Topics in Prediction and Learning
Lecture 1:
Optimal Universal Prediction—Quadratic Loss

Peter Bartlett

Computer Science and Statistics
University of California at Berkeley

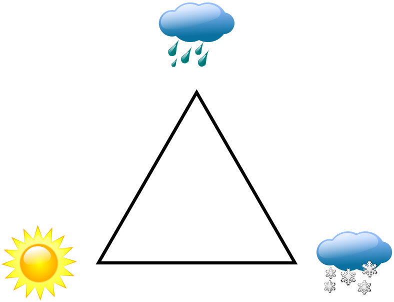
Mathematical Sciences
Queensland University of Technology

27 February–9 March, 2017
CREST, ENSAE

Online Prediction as a Zero-Sum Game

A repeated game:

At round t:

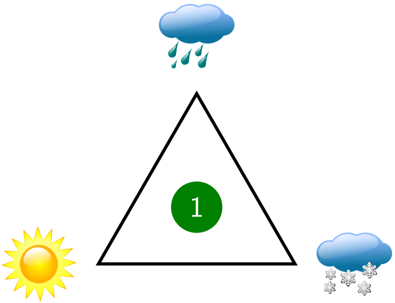


Online Prediction as a Zero-Sum Game

A repeated game:

At round t:

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

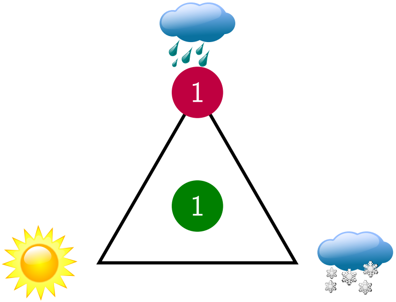


Online Prediction as a Zero-Sum Game

A repeated game:

At round t:

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

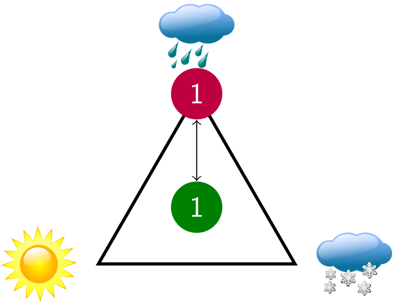


Online Prediction as a Zero-Sum Game

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 $\ell(a_t, y_t)$.



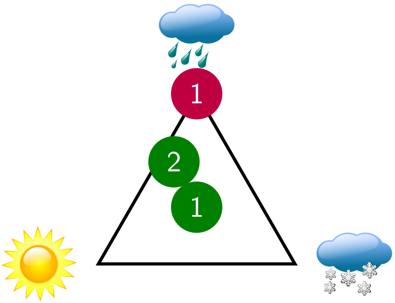
$$\ell(a_t, y_t) = \|a_t - y_t\|^2.$$

Online Prediction as a Zero-Sum Game

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 $\ell(a_t, y_t)$.

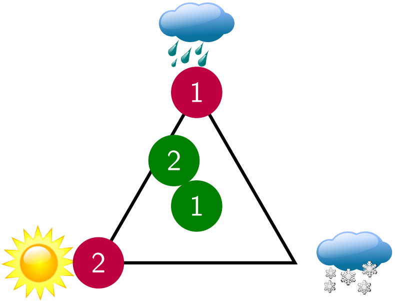


Online Prediction as a Zero-Sum Game

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 $\ell(a_t, y_t)$.

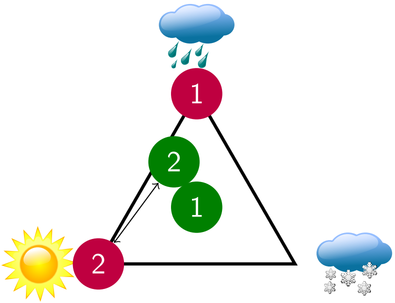


Online Prediction as a Zero-Sum Game

A repeated game:

At round t:

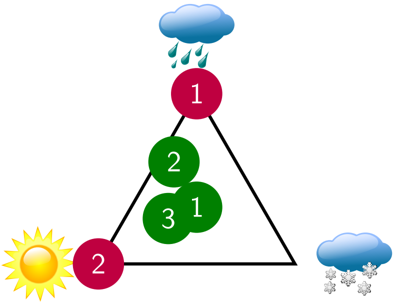
1. Player chooses prediction $a_t \in \mathcal{A}$.
2. Adversary chooses outcome $y_t \in \mathcal{Y}$.
3. Player incurs loss $\ell(a_t, y_t)$.



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 $\ell(a_t, y_t)$.

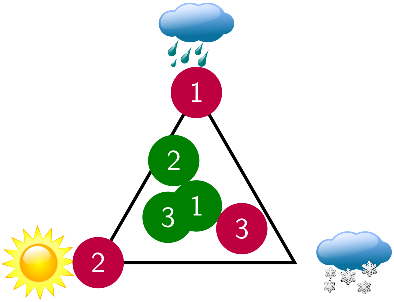


Online Prediction as a Zero-Sum Game

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 $\ell(a_t, y_t)$.

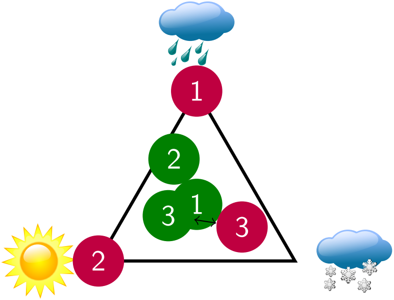


Online Prediction as a Zero-Sum Game

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 as a Zero-Sum Game

A repeated game:

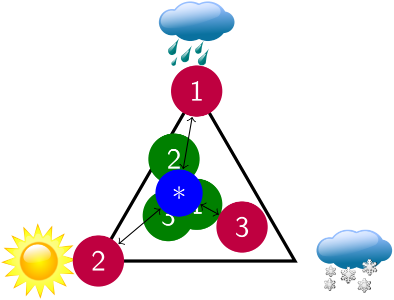
At round t:

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 as a Zero-Sum Game

A repeated game:

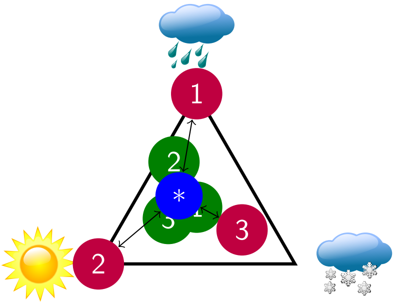
At round t:

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 $\mathcal{C}$:

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

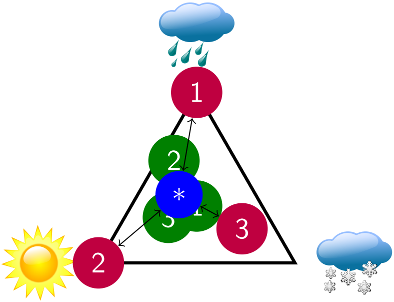


Online Prediction as a Zero-Sum Game

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 $\mathcal{C}$:

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

Examples:

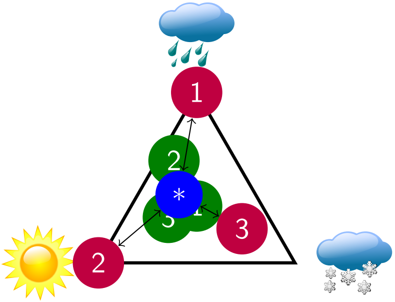
- $\mathcal{A} =$ simplex
  - calibration

Online Prediction as a Zero-Sum Game

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 $\mathcal{C}$:

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

Examples:

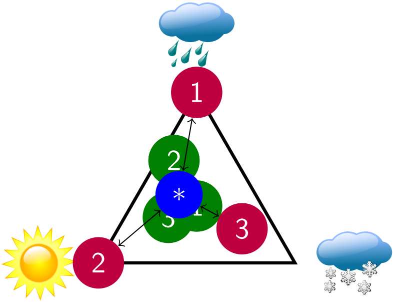
- $\mathcal{A}$ = simplex
  - calibration
  - prediction with expert advice

Online Prediction as a Zero-Sum Game

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 $\mathcal{C}$:

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

Examples:

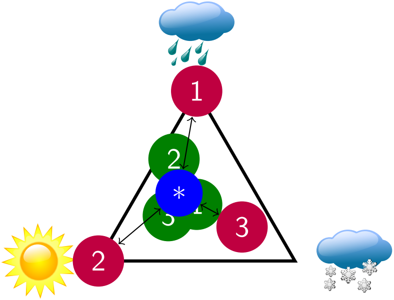
- $\mathcal{A}$ = simplex
  - calibration
  - prediction with expert advice (ℓ linear)

Online Prediction as a Zero-Sum Game

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 $\mathcal{C}$:

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

Examples:

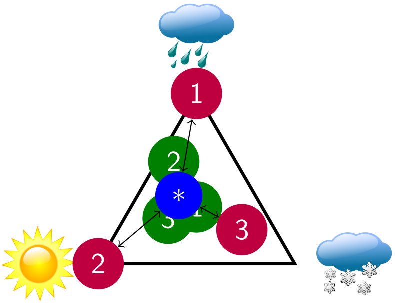
- $\mathcal{A}$ = simplex
  - calibration
  - prediction with expert advice (ℓ linear)
  - portfolio optimization

Online Prediction as a Zero-Sum Game

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 $\mathcal{C}$:

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

Examples:

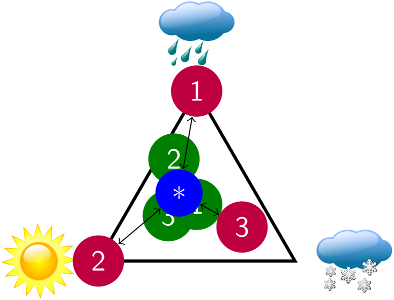
- $\mathcal{C}$ = linear model

Online Prediction as a Zero-Sum Game

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 $\mathcal{C}$:

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

Examples:

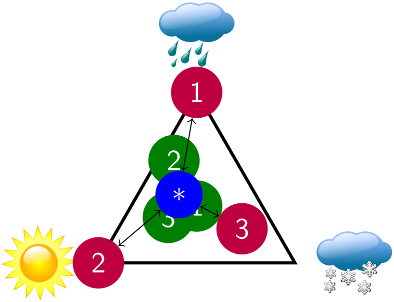
- $\mathcal{C}$ = linear model
  - linear regression

Online Prediction as a Zero-Sum Game

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 $\mathcal{C}$:

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

Examples:

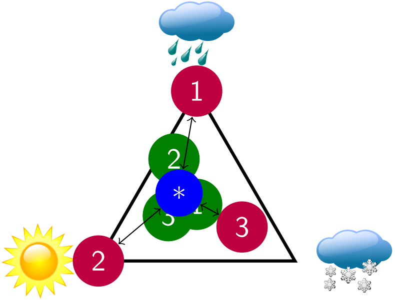
- $\mathcal{C}$ = smooth sequences

Online Prediction as a Zero-Sum Game

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 $\mathcal{C}$:

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

Examples:

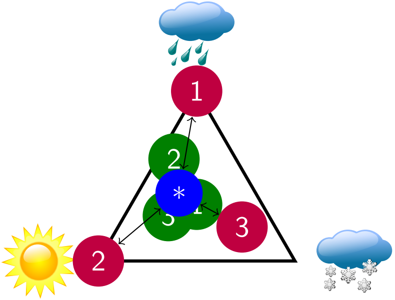
- $\mathcal{C}$ = smooth sequences
  - tracking

Online Prediction as a Zero-Sum Game

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 $\mathcal{C}$:

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

Examples:

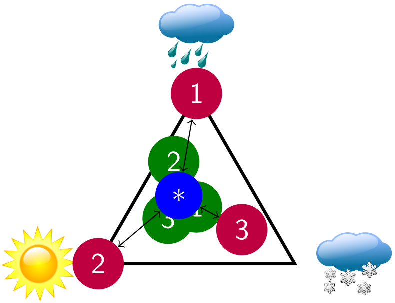
- $\mathcal{C}$ = smooth sequences
  - tracking
  - nonparametric regression

Online Prediction as a Zero-Sum Game

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 $\mathcal{C}$:

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

Examples:

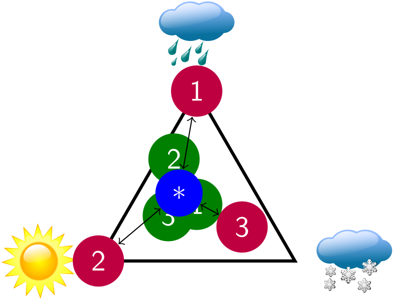
- $\mathcal{C}$ = probability model

Online Prediction as a Zero-Sum Game

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 $\mathcal{C}$:

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

Examples:

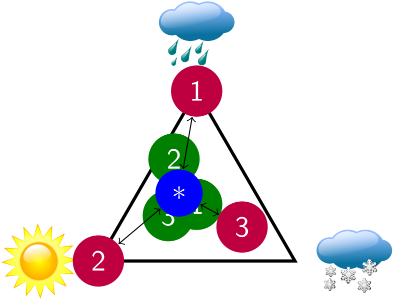
- $\mathcal{C}$ = probability model
  - density estimation

Online Prediction as a Zero-Sum Game

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 $\mathcal{C}$:

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

Examples:

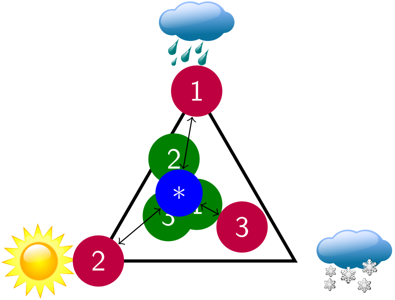
- $\mathcal{C} =$ option trades,
  $\mathcal{A} =$ asset trades

Online Prediction as a Zero-Sum Game

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 $\mathcal{C}$:

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

Examples:

- $\mathcal{C}$ = option trades,
  $\mathcal{A}$ = asset trades
  - option pricing

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.
- Can demonstrate robustness of statistical methods.

Online Prediction Games: Why?

- Universal prediction:
  very weak assumptions on process generating the data.
- Deterministic heart of a decision problem.
- Can demonstrate robustness of 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.
- Can demonstrate robustness of statistical methods.
- Typically streaming, so very scalable.
- Focus in this lecture: 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:

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

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

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

Questions

Online Prediction Games

Questions

- Minimax regret?

