldots, y_T) : \sum_{t=1}^{T} y_t^2 x_t^\top P_t x_t \leq R \right\}.$$

Minimax regret $= R$.

Minimax strategy: linear

$$\hat{y}_n^* = x_n^\top P_n s_{n-1}.$$

Linear regression

Ellipsoid constraints

$$\mathcal{Y}_R^T = \left\{ (y_1, \ldots, y_T) : \sum_{t=1}^{T} y_t^2 x_t^\top P_t x_t \leq R \right\}.$$

Minimax regret $= R$.

Minimax strategy: linear

$$\hat{y}_n^* = x_n^\top P_n s_{n-1}. \qquad \text{(MM)}$$

Equalizer property

For all $y_1, \ldots, y_T$,

$$\text{Regret of (MM)} := \sum_{t=1}^{T} (\hat{y}_t - y_t)^2 - \min_{\beta \in \mathbb{R}^p} \sum_{t=1}^{T} \left(\beta^\top x_t - y_t \right)^2$$

Linear regression

Ellipsoid constraints

$$\mathcal{Y}_R^T = \left\{ (y_1, \ldots, y_T) : \sum_{t=1}^{T} y_t^2 x_t^\top P_t x_t \leq R \right\}.$$

Minimax regret $= R$.

Minimax strategy: linear

$$\hat{y}_n^* = x_n^\top P_n s_{n-1}. \qquad \text{(MM)}$$

Equalizer property

For all $y_1, \ldots, y_T$,

$$\begin{aligned} \text{Regret of (MM)} &:= \sum_{t=1}^{T} (\hat{y}_t - y_t)^2 - \min_{\beta \in \mathbb{R}^p} \sum_{t=1}^{T} \left(\beta^\top x_t - y_t \right)^2 \\ &= \sum_{t=1}^{T} y_t^2 x_t^\top P_t x_t. \end{aligned}$$

Linear regression: Adversarial covariates

Recall:

$$P_T^{-1} = \sum_{t=1}^{T} x_t x_t^\top,$$

$$P_n^{-1} = \sum_{t=1}^{n} x_t x_t^\top + \sum_{t=n+1}^{T} \frac{x_t^\top P_t x_t}{1 + x_t^\top P_t x_t} x_t x_t^\top.$$

Define

$$P_0^{-1} = \sum_{q=1}^{T} \frac{x_q^\top P_q x_q}{1 + x_q^\top P_q x_q} x_q x_q^\top \succeq 0.$$

A reformulation

$$P_0^{-1} = \sum_{q=1}^{T} \frac{x_q^\top P_q x_q}{1 + x_q^\top P_q x_q} x_q x_q^\top \succeq 0.$$

$$\boxed{P_{t+1} = P_t - \frac{a_t}{b_t^2} P_t x_{t+1} x_{t+1}^\top P_t,}$$

where
$$a_t = \frac{\sqrt{4b_t^2 + 1} - 1}{\sqrt{4b_t^2 + 1} + 1},$$
$$b_t^2 = x_{t+1}^\top P_t x_{t+1}.$$

Linear regression

Legal covariate sequences

For any $t \geq 0$, any $x_1, \ldots, x_t$ and any P_t, the following two conditions are equivalent.

1. There is a $T \geq t$ and a sequence $x_{t+1}, \ldots, x_T$ such that
$$P_T^{-1} = \sum_{q=1}^{T} x_q x_q^\top .$$
2. $P_t^{-1} \succeq \sum_{q=1}^{t} x_q x_q^\top$.

Adversarial covariates

Thus, each $P_0 \succeq 0$ (a 'covariance budget') defines a set of sequences $x_1, \ldots, x_T$ (and corresponding suitable bounds on $y_1, \ldots, y_T$).
The same strategy is optimal for each of these sequences.

Linear regression

One-step constraint

Suppose we have $P_t^{-1} \succeq \sum_{q=1}^{t} x_q x_q^\top$.
Then x_{t+1} satisfies the consistency condition

$$P_{t+1}^{-1} \succeq \sum_{q=1}^{t+1} x_q x_q^\top,$$

iff

1. x_{t+1} is orthogonal to the kernel of $P_t^{-1} - \sum_{q=1}^{t} x_q x_q^\top$, and
2. $x_{t+1}^\top P_t x_{t+1} \leq d(\hat{x}_{t+1}) + \sqrt{d(\hat{x}_{t+1})}$,

where $\hat{x}_{t+1} = x_{t+1}/\|x_{t+1}\|$ and

$$d(\hat{x}) = \frac{\hat{x}^\top P_t \hat{x}}{\hat{x}^\top \left(P_t + P_t \left[\left(\sum_{q=1}^{t} x_q x_q^\top \right)^{-1} - P_t \right]^{-1} P_t \right) \hat{x}}.$$

Linear regression

$$\hat{y}_n^* = x_n^\top P_n s_{n-1}$$

- Minimax optimal for two families of label constraints: box constraints and problem-weighted ℓ_2 norm constraints.

Linear regression

$$\hat{y}_n^* = x_n^\top P_n s_{n-1}$$

- Minimax optimal for two families of label constraints: box constraints and problem-weighted ℓ_2 norm constraints.
- Strategy does not need to know the constraints.

Linear regression

$$\hat{y}_n^* = x_n^\top P_n s_{n-1}$$

- Minimax optimal for two families of label constraints: box constraints and problem-weighted ℓ_2 norm constraints.
- Strategy does not need to know the constraints.
- Regret is $O(p \log T)$.

Linear regression

$$\hat{y}_n^* = x_n^\top P_n s_{n-1}$$

- Minimax optimal for two families of label constraints: box constraints and problem-weighted ℓ_2 norm constraints.
- Strategy does not need to know the constraints.
- Regret is $O(p \log T)$.
- Same strategy is optimal for covariate sequences consistent with some 'covariance budget' P_0.

Outline

- Computing minimax optimal strategies.
- Part 1: Euclidean loss.
- Part 2: Linear regression.
- **Time series forecasting.**

Other games with efficient minimax optimal strategies

Time series forecasting

(with Yasin Abbasi-Yadkori, Wouter Koolen, Alan Malek)

$$\min_{a_1}\max_{x_1}\cdots\min_{a_T}\max_{x_T}\underbrace{\sum_{t=1}^{T}\|a_t-x_t\|^2}_{\text{Loss of Learner}}$$

$$-\min_{\hat{a}_1,\ldots,\hat{a}_T}\underbrace{\sum_{t=1}^{T}\|\hat{a}_t-x_t\|^2}_{\text{Loss of Comparator}}+\lambda_T\underbrace{\sum_{t=1}^{T+1}\|\hat{a}_t-\hat{a}_{t-1}\|^2}_{\text{Comparator Complexity}}.$$

- Expression for regret when x_t bounded. (And a bound when it is not.)
- Minimax strategy makes linear predictions.
- Regret is $\Theta\left(\frac{T}{\sqrt{1+\lambda_T}}\right)$.
- More generally, penalize comparator by the energy of the innovations of a time series model. Efficient linear minimax strategy. Regret?

Outline

- Computing minimax optimal strategies.
- Part 1: Euclidean loss.
- Part 2: Linear regression.
- Time series forecasting.