Online Prediction Games

Questions

- Minimax regret? (or rates?)

Online Prediction Games

Questions

- Minimax regret? (or rates?)
- Optimal player strategy?

Online Prediction Games

Questions

- Minimax regret? (or rates?)
- Optimal player strategy? (rate optimal?)

Online Prediction Games

Questions

- Minimax regret? (or rates?)
- Optimal player strategy? (rate optimal?)
- Efficient algorithms?

Online Prediction Games

Questions

- Minimax regret? (or rates?)
- Optimal player strategy? (rate optimal?)
- Efficient algorithms?
- Horizon-free?

Online Prediction Games

Questions

- Minimax regret? (or rates?)
- Optimal player strategy? (rate optimal?)
- Efficient algorithms?
- Horizon-free?
- Optimal adversary strategy?

Online Prediction Games

Questions

- Minimax regret? (or rates?)
- Optimal player strategy? (rate optimal?)
- Efficient algorithms?
- Horizon-free?
- Optimal adversary strategy?
- How do they depend on ℓ?

Online Prediction Games

Questions

- Minimax regret? (or rates?)
- Optimal player strategy? (rate optimal?)
- Efficient algorithms?
- Horizon-free?
- Optimal adversary 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? (or rates?)
- Optimal player strategy? (rate optimal?)
- Efficient algorithms?
- Horizon-free?
- Optimal adversary 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? (or rates?)
- Optimal player strategy? (rate optimal?)
- Efficient algorithms?
- Horizon-free?
- Optimal adversary 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. $-\log a(y)$, $a \in \{p_\theta : \theta \in \Theta\}$.

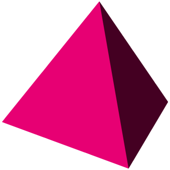
Online Prediction Games

Questions

- Minimax regret? (or rates?)
- Optimal player strategy? (rate optimal?)
- Efficient algorithms?
- Horizon-free?
- Optimal adversary strategy?
- How do they depend on ℓ, $\mathcal{Y}$, $\mathcal{A}$, $\mathcal{C}$?

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

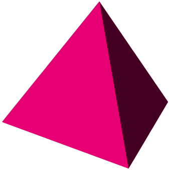
Online Prediction Games

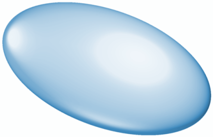
Questions

- Minimax regret? (or rates?)
- Optimal player strategy? (rate optimal?)
- Efficient algorithms?
- Horizon-free?
- Optimal adversary strategy?
- How do they depend on ℓ, $\mathcal{Y}$, $\mathcal{A}$, $\mathcal{C}$?

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

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





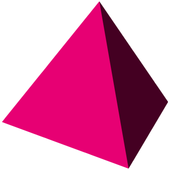
Online Prediction Games

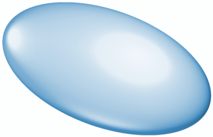
Questions

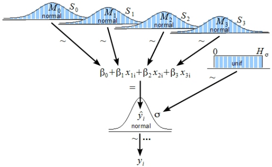
- Minimax regret? (or rates?)
- Optimal player strategy? (rate optimal?)
- Efficient algorithms?
- Horizon-free?
- Optimal adversary strategy?
- How do they depend on ℓ, $\mathcal{Y}$, $\mathcal{A}$, $\mathcal{C}$?

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 for Today's Lecture

Outline for Today's Lecture

- Computing minimax optimal strategies.

Outline for Today's Lecture

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

Outline for Today's Lecture

- 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.

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!
$\{(y_1, \ldots, y_t)\}$ is large.
V might be complex.

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!
$\{(y_1, \ldots, y_t)\}$ is large.
V might be complex.

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.**
  - $\mathcal{Y} =$ ball
  - $\mathcal{Y} =$ simplex
  - Closed, bounded $\mathcal{Y}$
  - Hilbert space
  - $\mathcal{Y} =$ ellipsoid
- Part 2: Linear regression.

Online prediction with quadratic loss

(Takimoto, Warmuth, 2000), (Koolen, Malek, B., 2014)

Online prediction with quadratic loss

Euclidean loss

$$\ell(\hat{y}, y) = \frac{1}{2}\|\hat{y} - y\|^2.$$

(Takimoto, Warmuth, 2000), (Koolen, Malek, B., 2014)

Online prediction with quadratic loss

Constraints

Strategy chooses $\hat{y}_n \in \mathbb{R}^d$.

Euclidean loss

$$\ell(\hat{y}, y) = \frac{1}{2} \|\hat{y} - y\|^2 .$$

(Takimoto, Warmuth, 2000), (Koolen, Malek, B., 2014)

Online prediction with quadratic loss

Constraints

Strategy chooses $\hat{y}_n \in \mathbb{R}^d$.
Adversary chooses $y_n \in \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathbb{R}^d$.

Euclidean loss

$$\ell(\hat{y}, y) = \frac{1}{2} \|\hat{y} - y\|^2 .$$

(Takimoto, Warmuth, 2000), (Koolen, Malek, B., 2014)

Online prediction with quadratic loss

Constraints

Strategy chooses $\hat{y}_n \in \mathbb{R}^d$.
Adversary chooses $y_n \in \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathbb{R}^d$.

Euclidean loss

$$\ell(\hat{y}, y) = \frac{1}{2} \|\hat{y} - y\|^2 .$$

$$\text{Regret} \;=$$

(Takimoto, Warmuth, 2000), (Koolen, Malek, B., 2014)

Online prediction with quadratic loss

Constraints

Strategy chooses $\hat{y}_n \in \mathbb{R}^d$.
Adversary chooses $y_n \in \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathbb{R}^d$.

Euclidean loss

$$\ell(\hat{y}, y) = \frac{1}{2} \|\hat{y} - y\|^2 .$$

$$\text{Regret} = \sum_{t=1}^{n} \ell(\hat{y}_t, y_t) -$$

(Takimoto, Warmuth, 2000), (Koolen, Malek, B., 2014)

Online prediction with quadratic loss

Constraints

Strategy chooses $\hat{y}_n \in \mathbb{R}^d$.
Adversary chooses $y_n \in \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathbb{R}^d$.

Euclidean loss

$$\ell(\hat{y}, y) = \frac{1}{2} \|\hat{y} - y\|^2 .$$

$$\text{Regret} = \sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{a \in \mathbb{R}^d} \sum_{t=1}^{n} \ell(a, y_t).$$

(Takimoto, Warmuth, 2000), (Koolen, Malek, B., 2014)

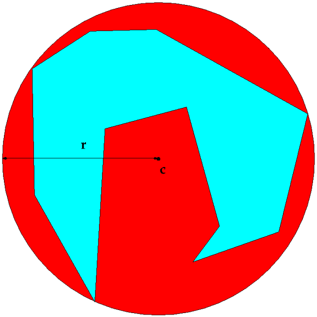
Online prediction with quadratic loss

Online prediction with quadratic loss

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

Online prediction with quadratic loss

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



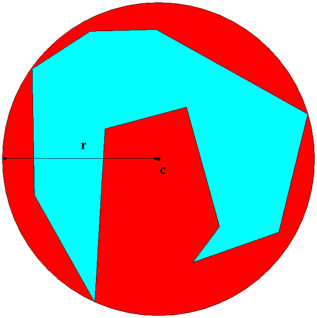
Online prediction with quadratic loss

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

Theorem

- For closed, bounded $\mathcal{Y} \subset \mathbb{R}^d$, minimax strategy is empirical minimizer, shrunk towards c:

$$a^*_{t+1} = t\alpha_{t+1}\bar{y}_t + (1 - t\alpha_{t+1})c.$$



Online prediction with quadratic loss

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

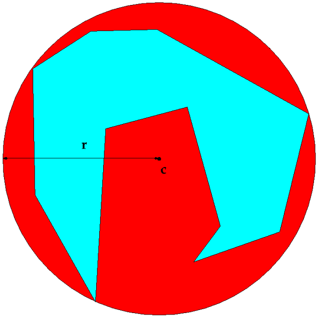
Theorem

- For closed, bounded $\mathcal{Y} \subset \mathbb{R}^d$, minimax strategy is empirical minimizer, shrunk towards c:

$$a^*_{t+1} = t\alpha_{t+1}\bar{y}_t + (1 - t\alpha_{t+1})c.$$

- Optimal regret: $\displaystyle \frac{r^2}{2} \sum_{t=1}^{T} \alpha_t;$

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



Outline

- Computing minimax optimal strategies.
- Part 1: Euclidean loss.
  - $\mathcal{Y}$ = **ball**
  - $\mathcal{Y}$ = simplex
  - Closed, bounded $\mathcal{Y}$
  - Hilbert space
  - $\mathcal{Y}$ = ellipsoid
- Part 2: Linear regression.

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

Value-to-go: quadratic

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

Online prediction with quadratic loss on the ball

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

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

Value-to-go: quadratic

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

$$\alpha_T = \frac{1}{T}, \quad \alpha_n = \alpha_{n+1}^2 + \alpha_{n+1}$$

Online prediction with quadratic loss on the ball

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

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

Value-to-go: quadratic

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

$$\alpha_T = \frac{1}{T}, \quad \alpha_n = \alpha_{n+1}^2 + \alpha_{n+1}$$

Minimax strategy: affine

$$a_n^* = c + \alpha_n \sum_{t=1}^{n-1}(y_t - c).$$

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

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

$$\alpha_T = \frac{1}{T}, \quad \alpha_n = \alpha_{n+1}^2 + \alpha_{n+1}$$

Minimax strategy: affine

$$a_n^* = c + \alpha_n \sum_{t=1}^{n-1}(y_t - c).$$

$$a_n^* = (n-1)\alpha_n \bar{y}_{n-1} + (1-(n-1)\alpha_n)c.$$

Online prediction with quadratic loss on the ball

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

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

Value-to-go: quadratic

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

$$\alpha_T = \frac{1}{T}, \quad \alpha_n = \alpha_{n+1}^2 + \alpha_{n+1} \le \frac{1}{n}.$$

Minimax strategy: affine

$$a_n^* = c + \alpha_n\sum_{t=1}^{n-1}(y_t - c).$$

$$a_n^* = (n-1)\alpha_n\bar{y}_{n-1} + (1-(n-1)\alpha_n)c.$$

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

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

$$\alpha_T = \frac{1}{T}, \quad \alpha_n = \alpha_{n+1}^2 + \alpha_{n+1} \le \frac{1}{n}.$$

Minimax strategy: affine

$$a_n^* = c + \alpha_n \sum_{t=1}^{n-1}(y_t - c).$$

$$a_n^* = (n-1)\alpha_n \bar{y}_{n-1} + (1-(n-1)\alpha_n)c.$$

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

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

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

$$\alpha_T = \frac{1}{T}, \quad \alpha_n = \alpha_{n+1}^2 + \alpha_{n+1} \leq \frac{1}{n}.$$

Minimax strategy: affine

$$a_n^* = c + \alpha_n \sum_{t=1}^{n-1}(y_t - c).$$

$$a_n^* = (n-1)\alpha_n \bar{y}_{n-1} + (1-(n-1)\alpha_n)c.$$

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

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

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

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: wlog, $c = 0$.

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: wlog, $c = 0$.

$$V(y_1, \ldots, y_T) := -\min_a \frac{1}{2} \sum_{t=1}^{T} \|a - y_t\|^2$$

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: wlog, $c = 0$.

$$V(y_1, \ldots, y_T) := -\min_a \frac{1}{2} \sum_{t=1}^{T} \|a - y_t\|^2 = -\frac{1}{2} \sum_{t=1}^{T} \|\bar{y} - y_t\|^2$$

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: wlog, $c = 0$.

$$V(y_1, \ldots, y_T) := -\min_a \frac{1}{2} \sum_{t=1}^{T} \|a - y_t\|^2 = -\frac{1}{2} \sum_{t=1}^{T} \|\bar{y} - y_t\|^2$$

$$= -\frac{1}{2} \sum_{t=1}^{T} \|\frac{s_T}{T} - y_t\|^2$$

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: wlog, $c = 0$.

$$V(y_1, \ldots, y_T) := -\min_a \frac{1}{2} \sum_{t=1}^{T} \|a - y_t\|^2 = -\frac{1}{2} \sum_{t=1}^{T} \|\bar{y} - y_t\|^2$$

$$= -\frac{1}{2} \sum_{t=1}^{T} \|\frac{s_T}{T} - y_t\|^2 = \frac{1}{2}\left(\frac{1}{T}\|s_T\|^2 - \sigma_T^2\right).$$

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: wlog, $c = 0$.

$$V(y_1, \ldots, y_T) := -\min_a \frac{1}{2} \sum_{t=1}^{T} \|a - y_t\|^2 = -\frac{1}{2} \sum_{t=1}^{T} \|\bar{y} - y_t\|^2$$

$$= -\frac{1}{2} \sum_{t=1}^{T} \left\|\frac{s_T}{T} - y_t\right\|^2 = \frac{1}{2} \left(\frac{1}{T}\|s_T\|^2 - \sigma_T^2\right).$$

i.e., $\quad V(s_T, \sigma_T^2, T) = \frac{1}{2} \left(\alpha_T \|s_T\|^2 - \sigma_T^2\right).$

Online prediction with quadratic loss on the ball

Proof idea

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

Online prediction with quadratic loss on the ball

Proof idea

$$
\begin{aligned}
V(y_1, \ldots, y_{t-1}) &= \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right) \\
&= \frac{1}{2} \min_{a_t} \max_{y_t} \Bigg(\|a_t - y_t\|^2 + \alpha_t \|s_{t-1} + y_t\|^2 \\
&\qquad - \sigma_{t-1}^2 - \|y_t\|^2 + r^2 \sum_{n=t+1}^{T} \alpha_n \Bigg)
\end{aligned}
$$

Online prediction with quadratic loss on the ball

Proof idea

$$
\begin{aligned}
V(y_1, \ldots, y_{t-1}) &= \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right) \\
&= \frac{1}{2} \min_{a_t} \max_{y_t} \Bigg(\|a_t - y_t\|^2 + \alpha_t \|s_{t-1} + y_t\|^2 \\
&\qquad - \sigma_{t-1}^2 - \|y_t\|^2 + r^2 \sum_{n=t+1}^{T} \alpha_n \Bigg) \\
&= \frac{1}{2} \min_{a_t} \max_{y_t} \Bigg(\|a_t\|^2 + 2 \left(\alpha_t s_{t-1} - a_t \right)^\top y_t + \alpha_t \|y_t\|^2 \\
&\qquad + \alpha_t \|s_{t-1}\|^2 - \sigma_{t-1}^2 + r^2 \sum_{n=t+1}^{T} \alpha_n \Bigg)
\end{aligned}
$$

Online prediction with quadratic loss on the ball

Proof idea

$$
\begin{aligned}
V(y_1, \ldots, y_{t-1}) &= \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right) \\
&= \frac{1}{2} \min_{a_t} \max_{y_t} \Bigg(\|a_t - y_t\|^2 + \alpha_t \|s_{t-1} + y_t\|^2 \\
&\qquad - \sigma_{t-1}^2 - \|y_t\|^2 + r^2 \sum_{n=t+1}^{T} \alpha_n \Bigg) \\
&= \frac{1}{2} \min_{a_t} \max_{y_t} \Bigg(\|a_t\|^2 + 2 \left(\alpha_t s_{t-1} - a_t \right)^\top y_t + \alpha_t \|y_t\|^2 \\
&\qquad + \alpha_t \|s_{t-1}\|^2 - \sigma_{t-1}^2 + r^2 \sum_{n=t+1}^{T} \alpha_n \Bigg)
\end{aligned}
$$

Optimization of y_t: maximize a convex function over the ball.

Online prediction with quadratic loss on the ball

Proof idea

$$
\begin{aligned}
V(y_1,\ldots,y_{t-1}) &= \min_{a_t}\max_{y_t}\left(\ell(a_t,y_t) + V(y_1,\ldots,y_t)\right) \\
&= \frac{1}{2}\min_{a_t}\max_{y_t}\Bigg(\|a_t - y_t\|^2 + \alpha_t\|s_{t-1}+y_t\|^2 \\
&\qquad - \sigma_{t-1}^2 - \|y_t\|^2 + r^2\sum_{n=t+1}^{T}\alpha_n\Bigg) \\
&= \frac{1}{2}\min_{a_t}\max_{y_t}\Bigg(\|a_t\|^2 + 2\left(\alpha_t s_{t-1} - a_t\right)^\top y_t + \alpha_t\|y_t\|^2 \\
&\qquad + \alpha_t\|s_{t-1}\|^2 - \sigma_{t-1}^2 + r^2\sum_{n=t+1}^{T}\alpha_n\Bigg)
\end{aligned}
$$

Optimization of y_t: maximize a convex function over the ball.
But the solution is easy: choose y_t on the sphere, aligned with $\alpha_t s_{t-1} - a_t$.

Online prediction with quadratic loss on the ball

Proof idea

$$V(y_1, \ldots, y_{t-1}) = \frac{1}{2} \min_{a_t} \max_{y_t} \Bigg(\|a_t\|^2 + 2\left(\alpha_t s_{t-1} - a_t\right)^\top y_t + \alpha_t \|y_t\|^2 + \alpha_t \|s_{t-1}\|^2 - \sigma_{t-1}^2 + r^2 \sum_{n=t+1}^{T} \alpha_n \Bigg)$$

Online prediction with quadratic loss on the ball

Proof idea

$$V(y_1, \ldots, y_{t-1}) = \frac{1}{2} \min_{a_t} \max_{y_t} \left(\|a_t\|^2 + 2\left(\alpha_t s_{t-1} - a_t\right)^\top y_t + \alpha_t \|y_t\|^2 + \alpha_t \|s_{t-1}\|^2 - \sigma_{t-1}^2 + r^2 \sum_{n=t+1}^{T} \alpha_n \right)$$

$$= \frac{1}{2} \min_{a_t} \left(\|a_t\|^2 + 2r \left\|\alpha_t s_{t-1} - a_t\right\| + \alpha_t \|s_{t-1}\|^2 - \sigma_{t-1}^2 + r^2 \sum_{n=t}^{T} \alpha_n \right)$$

Online prediction with quadratic loss on the ball

Proof idea

$$V(y_1, \ldots, y_{t-1}) = \frac{1}{2} \min_{a_t} \max_{y_t} \left(\|a_t\|^2 + 2\left(\alpha_t s_{t-1} - a_t\right)^\top y_t + \alpha_t \|y_t\|^2 \right.$$
$$\left. + \alpha_t \|s_{t-1}\|^2 - \sigma_{t-1}^2 + r^2 \sum_{n=t+1}^{T} \alpha_n \right)$$
$$= \frac{1}{2} \min_{a_t} \left(\|a_t\|^2 + 2r \left\|\alpha_t s_{t-1} - a_t\right\| \right.$$
$$\left. + \alpha_t \|s_{t-1}\|^2 - \sigma_{t-1}^2 + r^2 \sum_{n=t}^{T} \alpha_n \right)$$

$(a_t^* := \alpha_t s_{t-1})$

Online prediction with quadratic loss on the ball

Proof idea

$$V(y_1,\ldots,y_{t-1}) = \frac{1}{2}\min_{a_t}\max_{y_t}\Bigg(\|a_t\|^2 + 2\left(\alpha_t s_{t-1} - a_t\right)^\top y_t + \alpha_t\|y_t\|^2 + \alpha_t\|s_{t-1}\|^2 - \sigma_{t-1}^2 + r^2\sum_{n=t+1}^{T}\alpha_n\Bigg)$$

$$= \frac{1}{2}\min_{a_t}\Bigg(\|a_t\|^2 + 2r\left\|\alpha_t s_{t-1} - a_t\right\| + \alpha_t\|s_{t-1}\|^2 - \sigma_{t-1}^2 + r^2\sum_{n=t}^{T}\alpha_n\Bigg)$$

$(a_t^* := \alpha_t s_{t-1})$ $$= \frac{1}{2}\Bigg(\alpha_t^2\|s_{t-1}\|^2 + \alpha_t\|s_{t-1}\|^2 - \sigma_{t-1}^2 + r^2\sum_{n=t}^{T}\alpha_n\Bigg).$$

Online prediction with quadratic loss on the ball

Proof idea

$$V(y_1, \ldots, y_{t-1}) = \frac{1}{2} \min_{a_t} \max_{y_t} \Bigg(\|a_t\|^2 + 2\left(\alpha_t s_{t-1} - a_t\right)^\top y_t + \alpha_t \|y_t\|^2$$

$$+ \alpha_t \|s_{t-1}\|^2 - \sigma_{t-1}^2 + r^2 \sum_{n=t+1}^{T} \alpha_n \Bigg)$$

$$= \frac{1}{2} \min_{a_t} \Bigg(\|a_t\|^2 + 2r \left\|\alpha_t s_{t-1} - a_t\right\|$$

$$+ \alpha_t \|s_{t-1}\|^2 - \sigma_{t-1}^2 + r^2 \sum_{n=t}^{T} \alpha_n \Bigg)$$

$(a_t^* := \alpha_t s_{t-1})$ $$= \frac{1}{2} \left(\alpha_t^2 \|s_{t-1}\|^2 + \alpha_t \|s_{t-1}\|^2 - \sigma_{t-1}^2 + r^2 \sum_{n=t}^{T} \alpha_n \right).$$

Principle of indifference: Any $\|y_t\| = r$ is a best response.

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

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

$$\alpha_T = \frac{1}{T}, \quad \alpha_n = \alpha_{n+1}^2 + \alpha_{n+1} \le \frac{1}{n}.$$

Minimax strategy: affine in state

$$a_n^* = c + \alpha_n \sum_{t=1}^{n-1}(y_t - c).$$

$$a_n^* = (n-1)\alpha_n \bar{y}_{n-1} + (1-(n-1)\alpha_n)c.$$

Maximin distribution: same mean.

Minimax regret for ball

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

Outline

- Computing minimax optimal strategies.
- Part 1: Euclidean loss.
  - $\mathcal{Y}$ = ball
  - $\mathcal{Y}$ = **simplex**
  - Closed, bounded $\mathcal{Y}$
  - Hilbert space
  - $\mathcal{Y}$ = ellipsoid
- Part 2: Linear regression.

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

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

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

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

$$\alpha_T = \frac{1}{T}, \qquad \alpha_n = \alpha_{n+1}^2 + \alpha_{n+1}$$

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

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

$$a_n^* = c + \alpha_n \sum_{t=1}^{n-1}(y_t - c).$$

$$\alpha_T = \frac{1}{T}, \qquad \alpha_n = \alpha_{n+1}^2 + \alpha_{n+1}$$

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

$$a_n^* = c + \alpha_n \sum_{t=1}^{n-1} (y_t - c).$$

$$a_n^* = (n-1)\alpha_n \bar{y}_{n-1} + (1-(n-1)\alpha_n)c.$$

$$\alpha_T = \frac{1}{T}, \qquad \alpha_n = \alpha_{n+1}^2 + \alpha_{n+1}$$

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

$$a_n^* = c + \alpha_n \sum_{t=1}^{n-1} (y_t - c).$$

$$a_n^* = (n-1)\alpha_n \bar{y}_{n-1} + (1-(n-1)\alpha_n)c.$$

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

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

$$a_n^* = c + \alpha_n \sum_{t=1}^{n-1} (y_t - c).$$

$$a_n^* = (n-1)\alpha_n \bar{y}_{n-1} + (1-(n-1)\alpha_n)c.$$

Maximin distribution: same mean.

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

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:

$$V(y_1, \ldots, y_T) := -\min_a \frac{1}{2} \sum_{t=1}^{T} \|a - y_t\|^2 = -\frac{1}{2} \sum_{t=1}^{T} \|\bar{y} - y_t\|^2$$

$$= -\frac{1}{2} \sum_{t=1}^{T} \| \frac{s_T}{T} + c - y_t \|^2$$

i.e., $\quad V(s_T, \sigma_T^2, T) = \frac{1}{2} \left(\alpha_T \|s_T\|^2 - \sigma_T^2 \right).$

Online prediction with quadratic loss on the simplex

Proof idea

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

Online prediction with quadratic loss on the simplex

Proof idea

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

Online prediction with quadratic loss on the simplex

Proof idea

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

Online prediction with quadratic loss on the simplex

Proof idea

$$
\begin{aligned}
V(y_1,\ldots,y_{t-1}) &:= \min_{a_t}\max_{y_t}\left(\ell(a_t,y_t)+V(y_1,\ldots,y_t)\right)\\
&= \min_{a_t}\max_{p_t}\mathbb{E}_{y_t\sim p_t}\left(\ell(a_t,y_t)+V(y_1,\ldots,y_t)\right)\\
&= \max_{p_t}\min_{a_t}\mathbb{E}_{y_t\sim p_t}\left(\ell(a_t,y_t)+V(y_1,\ldots,y_t)\right)\\
&= \frac{1}{2}\max_{p_t}\min_{a_t}\mathbb{E}_{y_t\sim p_t}\Bigg(\|a_t-y_t\|^2+\alpha_t\|s_{t-1}+y_t-c\|^2\\
&\qquad -\sigma_{t-1}^2-\|y_t-c\|^2+r^2\sum_{n=t+1}^{T}\alpha_n\Bigg)
\end{aligned}
$$

Online prediction with quadratic loss on the simplex

Proof idea

$$
\begin{aligned}
V(y_1,\ldots,y_{t-1}) &:= \min_{a_t}\max_{y_t}\left(\ell(a_t,y_t)+V(y_1,\ldots,y_t)\right)\\
&= \min_{a_t}\max_{p_t}\mathbb{E}_{y_t\sim p_t}\left(\ell(a_t,y_t)+V(y_1,\ldots,y_t)\right)\\
&= \max_{p_t}\min_{a_t}\mathbb{E}_{y_t\sim p_t}\left(\ell(a_t,y_t)+V(y_1,\ldots,y_t)\right)\\
&= \frac{1}{2}\max_{p_t}\min_{a_t}\mathbb{E}_{y_t\sim p_t}\Bigg(\|a_t-y_t\|^2+\alpha_t\|s_{t-1}+y_t-c\|^2\\
&\qquad -\sigma_{t-1}^2-\|y_t-c\|^2+r^2\sum_{n=t+1}^{T}\alpha_n\Bigg)\\
&= \frac{1}{2}\max_{p_t}\sum_{i=1}^{d+1}p_t(i)\left(\|Zp_t-z_i\|^2+C_i\right),
\end{aligned}
$$

where $Z=[z_1\cdots z_{d+1}]$ contains the simplex vertices.

Online prediction with quadratic loss on the simplex

Proof idea

$$V(y_1, \ldots, y_{t-1}) = \frac{1}{2} \max_{p_t} \sum_{i=1}^{d+1} p_t(i) \left(\|Zp_t - z_i\|^2 + C_i \right)$$

Online prediction with quadratic loss on the simplex

Proof idea

$$
\begin{aligned}
V(y_1, \ldots, y_{t-1}) &= \frac{1}{2} \max_{p_t} \sum_{i=1}^{d+1} p_t(i) \left(\|Zp_t - z_i\|^2 + C_i \right) \\
&= \frac{1}{2} \max_{p_t} \left(-p_t^\top Z^\top Z p_t + \text{linear}(p_t) + \text{constant} \right).
\end{aligned}
$$

Online prediction with quadratic loss on the simplex

Proof idea

$$V(y_1, \ldots, y_{t-1}) = \frac{1}{2} \max_{p_t} \sum_{i=1}^{d+1} p_t(i) \left(\|Zp_t - z_i\|^2 + C_i \right)$$

$$= \frac{1}{2} \max_{p_t} \left(-p_t^\top Z^\top Z p_t + \text{linear}(p_t) + \text{constant} \right).$$

It's clear that, 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

$$V(y_1, \ldots, y_{t-1}) = \frac{1}{2} \max_{p_t} \sum_{i=1}^{d+1} p_t(i) \left(\|Zp_t - z_i\|^2 + C_i \right)$$
$$= \frac{1}{2} \max_{p_t} \left(-p_t^\top Z^\top Z p_t + \text{linear}(p_t) + \text{constant} \right).$$

It's clear that, 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.
It turns out that when the simplex points z_i are on the surface of the smallest ball, the maximizer is a probability distribution.

Proof idea

- Solving this quadratic minimization gives:

Online prediction with quadratic loss on the simplex

Proof idea

- Solving this quadratic minimization gives:
  - the value function,

Online prediction with quadratic loss on the simplex

Proof idea

- Solving this quadratic minimization gives:
  - the value function,
  - hence the recurrence relation determining the α_t sequence, and

Online prediction with quadratic loss on the simplex

Proof idea

- Solving this quadratic minimization gives:
  - the value function,
  - hence the recurrence relation determining the α_t sequence, and
  - the optimal a_t^*.

Online prediction with quadratic loss on the simplex

Proof idea

- Solving this quadratic minimization gives:
  - the value function,
  - hence the recurrence relation determining the α_t sequence, and
  - the optimal a_t^*.
- It's clear from the proof that the maximin distribution is concentrated on the vertices of the simplex, and that a_t^* is its expectation.

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

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

$$a_n^* = c + \alpha_n \sum_{t=1}^{n-1}(y_t - c).$$

$$a_n^* = (n-1)\alpha_n \bar{y}_{n-1} + (1-(n-1)\alpha_n)c.$$

Maximin distribution: same mean.

$$\alpha_T = \frac{1}{T}, \qquad \alpha_n = \alpha_{n+1}^2 + \alpha_{n+1} \leq \frac{1}{n}.$$

Outline

- Computing minimax optimal strategies.
- Part 1: Euclidean loss.
  - $\mathcal{Y} =$ ball
  - $\mathcal{Y} =$ simplex
  - **Closed, bounded** $\mathcal{Y}$
  - Hilbert space
  - $\mathcal{Y} =$ ellipsoid
- Part 2: Linear regression.

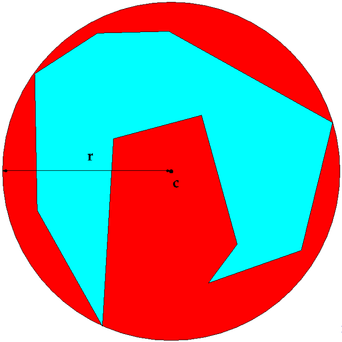
Online prediction with quadratic loss

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

Online prediction with quadratic loss

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

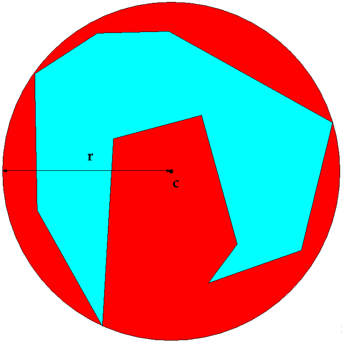
Recall: the smallest ball containing $\mathcal{Y}$ is $B_{\mathcal{Y}} = \{x \in \mathbb{R}^d : \|x - c\| \leq r\}$.



Online prediction with quadratic loss

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

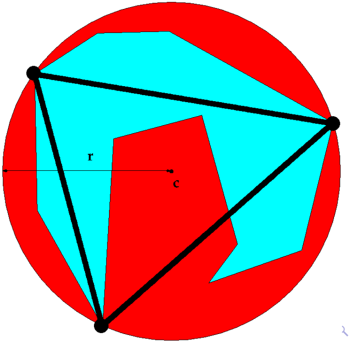
Recall: the smallest ball containing $\mathcal{Y}$ is $B_{\mathcal{Y}} = \{x \in \mathbb{R}^d : \|x - c\| \leq r\}$.
A Lagrange dual argument shows that the optimal center is in the convex hull of a set of *contact points* of $\mathcal{Y}$ at radius r.



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

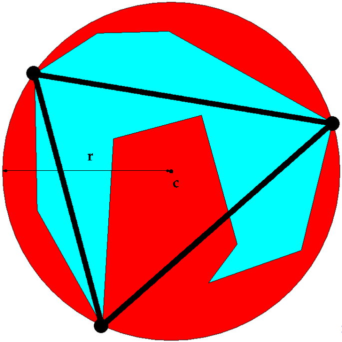
The general case: closed, bounded $\mathcal{Y} \subset \mathbb{R}^d$

Recall: the smallest ball containing $\mathcal{Y}$ is $B_{\mathcal{Y}} = \{x \in \mathbb{R}^d : \|x - c\| \le r\}$.
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 $|S| \le d + 1$.

From below

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

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



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

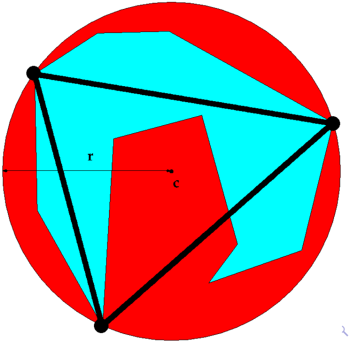
$\mathcal{Y} \supseteq S$, so

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

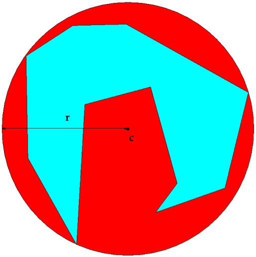
From above

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

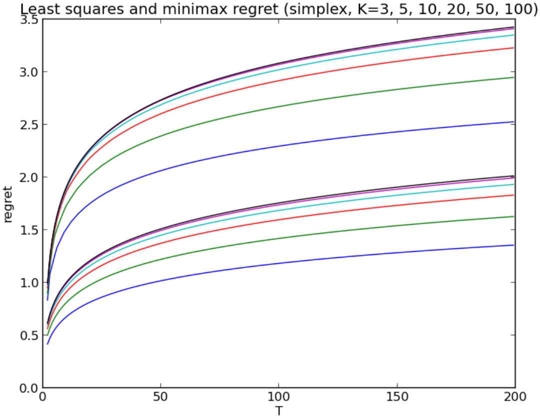
Minimax regret

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

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.
  - $\mathcal{Y} =$ ball
  - $\mathcal{Y} =$ simplex
  - Closed, bounded $\mathcal{Y}$
  - **Hilbert space**
  - $\mathcal{Y} =$ ellipsoid
- Part 2: Linear regression.

Online prediction in Hilbert space

Online prediction in Hilbert space

Loss

$$\ell(\hat{y}, y) = \frac{1}{2} \|\hat{y} - y\|^2 .$$

Online prediction in Hilbert space

Constraints

Strategy chooses $\hat{y}_n \in \mathcal{H}$, a Hilbert space (separable, complete, inner product space).

$$\|\hat{y} - y\|^2 = \langle \hat{y} - y, \hat{y} - y \rangle .$$

Loss

$$\ell(\hat{y}, y) = \frac{1}{2} \|\hat{y} - y\|^2 .$$

Online prediction in Hilbert space

Constraints

Strategy chooses $\hat{y}_n \in \mathcal{H}$, a Hilbert space (separable, complete, inner product space).

$$\|\hat{y} - y\|^2 = \langle \hat{y} - y, \hat{y} - y \rangle .$$

Adversary chooses $y_n \in \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathcal{H}$ is a closed, bounded, convex set.

Loss

$$\ell(\hat{y}, y) = \frac{1}{2} \|\hat{y} - y\|^2 .$$

Online prediction in Hilbert space

Constraints

Strategy chooses $\hat{y}_n \in \mathcal{H}$, a Hilbert space (separable, complete, inner product space).

$$\|\hat{y} - y\|^2 = \langle \hat{y} - y, \hat{y} - y \rangle .$$

Adversary chooses $y_n \in \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathcal{H}$ is a closed, bounded, convex set.

Loss

$$\ell(\hat{y}, y) = \frac{1}{2} \|\hat{y} - y\|^2 .$$

$$\text{Regret} = \sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{a \in \mathcal{H}} \sum_{t=1}^{n} \ell(a, y_t).$$

Online prediction in Hilbert space

Fact

The smallest enclosing ball of a closed, bounded, convex $\mathcal{Y}$ is well-defined, with a unique center and radius.

Online prediction in Hilbert space

Fact

The smallest enclosing ball of a closed, bounded, convex $\mathcal{Y}$ is well-defined, with a unique center and radius.

Recall

If $\mathcal{Y}$ lies in a finite-dimensional subset of $\mathcal{H}$, with smallest enclosing ball $B(c, r)$, the minimax strategy is

$$a_n^* = c + \alpha_n \sum_{t=1}^{n-1} (y_t - c).$$

Online prediction in Hilbert space

Theorem

For any $d \in \mathcal{H}$, the strategy

$$a_n = d + \alpha_n \sum_{t=1}^{n-1} (y_t - c)$$

Online prediction in Hilbert space

Theorem

For any $d \in \mathcal{H}$, the strategy

$$a_n = d + \alpha_n \sum_{t=1}^{n-1} (y_t - c)$$

has regret

$$R_T = \frac{1}{2} \sum_{t=1}^{T} \alpha_t \|y_t - d\|^2$$

Online prediction in Hilbert space

Theorem

For any $d \in \mathcal{H}$, the strategy

$$a_n = d + \alpha_n \sum_{t=1}^{n-1} (y_t - c)$$

has regret

$$\begin{aligned} R_T &= \frac{1}{2} \sum_{t=1}^{T} \alpha_t \|y_t - d\|^2 \\ &\leq \frac{1}{2} \sup_{z \in \mathcal{Y}} \|z - d\|^2 \sum_{t=1}^{T} \alpha_t . \end{aligned}$$

Online prediction in Hilbert space

Theorem

For any $d \in \mathcal{H}$, the strategy

$$a_n = d + \alpha_n \sum_{t=1}^{n-1} (y_t - c)$$

has regret

$$\begin{aligned} R_T &= \frac{1}{2} \sum_{t=1}^{T} \alpha_t \|y_t - d\|^2 \\ &\leq \frac{1}{2} \sup_{z \in \mathcal{Y}} \|z - d\|^2 \sum_{t=1}^{T} \alpha_t . \end{aligned}$$

Proof: a straightforward calculation.

Online prediction in Hilbert space

Corollary

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

(By setting $d = c$.)

Online prediction in Hilbert space

Theorem

For a closed, bounded, convex $\mathcal{Y}$

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

Online prediction in Hilbert space

Lower bound proof idea

We construct a sequence of finite sets $C_1, C_2, \ldots \subseteq \mathcal{Y}$ so that

$$\frac{r(C_i)}{r(\mathcal{Y})} \geq 1 - \sqrt{\frac{2}{i}},$$

where $r(C_i)$ is the radius of the smallest ball containing C_i.

Online prediction in Hilbert space

Lower bound proof idea

We construct a sequence of finite sets $C_1, C_2, \ldots \subseteq \mathcal{Y}$ so that

$$\frac{r(C_i)}{r(\mathcal{Y})} \geq 1 - \sqrt{\frac{2}{i}},$$

where $r(C_i)$ is the radius of the smallest ball containing C_i.

Online prediction in Hilbert space

Lower bound proof idea

We construct a sequence of finite sets $C_1, C_2, \ldots \subseteq \mathcal{Y}$ so that

$$\frac{r(C_i)}{r(\mathcal{Y})} \geq 1 - \sqrt{\frac{2}{i}},$$

where $r(C_i)$ is the radius of the smallest ball containing C_i.
Since $V_T(C_i) \leq V_T(\mathcal{Y})$, this gives the result.
To construct the C_i:

Online prediction in Hilbert space

Lower bound proof idea

We construct a sequence of finite sets $C_1, C_2, \ldots \subseteq \mathcal{Y}$ so that

$$\frac{r(C_i)}{r(\mathcal{Y})} \geq 1 - \sqrt{\frac{2}{i}},$$

where $r(C_i)$ is the radius of the smallest ball containing C_i.
Since $V_T(C_i) \leq V_T(\mathcal{Y})$, this gives the result.
To construct the C_i:

1. Start with $C_1 = \{y_1\}$.

Online prediction in Hilbert space

Lower bound proof idea

We construct a sequence of finite sets $C_1, C_2, \ldots \subseteq \mathcal{Y}$ so that

$$\frac{r(C_i)}{r(\mathcal{Y})} \geq 1 - \sqrt{\frac{2}{i}},$$

where $r(C_i)$ is the radius of the smallest ball containing C_i.
Since $V_T(C_i) \leq V_T(\mathcal{Y})$, this gives the result.
To construct the C_i:

1. Start with $C_1 = \{y_1\}$.
2. Set $C_{i+1} = C_i \cup \{y_{i+1}\}$, where $\|c - y_{i+1}\| \geq r(\mathcal{Y})$, for c the center of the smallest enclosing ball for $\mathcal{Y}$.

Online prediction in Hilbert space

Lower bound proof idea

$$r^2(C_i) = \min_c \max_{y \in C_i} \|y - c\|^2$$

Online prediction in Hilbert space

Lower bound proof idea

$$r^2(C_i) = \min_c \max_{y \in C_i} \|y - c\|^2$$
$$= \max_p \min_c \sum_{z \in C_i} p_z \|z - c\|^2.$$

Lower bound proof idea

$$
\begin{aligned}
r^2(C_i) &= \min_c \max_{y \in C_i} \|y - c\|^2 \\
&= \max_p \min_c \sum_{z \in C_i} p_z \|z - c\|^2.
\end{aligned}
$$

(We can apply the min-max theorem because C_i is finite.)

Online prediction in Hilbert space

Lower bound proof idea

$$
\begin{aligned}
r^2(C_i) &= \min_c \max_{y \in C_i} \|y - c\|^2 \\
&= \max_p \min_c \sum_{z \in C_i} p_z \|z - c\|^2.
\end{aligned}
$$

(We can apply the min-max theorem because C_i is finite.)
Now, consider a distribution q on $C_{i+1} = C_i \cup \{y_{i+1}\}$, with

$$
q(z) = \begin{cases} (1-\lambda)p(z) & \text{for } z \in C_i, \\ \lambda & \text{for } z = y_{i+1}. \end{cases}
$$

Online prediction in Hilbert space

Lower bound proof idea

$$
\begin{aligned}
r^2(C_i) &= \min_c \max_{y \in C_i} \|y - c\|^2 \\
&= \max_p \min_c \sum_{z \in C_i} p_z \|z - c\|^2.
\end{aligned}
$$

(We can apply the min-max theorem because C_i is finite.)
Now, consider a distribution q on $C_{i+1} = C_i \cup \{y_{i+1}\}$, with

$$
q(z) = \begin{cases} (1-\lambda)p(z) & \text{for } z \in C_i, \\ \lambda & \text{for } z = y_{i+1}. \end{cases}
$$

Evaluating $r(C_{i+1})$ and optimizing over λ gives the result.

Online prediction in Hilbert space

Theorem

For a closed, bounded, convex $\mathcal{Y}$

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

Online prediction in Hilbert space

Theorem

For a closed, bounded, convex $\mathcal{Y}$

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

and this is achieved by the minimax optimal strategy

$$a_n^* = c + \alpha_n \sum_{t=1}^{n-1} (y_t - c)$$

Outline

- Computing minimax optimal strategies.
- Part 1: Euclidean loss.
  - $\mathcal{Y} =$ ball
  - $\mathcal{Y} =$ simplex
  - Closed, bounded $\mathcal{Y}$
  - Hilbert space
  - $\mathcal{Y} =$ **ellipsoid**
- Part 2: Linear regression.

Online prediction with quadratic loss on an ellipsoid

Ellipsoid:

$$\mathcal{Y} = \{y : (y - c)^\top W^{-1}(y - c) \leq 1\}$$

Here, $W \succeq 0$.

Online prediction with quadratic loss on an ellipsoid

Ellipsoid:

$$\mathcal{Y} = \{y : (y - c)^\top W^{-1}(y - c) \leq 1\}$$

Here, $W \succeq 0$. Without loss of generality, W is symmetric.

Online prediction with quadratic loss on an ellipsoid

Ellipsoid:

$$\mathcal{Y} = \{y : (y-c)^\top W^{-1}(y-c) \le 1\}$$

Here, $W \succeq 0$. Without loss of generality, W is symmetric.

Write $W = \sum_{i=1}^{d} \nu_i v_i v_i^\top$, with $\nu_1 \ge \nu_2 \ge \cdots \ge \nu_d \ge 0$.

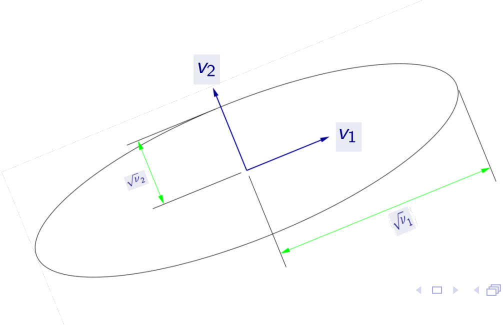
Online prediction with quadratic loss on an ellipsoid

Ellipsoid:

$$\mathcal{Y} = \{y : (y-c)^\top W^{-1}(y-c) \le 1\}$$

Here, $W \succeq 0$. Without loss of generality, W is symmetric.

Write $W = \sum_{i=1}^{d} \nu_i v_i v_i^\top$, with $\nu_1 \ge \nu_2 \ge \cdots \ge \nu_d \ge 0$.



Online prediction with quadratic loss on an ellipsoid

Ellipsoid: $\mathcal{Y} = \{y : (y-c)^\top W^{-1}(y-c) \leq 1\}$ $(W \succeq 0)$

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

Online prediction with quadratic loss on an ellipsoid

Ellipsoid: $\mathcal{Y} = \{y : (y-c)^\top W^{-1}(y-c) \le 1\}$ $(W \succeq 0)$

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.

$$V(s_n, \sigma_n^2) = \frac{1}{2}\left(s_n^\top A_n s_n - \sigma_n^2 + \nu_1 \sum_{t=n+1}^{T} \alpha_n\right).$$

Online prediction with quadratic loss on an ellipsoid

Ellipsoid: $\mathcal{Y} = \{y : (y-c)^\top W^{-1}(y-c) \le 1\}$ $\quad (W \succeq 0)$

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.

$$V(s_n, \sigma_n^2) = \frac{1}{2}\left(s_n^\top A_n s_n - \sigma_n^2 + \nu_1 \sum_{t=n+1}^{T} \alpha_n\right).$$

$$\nu_1 = \lambda_{\max}(W),$$

Online prediction with quadratic loss on an ellipsoid

Ellipsoid: $\mathcal{Y} = \{y : (y-c)^\top W^{-1}(y-c) \leq 1\}$ $(W \succeq 0)$

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.

$$V(s_n, \sigma_n^2) = \frac{1}{2}\left(s_n^\top A_n s_n - \sigma_n^2 + \nu_1 \sum_{t=n+1}^{T} \alpha_n\right).$$

$$\nu_1 = \lambda_{\max}(W),$$

$$A_T = \frac{1}{T}I, \qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1}$$

Online prediction with quadratic loss on an ellipsoid

Ellipsoid: $\mathcal{Y} = \{y : (y-c)^\top W^{-1}(y-c) \le 1\}$ $\qquad (W \succeq 0)$

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.

$$V(s_n, \sigma_n^2) = \frac{1}{2}\left(s_n^\top A_n s_n - \sigma_n^2 + \nu_1 \sum_{t=n+1}^{T} \alpha_n\right).$$

$$\alpha_n = \lambda_{\max}(A_n), \qquad \nu_1 = \lambda_{\max}(W),$$

$$A_T = \frac{1}{T} I, \qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1}$$

Online prediction with quadratic loss on an ellipsoid

Ellipsoid: $\mathcal{Y} = \{y : (y-c)^\top W^{-1}(y-c) \le 1\}$ $\qquad (W \succeq 0)$

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. Minimax strategy: affine in state

$$V(s_n, \sigma_n^2) = \frac{1}{2}\left(s_n^\top A_n s_n - \sigma_n^2 + \nu_1 \sum_{t=n+1}^{T} \alpha_n\right).$$

$$a_n^* - c = \left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} \sum_{t=1}^{n-1}(y_t - c).$$

$\alpha_n = \lambda_{\max}(A_n)$, $\qquad \nu_1 = \lambda_{\max}(W)$,

$A_T = \frac{1}{T} I$, $\qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1}$

Online prediction with quadratic loss on an ellipsoid

Ellipsoid: $\mathcal{Y} = \{y : (y-c)^\top W^{-1}(y-c) \le 1\}$ $\qquad (W \succeq 0)$

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. Minimax strategy: affine in state

$$V(s_n, \sigma_n^2) = \frac{1}{2}\left(s_n^\top A_n s_n - \sigma_n^2 + \nu_1 \sum_{t=n+1}^{T} \alpha_n\right).$$

$$a_n^* - c = \left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} \sum_{t=1}^{n-1} (y_t - c).$$

$\alpha_n = \lambda_{\max}(A_n)$, $\qquad \nu_1 = \lambda_{\max}(W)$,

$A_T = \frac{1}{T} I$, $\qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1}$

$$V(\mathcal{Y}) = \frac{\nu_1}{2}\sum_{t=1}^{T} \alpha_t.$$

Online prediction with quadratic loss on an ellipsoid

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 an ellipsoid

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, as before:

$$V(y_1, \dots, y_T) = \frac{1}{2}\left(s_T^\top A_T s_T - \sigma_T^2\right).$$

Online prediction with quadratic loss on an ellipsoid

Proof idea

$$
\begin{aligned}
V(y_1, \ldots, y_{t-1}) &= \min_{a_t} \max_{y_t} \left(\ell(a_t, y_t) + V(y_1, \ldots, y_t) \right) \\
&= \frac{1}{2} \min_{a_t} \max_{y_t} \Bigg(\|a_t - y_t\|^2 + (s_{t-1} + y_t)^\top A_n (s_{t-1} + y_t) \\
&\qquad - \sigma_{t-1}^2 - \|y_t\|^2 + \nu_1 \sum_{t=n+1}^{T} \alpha_n \Bigg) .
\end{aligned}
$$

Online prediction with quadratic loss on an ellipsoid

Proof idea

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

$$= \frac{1}{2}\min_{a_t}\max_{y_t}\Bigg(\|a_t - y_t\|^2 + (s_{t-1}+y_t)^\top A_n (s_{t-1}+y_t)$$

$$-\sigma_{t-1}^2 - \|y_t\|^2 + \nu_1 \sum_{t=n+1}^{T}\alpha_n\Bigg).$$

- At each step, the inner maximum is of a (convex) quadratic criterion with a single quadratic constraint.

Online prediction with quadratic loss on an ellipsoid

Proof idea

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

$$= \frac{1}{2} \min_{a_t} \max_{y_t} \Bigg(\|a_t - y_t\|^2 + (s_{t-1} + y_t)^\top A_n (s_{t-1} + y_t)$$

$$- \sigma_{t-1}^2 - \|y_t\|^2 + \nu_1 \sum_{t=n+1}^{T} \alpha_n \Bigg).$$

- 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 an ellipsoid

Proof idea

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

$$= \frac{1}{2} \min_{a_t} \max_{y_t} \Bigg(\|a_t - y_t\|^2 + (s_{t-1} + y_t)^\top A_n (s_{t-1} + y_t)$$

$$- \sigma_{t-1}^2 - \|y_t\|^2 + \nu_1 \sum_{t=n+1}^{T} \alpha_n \Bigg).$$

- 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 on an ellipsoid

Minimax strategy:

$$a_n^* - c = \left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} \sum_{t=1}^{n-1} (y_t - c).$$

$$\nu_1 = \lambda_{\max}(W), \quad \alpha_{n+1} = \lambda_{\max}(A_{n+1}),$$

$$A_T = \frac{1}{T} I, \qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1}.$$

Online prediction with quadratic loss on an ellipsoid

Minimax strategy:

$$a_n^* - c = \left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} \sum_{t=1}^{n-1} (y_t - c).$$

$$\nu_1 = \lambda_{\max}(W), \quad \alpha_{n+1} = \lambda_{\max}(A_{n+1}),$$

$$A_T = \frac{1}{T} I, \qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1}.$$

How does shrinkage behave?

Write $W = \sum_{i=1}^{d} \nu_i v_i v_i^\top$, with $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_d$.

Online prediction with quadratic loss on an ellipsoid

Minimax strategy:

$$a_n^* - c = \left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} \sum_{t=1}^{n-1} (y_t - c).$$

$\nu_1 = \lambda_{\max}(W), \quad \alpha_{n+1} = \lambda_{\max}(A_{n+1}),$

$A_T = \frac{1}{T} I, \qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1}.$

How does shrinkage behave?

Write $W = \sum_{i=1}^{d} \nu_i v_i v_i^\top$, with $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_d$.

It's easy to see that W determines the eigenspace of A_n

Online prediction with quadratic loss on an ellipsoid

Minimax strategy:

$$a_n^* - c = \left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} \sum_{t=1}^{n-1}(y_t - c).$$

$$\nu_1 = \lambda_{\max}(W), \quad \alpha_{n+1} = \lambda_{\max}(A_{n+1}),$$

$$A_T = \frac{1}{T} I, \qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1}.$$

How does shrinkage behave?

Write $W = \sum_{i=1}^{d} \nu_i v_i v_i^\top$, with $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_d$.

It's easy to see that W determines the eigenspace of A_n, so we can write

$$A_n = \sum_{i=1}^{d} \lambda_i^{(n)} v_i v_i^\top.$$

Online prediction with quadratic loss on an ellipsoid

Minimax strategy:

$$a_n^* - c = \left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} \sum_{t=1}^{n-1} (y_t - c).$$

$\nu_1 = \lambda_{\max}(W), \quad \alpha_{n+1} = \lambda_{\max}(A_{n+1}),$

$A_T = \frac{1}{T} I, \qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1}.$

How does shrinkage behave?

Write $W = \sum_{i=1}^{d} \nu_i v_i v_i^\top$, with $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_d$.

It's easy to see that W determines the eigenspace of A_n, so we can write

$$A_n = \sum_{i=1}^{d} \lambda_i^{(n)} v_i v_i^\top.$$

How do the $\lambda_i^{(n)}$ evolve?

Online prediction with quadratic loss on an ellipsoid

$$a_n^* - c = \left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} \sum_{t=1}^{n-1} (y_t - c).$$

$$\alpha_{n+1} = \lambda_{\max}(A_{n+1}), \qquad \nu_1 = \lambda_{\max}(W),$$

$$A_T = \frac{1}{T} I, \qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1},$$

$$W = \sum_{i=1}^{d} \nu_i v_i v_i^\top, \qquad \nu_1 \geq \nu_2 \geq \cdots \geq \nu_d, \qquad A_n = \sum_{i=1}^{d} \lambda_i^{(n)} v_i v_i^\top.$$

Online prediction with quadratic loss on an ellipsoid

$$a_n^* - c = \left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} \sum_{t=1}^{n-1}(y_t - c).$$

$$\alpha_{n+1} = \lambda_{\max}(A_{n+1}), \qquad \nu_1 = \lambda_{\max}(W),$$

$$A_T = \frac{1}{T} I, \qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1},$$

$$W = \sum_{i=1}^{d} \nu_i v_i v_i^\top, \qquad \nu_1 \geq \nu_2 \geq \cdots \geq \nu_d, \qquad A_n = \sum_{i=1}^{d} \lambda_i^{(n)} v_i v_i^\top.$$

How do the eigenvalues evolve?

Online prediction with quadratic loss on an ellipsoid

$$a_n^* - c = \left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} \sum_{t=1}^{n-1} (y_t - c).$$

$$\alpha_{n+1} = \lambda_{\max}(A_{n+1}), \qquad \nu_1 = \lambda_{\max}(W),$$

$$A_T = \frac{1}{T} I, \qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1},$$

$$W = \sum_{i=1}^{d} \nu_i v_i v_i^\top, \qquad \nu_1 \geq \nu_2 \geq \cdots \geq \nu_d, \qquad A_n = \sum_{i=1}^{d} \lambda_i^{(n)} v_i v_i^\top.$$

How do the eigenvalues evolve?

$$\lambda_i^{(T)} = \frac{1}{T},$$

Online prediction with quadratic loss on an ellipsoid

$$a_n^* - c = \left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} \sum_{t=1}^{n-1}(y_t - c).$$

$$\alpha_{n+1} = \lambda_{\max}(A_{n+1}), \qquad \nu_1 = \lambda_{\max}(W),$$

$$A_T = \frac{1}{T} I, \qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1},$$

$$W = \sum_{i=1}^{d} \nu_i v_i v_i^\top, \qquad \nu_1 \geq \nu_2 \geq \cdots \geq \nu_d, \qquad A_n = \sum_{i=1}^{d} \lambda_i^{(n)} v_i v_i^\top.$$

How do the eigenvalues evolve?

$$\lambda_i^{(T)} = \frac{1}{T}, \qquad \lambda_i^{(n)} = \frac{1}{1 + \lambda_1^{(n+1)}\nu_1/\nu_i - \lambda_i^{(n+1)}} \left(\lambda_i^{(n+1)}\right)^2 + \lambda_i^{(n+1)}.$$

Online prediction with quadratic loss on an ellipsoid

$$a_n^* - c = \left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} \sum_{t=1}^{n-1} (y_t - c).$$

$$\alpha_{n+1} = \lambda_{\max}(A_{n+1}), \qquad \nu_1 = \lambda_{\max}(W),$$

$$A_T = \frac{1}{T} I, \qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1},$$

$$W = \sum_{i=1}^{d} \nu_i v_i v_i^\top, \qquad \nu_1 \geq \nu_2 \geq \cdots \geq \nu_d, \qquad A_n = \sum_{i=1}^{d} \lambda_i^{(n)} v_i v_i^\top.$$

How do the eigenvalues evolve?

$$\lambda_i^{(T)} = \frac{1}{T}, \qquad \lambda_i^{(n)} = \frac{1}{1 + \lambda_1^{(n+1)}\nu_1/\nu_i - \lambda_i^{(n+1)}} \left(\lambda_i^{(n+1)}\right)^2 + \lambda_i^{(n+1)}.$$

- $\alpha_n = \lambda_1^{(n)}$

Online prediction with quadratic loss on an ellipsoid

$$a_n^* - c = \left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} \sum_{t=1}^{n-1}(y_t - c).$$

$$\alpha_{n+1} = \lambda_{\max}(A_{n+1}), \qquad \nu_1 = \lambda_{\max}(W),$$

$$A_T = \frac{1}{T} I, \qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1},$$

$$W = \sum_{i=1}^{d} \nu_i v_i v_i^\top, \qquad \nu_1 \geq \nu_2 \geq \cdots \geq \nu_d, \qquad A_n = \sum_{i=1}^{d} \lambda_i^{(n)} v_i v_i^\top.$$

How do the eigenvalues evolve?

$$\lambda_i^{(T)} = \frac{1}{T}, \qquad \lambda_i^{(n)} = \frac{1}{1 + \lambda_1^{(n+1)}\nu_1/\nu_i - \lambda_i^{(n+1)}}\left(\lambda_i^{(n+1)}\right)^2 + \lambda_i^{(n+1)}.$$

- $\alpha_n = \lambda_1^{(n)} \geq \lambda_2^{(n)} \geq \cdots \geq \lambda_d^{(n)}$

Online prediction with quadratic loss on an ellipsoid

$$a_n^* - c = \left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} \sum_{t=1}^{n-1} (y_t - c).$$

$$\alpha_{n+1} = \lambda_{\max}(A_{n+1}), \qquad \nu_1 = \lambda_{\max}(W),$$

$$A_T = \frac{1}{T} I, \qquad A_n = A_{n+1}\left(\alpha_{n+1}\nu_1 W^{-1} + I - A_{n+1}\right)^{-1} A_{n+1} + A_{n+1},$$

$$W = \sum_{i=1}^{d} \nu_i v_i v_i^\top, \qquad \nu_1 \geq \nu_2 \geq \cdots \geq \nu_d, \qquad A_n = \sum_{i=1}^{d} \lambda_i^{(n)} v_i v_i^\top.$$

How do the eigenvalues evolve?

$$\lambda_i^{(T)} = \frac{1}{T}, \qquad \lambda_i^{(n)} = \frac{1}{1 + \lambda_1^{(n+1)}\nu_1/\nu_i - \lambda_i^{(n+1)}} \left(\lambda_i^{(n+1)}\right)^2 + \lambda_i^{(n+1)}.$$

- $\alpha_n = \lambda_1^{(n)} \geq \lambda_2^{(n)} \geq \cdots \geq \lambda_d^{(n)}$; the gap increases with n for smaller ν_i.

Online prediction with quadratic loss on an ellipsoid

Eigenvalues of A_n

$$\lambda_i^{(T)} = \frac{1}{T}, \qquad \lambda_i^{(n)} = \frac{1}{1 + \lambda_1^{(n+1)} \nu_1 / \nu_i - \lambda_i^{(n+1)}} \left(\lambda_i^{(n+1)} \right)^2 + \lambda_i^{(n+1)}.$$

$$\alpha_n = \lambda_1^{(n)} \geq \lambda_2^{(n)} \geq \cdots \geq \lambda_d^{(n)}.$$

Online prediction with quadratic loss on an ellipsoid

Eigenvalues of A_n

$$\lambda_i^{(T)} = \frac{1}{T}, \qquad \lambda_i^{(n)} = \frac{1}{1 + \lambda_1^{(n+1)}\nu_1/\nu_i - \lambda_i^{(n+1)}} \left(\lambda_i^{(n+1)}\right)^2 + \lambda_i^{(n+1)}.$$

$$\alpha_n = \lambda_1^{(n)} \geq \lambda_2^{(n)} \geq \cdots \geq \lambda_d^{(n)}.$$

Optimal strategy

Online prediction with quadratic loss on an ellipsoid

Eigenvalues of A_n

$$\lambda_i^{(T)} = \frac{1}{T}, \qquad \lambda_i^{(n)} = \frac{1}{1 + \lambda_1^{(n+1)} \nu_1/\nu_i - \lambda_i^{(n+1)}} \left(\lambda_i^{(n+1)}\right)^2 + \lambda_i^{(n+1)}.$$

$$\alpha_n = \lambda_1^{(n)} \geq \lambda_2^{(n)} \geq \cdots \geq \lambda_d^{(n)}.$$

Optimal strategy

- $a_n^* - c = \sum_{i=1}^{d} \frac{\lambda_i^{(n)}}{1 + \lambda_1^{(n)} \nu_1/\nu_i - \lambda_i^{(n)}} v_i^\top s_{n-1} v_i.$

Online prediction with quadratic loss on an ellipsoid

Eigenvalues of A_n

$$\lambda_i^{(T)} = \frac{1}{T}, \qquad \lambda_i^{(n)} = \frac{1}{1 + \lambda_1^{(n+1)} \nu_1/\nu_i - \lambda_i^{(n+1)}} \left(\lambda_i^{(n+1)}\right)^2 + \lambda_i^{(n+1)}.$$

$$\alpha_n = \lambda_1^{(n)} \geq \lambda_2^{(n)} \geq \cdots \geq \lambda_d^{(n)}.$$

Optimal strategy

- $a_n^* - c = \sum_{i=1}^{d} \frac{\lambda_i^{(n)}}{1 + \lambda_1^{(n)} \nu_1/\nu_i - \lambda_i^{(n)}} v_i^\top s_{n-1} v_i.$
- The α_n sequence determines the shrinkage in the ellipsoid's major axis, as for the case of the ball.

Online prediction with quadratic loss on an ellipsoid

Eigenvalues of A_n

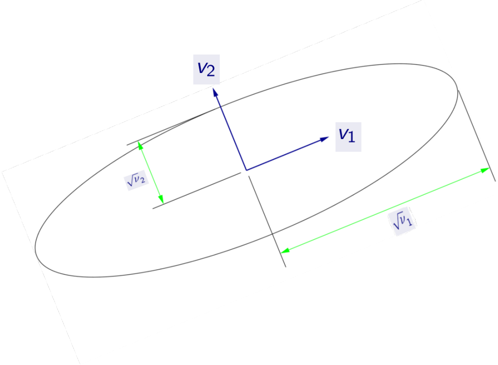
$$\lambda_i^{(T)} = \frac{1}{T}, \qquad \lambda_i^{(n)} = \frac{1}{1 + \lambda_1^{(n+1)}\nu_1/\nu_i - \lambda_i^{(n+1)}} \left(\lambda_i^{(n+1)}\right)^2 + \lambda_i^{(n+1)}.$$

$$\alpha_n = \lambda_1^{(n)} \geq \lambda_2^{(n)} \geq \cdots \geq \lambda_d^{(n)}.$$

Optimal strategy

- $a_n^* - c = \sum_{i=1}^{d} \frac{\lambda_i^{(n)}}{1 + \lambda_1^{(n)}\nu_1/\nu_i - \lambda_i^{(n)}} v_i^\top s_{n-1} v_i.$
- The α_n sequence determines the shrinkage in the ellipsoid's major axis, as for the case of the ball.
- There is *more* shrinkage in the other directions: smaller ν_i implies more shrinkage in the v_i direction.

Online prediction with quadratic loss on an ellipsoid



Optimal strategy

- The α_n sequence determines the shrinkage in the ellipsoid's major axis, as for the case of the ball.
- There is *more* shrinkage in the other directions: smaller ν_i implies more shrinkage in the v_i direction.

Online prediction with quadratic loss

Subgame optimality

Online prediction with quadratic loss

Subgame optimality

- What if the adversary is not optimal?

Online prediction with quadratic loss

Subgame optimality

- What if the adversary is not optimal?
- For $\mathcal{Y} =$ an ellipsoid, or $\mathcal{Y} =$ a simplex that touches its smallest enclosing ball, we have explicit expressions for the value to go and for the optimal prediction for any sequence of adversary choices, including suboptimal ones.

Online prediction with quadratic loss

Subgame optimality

- What if the adversary is not optimal?
- For $\mathcal{Y} =$ an ellipsoid, or $\mathcal{Y} =$ a simplex that touches its smallest enclosing ball, we have explicit expressions for the value to go and for the optimal prediction for any sequence of adversary choices, including suboptimal ones.
- For other $\mathcal{Y}$, an optimal adversary should play only in the intersection of $\mathcal{Y}$ and the surface of the smallest ball.

Online prediction with quadratic loss

Subgame optimality

- What if the adversary is not optimal?
- For $\mathcal{Y}$ = an ellipsoid, or $\mathcal{Y}$ = a simplex that touches its smallest enclosing ball, we have explicit expressions for the value to go and for the optimal prediction for any sequence of adversary choices, including suboptimal ones.
- For other $\mathcal{Y}$, an optimal adversary should play only in the intersection of $\mathcal{Y}$ and the surface of the smallest ball.
- If it does not, the ball strategy (or an enclosing ellipsoid strategy) might be suboptimal.

Online prediction with quadratic loss

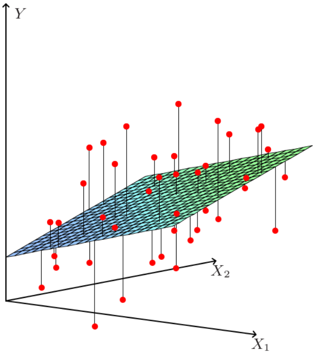
Subgame optimality

- What if the adversary is not optimal?
- For $\mathcal{Y} =$ an ellipsoid, or $\mathcal{Y} =$ a simplex that touches its smallest enclosing ball, we have explicit expressions for the value to go and for the optimal prediction for any sequence of adversary choices, including suboptimal ones.
- For other $\mathcal{Y}$, an optimal adversary should play only in the intersection of $\mathcal{Y}$ and the surface of the smallest ball.
- If it does not, the ball strategy (or an enclosing ellipsoid strategy) might be suboptimal.
- Consider, for example, playing as if $\mathcal{Y}$ is the smallest ball when it is an ellipsoid. If the adversary plays only on the major axis, the optimal strategies are identical. If the adversary is suboptimal, the smallest ball strategy will under-regularize.

Outline

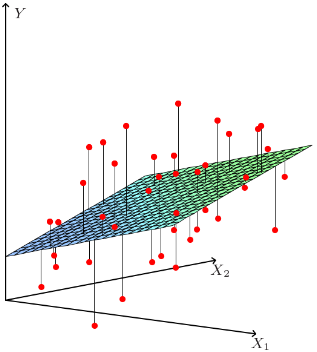
- Computing minimax optimal strategies.
- Part 1: Euclidean loss.
  - $\mathcal{Y}$ = ball
  - $\mathcal{Y}$ = simplex
  - Closed, bounded $\mathcal{Y}$
  - Hilbert space
  - $\mathcal{Y}$ = ellipsoid
- **Part 2: Linear regression.**
  - Fixed design.
  - Minimax strategy is regularized least squares.
  - Box and ellipsoid constraints.
  - Adversarial covariates.

Online fixed design linear regression



(B., Koolen, Malek, Takimoto, Warmuth, 2015)

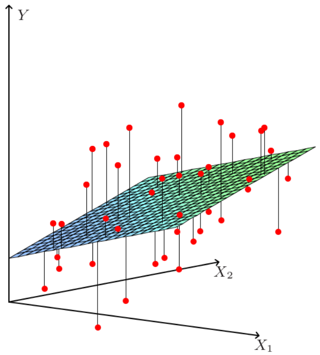
Online fixed design linear regression



Protocol

(B., Koolen, Malek, Takimoto, Warmuth, 2015)

Online fixed design linear regression

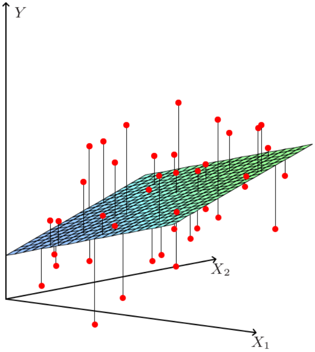


Protocol

Given: T;

(B., Koolen, Malek, Takimoto, Warmuth, 2015)

Online fixed design linear regression

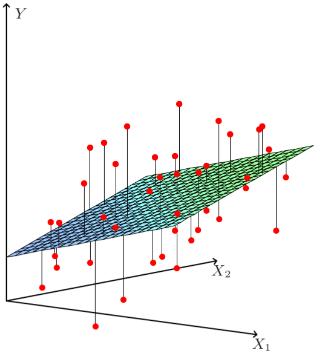


Protocol

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

(B., Koolen, Malek, Takimoto, Warmuth, 2015)

Online fixed design linear regression

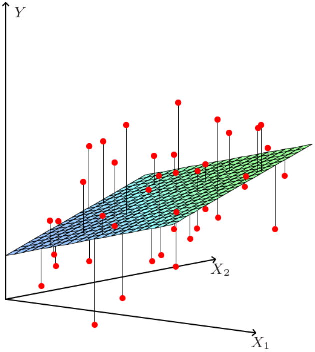


Protocol

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

(B., Koolen, Malek, Takimoto, Warmuth, 2015)

Online fixed design linear regression



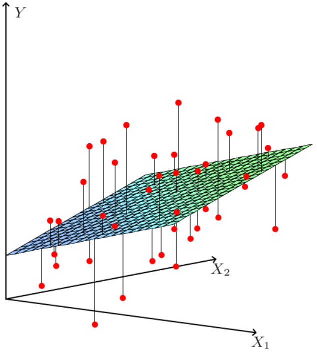
Protocol

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

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

(B., Koolen, Malek, Takimoto, Warmuth, 2015)

Online fixed design linear regression



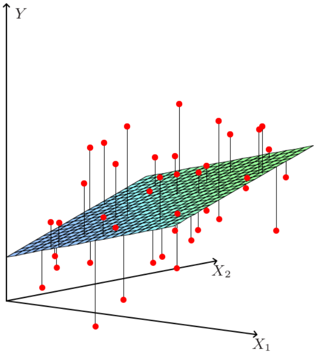
Protocol

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

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

(B., Koolen, Malek, Takimoto, Warmuth, 2015)

Online fixed design linear regression



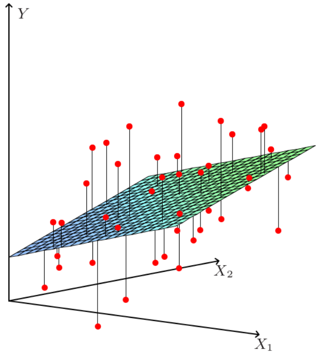
Protocol

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 $y_t \in \mathbb{R}$

(B., Koolen, Malek, Takimoto, Warmuth, 2015)

Online fixed design linear regression



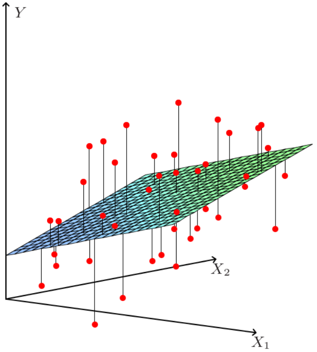
Protocol

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 $y_t \in \mathbb{R}$ $\quad (y_1^T \in \mathcal{Y}^T)$

(B., Koolen, Malek, Takimoto, Warmuth, 2015)

Online fixed design linear regression



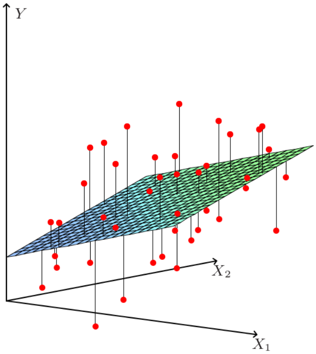
Protocol

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 $(\hat{y}_t - y_t)^2$.

(B., Koolen, Malek, Takimoto, Warmuth, 2015)

Online fixed design linear regression



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 $(\hat{y}_t - y_t)^2$.

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

(B., Koolen, Malek, Takimoto, Warmuth, 2015)

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

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

Minimax* strategy: linear

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

Online fixed design linear regression

Sufficient statistics

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

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

Minimax* strategy: linear

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

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

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

Minimax* strategy: linear

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

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

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:

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

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

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

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

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

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

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

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

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

Minimax strategy: linear

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

Linear regression

Box constraints

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

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

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

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

Minimax strategy: linear

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

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

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

Linear regression

Box constraints

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

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

Minimax strategy: linear

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

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

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

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

Minimax strategy: linear

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

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: Proof idea

Offline optimal:

$$V(s_T, \sigma_T^2, T) = -\min_{\beta \in \mathbb{R}^p} \sum_{t=1}^{T} \left(\beta^\top x_t - y_t \right)^2$$

Linear regression: Proof idea

Offline optimal:

$$V(s_T, \sigma_T^2, T) = -\min_{\beta \in \mathbb{R}^p} \sum_{t=1}^{T} \left(\beta^\top x_t - y_t\right)^2 \qquad \beta_T = P_T s_T,$$

Linear regression: Proof idea

Offline optimal:

$$V(s_T, \sigma_T^2, T) = -\min_{\beta \in \mathbb{R}^p} \sum_{t=1}^{T} \left(\beta^\top x_t - y_t\right)^2 \qquad \beta_T = P_T s_T,$$

$$P_T = \left(\sum_{t=1}^{T} x_t x_t^\top\right)^{-1},$$

Linear regression: Proof idea

Offline optimal:

$$V(s_T, \sigma_T^2, T) = -\min_{\beta \in \mathbb{R}^p} \sum_{t=1}^{T} \left(\beta^\top x_t - y_t\right)^2 \qquad \beta_T = P_T s_T,$$

$$P_T = \left(\sum_{t=1}^{T} x_t x_t^\top\right)^{-1},$$

$$s_T = \sum_{t=1}^{T} y_t x_t,$$

Linear regression: Proof idea

Offline optimal:

$$V(s_T, \sigma_T^2, T) = -\min_{\beta \in \mathbb{R}^p} \sum_{t=1}^{T} \left(\beta^\top x_t - y_t\right)^2 \qquad \beta_T = P_T s_T,$$

$$= -\sigma_T^2 + \beta_T^\top s_T \qquad P_T = \left(\sum_{t=1}^{T} x_t x_t^\top\right)^{-1},$$

$$s_T = \sum_{t=1}^{T} y_t x_t,$$

Linear regression: Proof idea

Offline optimal:

$$V(s_T, \sigma_T^2, T) = -\min_{\beta \in \mathbb{R}^p} \sum_{t=1}^{T} \left(\beta^\top x_t - y_t\right)^2 \qquad \beta_T = P_T s_T,$$

$$= -\sigma_T^2 + \beta_T^\top s_T \qquad P_T = \left(\sum_{t=1}^{T} x_t x_t^\top\right)^{-1},$$

$$s_T = \sum_{t=1}^{T} y_t x_t,$$

$$\sigma_T^2 = \sum_{t=1}^{T} y_t^2.$$

Linear regression: Proof idea

Offline optimal:

$$V(s_T, \sigma_T^2, T) = -\min_{\beta \in \mathbb{R}^p} \sum_{t=1}^{T} \left(\beta^\top x_t - y_t\right)^2 \qquad \beta_T = P_T s_T,$$

$$= -\sigma_T^2 + \beta_T^\top s_T \qquad P_T = \left(\sum_{t=1}^{T} x_t x_t^\top\right)^{-1},$$

$$= -\sigma_T^2 + s_T^\top P_T s_T. \qquad s_T = \sum_{t=1}^{T} y_t x_t,$$

$$\sigma_T^2 = \sum_{t=1}^{T} y_t^2.$$

Linear regression: Proof idea

Value to go

We'll show by induction that

$$V(s_t, \sigma_t^2, t) = s_t^\top P_t s_t - \sigma_t^2 + \gamma_t.$$

Linear regression: Proof idea

Value to go

We'll show by induction that

$$V(s_t, \sigma_t^2, t) = s_t^\top P_t s_t - \sigma_t^2 + \gamma_t.$$

It's true for $T = t$ with $\gamma_T = 0$.

Linear regression: Proof idea

Value to go

We'll show by induction that

$$V(s_t, \sigma_t^2, t) = s_t^\top P_t s_t - \sigma_t^2 + \gamma_t.$$

It's true for $T = t$ with $\gamma_T = 0$. Then

$$V(s_t, \sigma_t^2, t) = \min_{\hat{y}_{t+1}} \max_{y_{t+1}} (\hat{y}_{t+1} - y_{t+1})^2 + V(s_t + y_{t+1} x_{t+1}, \sigma_t^2 + y_{t+1}^2, t+1)$$

Linear regression: Proof idea

Value to go

We'll show by induction that

$$V(s_t, \sigma_t^2, t) = s_t^\top P_t s_t - \sigma_t^2 + \gamma_t.$$

It's true for $T = t$ with $\gamma_T = 0$. Then

$$\begin{aligned} V(s_t, \sigma_t^2, t) &= \min_{\hat{y}_{t+1}} \max_{y_{t+1}} (\hat{y}_{t+1} - y_{t+1})^2 + V(s_t + y_{t+1}x_{t+1}, \sigma_t^2 + y_{t+1}^2, t+1) \\ &= \min_{\hat{y}_{t+1}} \left(\hat{y}_{t+1}^2 + \max_{y_{t+1}} 2 \left(x_{t+1}^\top P_{t+1} s_t - \hat{y}_{t+1} \right) y_{t+1} \right. \\ &\qquad \left. + x_{t+1}^\top P_{t+1} x_{t+1} y_{t+1}^2 \right) + s_t^\top P_{t+1} s_t - \sigma_t^2 + \gamma_{t+1}. \end{aligned}$$

Linear regression: Proof idea

Value to go

The inner maximization is of a convex quadratic.

Linear regression: Proof idea

Value to go

The inner maximization is of a convex quadratic.
It is maximized by $y_{t+1} \in \{-B, B\}$:

Linear regression: Proof idea

Value to go

The inner maximization is of a convex quadratic.
It is maximized by $y_{t+1} \in \{-B, B\}$:

$$V(s_t, \sigma_t^2, t) = \min_{\hat{y}_{t+1}} \left(\hat{y}_{t+1}^2 + 2B \left| x_{t+1}^\top P_{t+1} s_t - \hat{y}_{t+1} \right| \right) + s_t^\top P_{t+1} s_t - \sigma_t^2 + \gamma_{t+1} + B^2 x_{t+1}^\top P_{t+1} x_{t+1}$$

Linear regression: Proof idea

Value to go

The inner maximization is of a convex quadratic.
It is maximized by $y_{t+1} \in \{-B, B\}$:

$$V(s_t, \sigma_t^2, t) = \min_{\hat{y}_{t+1}} \left(\hat{y}_{t+1}^2 + 2B \left| x_{t+1}^\top P_{t+1} s_t - \hat{y}_{t+1} \right| \right) + s_t^\top P_{t+1} s_t - \sigma_t^2 + \gamma_{t+1} + B^2 x_{t+1}^\top P_{t+1} x_{t+1}$$

Provided the problem is not too constrained (i.e., $B \geq |x_{t+1}^\top P_{t+1} s_t|$), the solution is $\hat{y}_{t+1} = x_{t+1}^\top P_{t+1} s_t$.

Linear regression: Proof idea

Value to go

The inner maximization is of a convex quadratic.
It is maximized by $y_{t+1} \in \{-B, B\}$:

$$V(s_t, \sigma_t^2, t) = \min_{\hat{y}_{t+1}} \left(\hat{y}_{t+1}^2 + 2B \left| x_{t+1}^\top P_{t+1} s_t - \hat{y}_{t+1} \right| \right) + s_t^\top P_{t+1} s_t - \sigma_t^2 + \gamma_{t+1} + B^2 x_{t+1}^\top P_{t+1} x_{t+1}$$

Provided the problem is not too constrained (i.e., $B \geq |x_{t+1}^\top P_{t+1} s_t|$), the solution is $\hat{y}_{t+1} = x_{t+1}^\top P_{t+1} s_t$.
(Otherwise, the player should clip $\hat{y}_{t+1}$ to B or $-B$.)

Linear regression: Proof idea

Value to go

The inner maximization is of a convex quadratic.
It is maximized by $y_{t+1} \in \{-B, B\}$:

$$\begin{aligned}
V(s_t, \sigma_t^2, t) &= \min_{\hat{y}_{t+1}} \left(\hat{y}_{t+1}^2 + 2B \left| x_{t+1}^\top P_{t+1} s_t - \hat{y}_{t+1} \right| \right) + s_t^\top P_{t+1} s_t \\
&\quad - \sigma_t^2 + \gamma_{t+1} + B^2 x_{t+1}^\top P_{t+1} x_{t+1} \\
&= s_t^\top P_{t+1} x_{t+1} x_{t+1}^\top P_{t+1} s_t + s_t^\top P_{t+1} s_t \\
&\quad - \sigma_t^2 + \gamma_{t+1} + B^2 x_{t+1}^\top P_{t+1} x_{t+1}.
\end{aligned}$$

Provided the problem is not too constrained (i.e., $B \geq |x_{t+1}^\top P_{t+1} s_t|$), the solution is $\hat{y}_{t+1} = x_{t+1}^\top P_{t+1} s_t$.
(Otherwise, the player should clip $\hat{y}_{t+1}$ to B or $-B$.)

Linear regression: Proof idea

Value to go

$$V(s_t, \sigma_t^2, t) = s_t^\top \left(P_{t+1} x_{t+1} x_{t+1}^\top P_{t+1} + P_{t+1} \right) s_t$$
$$- \sigma_t^2 + \gamma_{t+1} + B^2 x_{t+1}^\top P_{t+1} x_{t+1}.$$

Linear regression: Proof idea

Value to go

$$V(s_t, \sigma_t^2, t) = s_t^\top \left(P_{t+1} x_{t+1} x_{t+1}^\top P_{t+1} + P_{t+1}\right) s_t - \sigma_t^2 + \gamma_{t+1} + B^2 x_{t+1}^\top P_{t+1} x_{t+1}.$$

Optimal predictions

$$\hat{y}_{n+1} = x_{n+1}^\top P_{n+1} s_n, \qquad s_n = \sum_{t=1}^{n} y_t x_t,$$

$$P_T = \left(\sum_{t=1}^{T} x_t x_t^\top\right)^{-1}, \qquad P_n = P_{n+1} x_{n+1} x_{n+1}^\top P_{n+1} + P_{n+1}.$$

Linear regression: Proof idea

Optimal predictions

$$\hat{y}_{n+1} = x_{n+1}^{\top} P_{n+1} s_n, \qquad s_n = \sum_{t=1}^{n} y_t x_t,$$

$$P_T = \left(\sum_{t=1}^{T} x_t x_t^{\top} \right)^{-1}, \qquad P_n = P_{n+1} x_{n+1} x_{n+1}^{\top} P_{n+1} + P_{n+1}.$$

Linear regression: Proof idea

Optimal predictions

$$\hat{y}_{n+1} = x_{n+1}^\top P_{n+1} s_n, \qquad s_n = \sum_{t=1}^{n} y_t x_t,$$

$$P_T = \left(\sum_{t=1}^{T} x_t x_t^\top \right)^{-1}, \qquad P_n = P_{n+1} x_{n+1} x_{n+1}^\top P_{n+1} + P_{n+1}.$$

Example

Suppose we wish to estimate a mean, and the covariate is a fixed scalar, say $x_n = 1$.

Linear regression: Proof idea

Optimal predictions

$$\hat{y}_{n+1} = x_{n+1}^\top P_{n+1} s_n, \qquad s_n = \sum_{t=1}^{n} y_t x_t,$$

$$P_T = \left(\sum_{t=1}^{T} x_t x_t^\top \right)^{-1}, \qquad P_n = P_{n+1} x_{n+1} x_{n+1}^\top P_{n+1} + P_{n+1}.$$

Example

Suppose we wish to estimate a mean, and the covariate is a fixed scalar, say $x_n = 1$.
Then $P_T = 1/T$ and P_n evolves as α_n in the ball game.

Linear regression: Proof idea

An alternative recurrence

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

Proof

- The result is true for $n = T$:

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

Linear regression: Proof idea

Proof

- If it's true for n, we apply the Sherman-Morrison formula:

$$\left(A + uv^\top\right)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u},$$

Linear regression: Proof idea

Proof

- If it's true for n, we apply the Sherman-Morrison formula:

$$\left(A + uv^\top\right)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u},$$

which implies

$$P_{n-1}^{-1} = \left(P_n + P_n x_n x_n^\top P_n\right)^{-1}$$

Linear regression: Proof idea

Proof

- If it's true for n, we apply the Sherman-Morrison formula:

$$\left(A + uv^\top\right)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u},$$

which implies

$$\begin{aligned} P_{n-1}^{-1} &= \left(P_n + P_n x_n x_n^\top P_n\right)^{-1} \\ &= P_n^{-1} - \frac{x_n x_n^\top}{1 + x_n^\top P_n x_n} \end{aligned}$$

Linear regression: Proof idea

Proof

- If it's true for n, we apply the Sherman-Morrison formula:

$$\left(A + uv^\top\right)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u},$$

which implies

$$\begin{aligned}
P_{n-1}^{-1} &= \left(P_n + P_n x_n x_n^\top P_n\right)^{-1} \\
&= P_n^{-1} - \frac{x_n x_n^\top}{1 + x_n^\top P_n x_n} \\
&= \sum_{t=1}^{n} x_t x_t^\top + \sum_{t=n+1}^{T} \frac{x_t^\top P_t x_t}{1 + x_t^\top P_t x_t} x_t x_t^\top - \frac{x_n x_n^\top}{1 + x_n^\top P_n x_n}
\end{aligned}$$

Proof

- If it's true for n, we apply the Sherman-Morrison formula:

$$\left(A + uv^\top\right)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u},$$

which implies

$$\begin{aligned}
P_{n-1}^{-1} &= \left(P_n + P_n x_n x_n^\top P_n\right)^{-1} \\
&= P_n^{-1} - \frac{x_n x_n^\top}{1 + x_n^\top P_n x_n} \\
&= \sum_{t=1}^{n} x_t x_t^\top + \sum_{t=n+1}^{T} \frac{x_t^\top P_t x_t}{1 + x_t^\top P_t x_t} x_t x_t^\top - \frac{x_n x_n^\top}{1 + x_n^\top P_n x_n} \\
&= \sum_{t=1}^{n-1} x_t x_t^\top + \sum_{t=n}^{T} \frac{x_t^\top P_t x_t}{1 + x_t^\top P_t x_t} x_t x_t^\top.
\end{aligned}$$

Linear regression: Regret

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

Theorem

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

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.

Linear regression: Alternative constraints

Ellipsoid constraints (weighted 2-norm)

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

Linear regression: Alternative constraints

Ellipsoid constraints (weighted 2-norm)

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

Minimax strategy: linear

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

Linear regression: Alternative constraints

Ellipsoid constraints (weighted 2-norm)

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

Minimax regret $= R$.

Minimax strategy: linear

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

Linear regression: Alternative constraints

Ellipsoid constraints (weighted 2-norm)

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

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: Alternative constraints

Ellipsoid constraints (weighted 2-norm)

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

Minimax regret $= R$.

Minimax strategy: linear

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

Equalizer property

For all $y_1, \ldots, y_T$,

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

$$= \sum_{t=1}^{T} y_t^2 x_t^\top P_t x_t.$$

Linear regression: Alternative constraints

Equalizer property

For all $y_1, \ldots, y_T$,

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

(Proof is by induction.)

Linear regression: Alternative constraints

Equalizer property

For all $y_1, \ldots, y_T$,

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

(Proof is by induction.)

Corollary

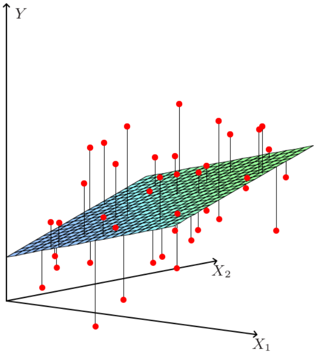
For every R, (MM) is minimax optimal on

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

Outline

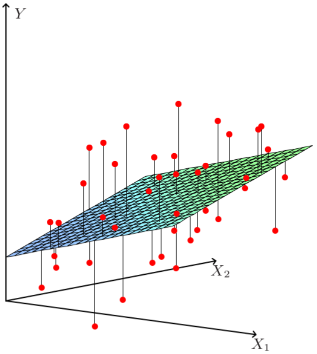
- Computing minimax optimal strategies.
- Part 1: Euclidean loss.
- Part 2: Linear regression.
  - Fixed design.
  - Minimax strategy is regularized least squares.
  - Box and ellipsoid constraints.
  - **Adversarial covariates.**

Linear regression: Adversarial covariates



(B., Malek, 2017)

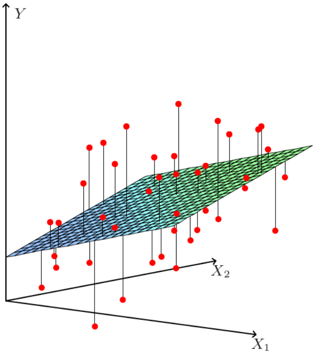
Linear regression: Adversarial covariates



Protocol

(B., Malek, 2017)

Linear regression: Adversarial covariates

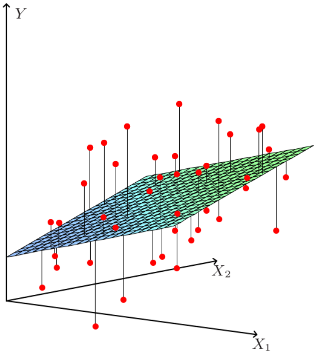


Protocol

Given: T;

(B., Malek, 2017)

Linear regression: Adversarial covariates

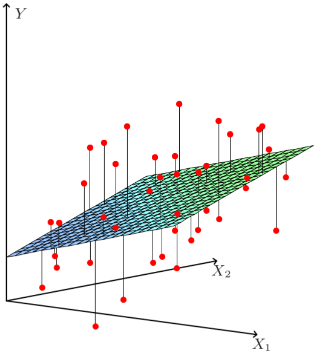


Protocol

Given: T; $\mathcal{X}^T \subset (\mathbb{R}^p)^T$;

(B., Malek, 2017)

Linear regression: Adversarial covariates

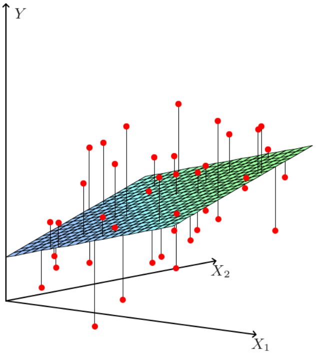


Protocol

Given: T; $\mathcal{X}^T \subset (\mathbb{R}^p)^T$; $\mathcal{Y}^T \subset \mathbb{R}^T$.

(B., Malek, 2017)

Linear regression: Adversarial covariates



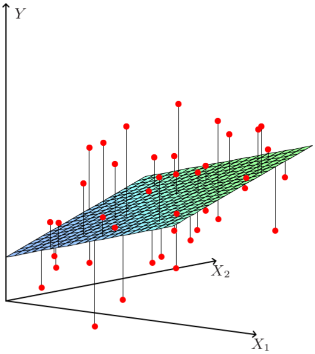
Protocol

Given: T; $\mathcal{X}^T \subset (\mathbb{R}^p)^T$; $\mathcal{Y}^T \subset \mathbb{R}^T$.

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

(B., Malek, 2017)

Linear regression: Adversarial covariates



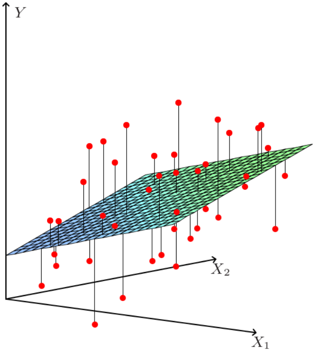
Protocol

Given: T; $\mathcal{X}^T \subset (\mathbb{R}^p)^T$; $\mathcal{Y}^T \subset \mathbb{R}^T$.
For $t = 1, 2, \ldots, T$:

- Adversary reveals $x_t \in \mathbb{R}^p$

(B., Malek, 2017)

Linear regression: Adversarial covariates



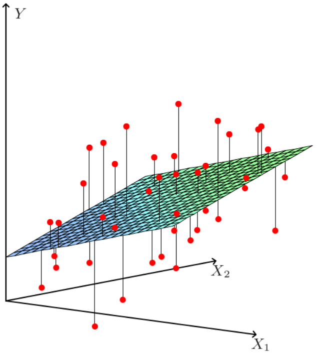
Protocol

Given: T; $\mathcal{X}^T \subset (\mathbb{R}^p)^T$; $\mathcal{Y}^T \subset \mathbb{R}^T$.
For $t = 1, 2, \ldots, T$:

- Adversary reveals $x_t \in \mathbb{R}^p$ $(x_1^T \in \mathcal{X}^T)$

(B., Malek, 2017)

Linear regression: Adversarial covariates



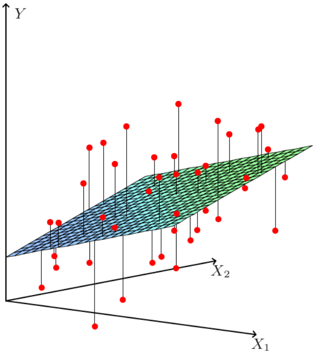
Protocol

Given: T; $\mathcal{X}^T \subset (\mathbb{R}^p)^T$; $\mathcal{Y}^T \subset \mathbb{R}^T$.
For $t = 1, 2, \ldots, T$:

- Adversary reveals $x_t \in \mathbb{R}^p$ ($x_1^T \in \mathcal{X}^T$)
- Learner predicts $\hat{y}_t \in \mathbb{R}$

(B., Malek, 2017)

Linear regression: Adversarial covariates



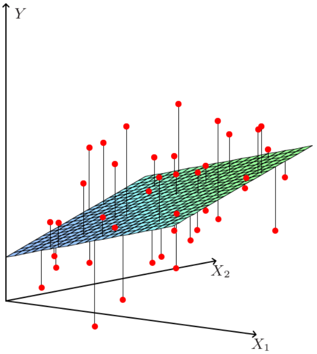
Protocol

Given: T; $\mathcal{X}^T \subset (\mathbb{R}^p)^T$; $\mathcal{Y}^T \subset \mathbb{R}^T$.
For $t = 1, 2, \ldots, T$:

- Adversary reveals $x_t \in \mathbb{R}^p$ $(x_1^T \in \mathcal{X}^T)$
- Learner predicts $\hat{y}_t \in \mathbb{R}$
- Adversary reveals $y_t \in \mathbb{R}$

(B., Malek, 2017)

Linear regression: Adversarial covariates



Protocol

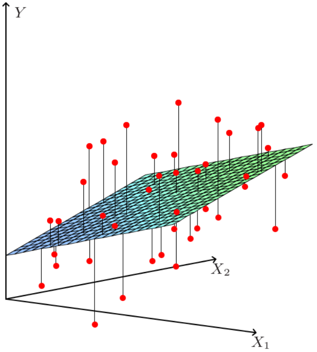
Given: T; $\mathcal{X}^T \subset (\mathbb{R}^p)^T$; $\mathcal{Y}^T \subset \mathbb{R}^T$.

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

- Adversary reveals $x_t \in \mathbb{R}^p$ $(x_1^T \in \mathcal{X}^T)$
- Learner predicts $\hat{y}_t \in \mathbb{R}$
- Adversary reveals $y_t \in \mathbb{R}$ $(y_1^T \in \mathcal{Y}^T)$

(B., Malek, 2017)

Linear regression: Adversarial covariates



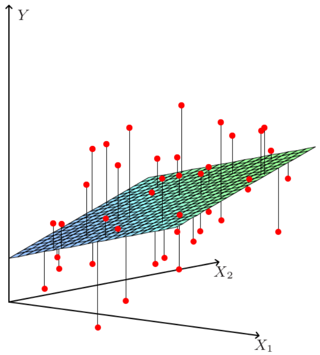
Protocol

Given: T; $\mathcal{X}^T \subset (\mathbb{R}^p)^T$; $\mathcal{Y}^T \subset \mathbb{R}^T$.
For $t = 1, 2, \ldots, T$:

- Adversary reveals $x_t \in \mathbb{R}^p$ $(x_1^T \in \mathcal{X}^T)$
- Learner predicts $\hat{y}_t \in \mathbb{R}$
- Adversary reveals $y_t \in \mathbb{R}$ $(y_1^T \in \mathcal{Y}^T)$
- Learner incurs loss $(\hat{y}_t - y_t)^2$.

(B., Malek, 2017)

Linear regression: Adversarial covariates



Protocol

Given: T; $\mathcal{X}^T \subset (\mathbb{R}^p)^T$; $\mathcal{Y}^T \subset \mathbb{R}^T$.
For $t = 1, 2, \ldots, T$:

- Adversary reveals $x_t \in \mathbb{R}^p$ $(x_1^T \in \mathcal{X}^T)$
- Learner predicts $\hat{y}_t \in \mathbb{R}$
- Adversary reveals $y_t \in \mathbb{R}$ $(y_1^T \in \mathcal{Y}^T)$
- Learner incurs loss $(\hat{y}_t - y_t)^2$.

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

(B., Malek, 2017)

Linear regression: Adversarial covariates

A covariance budget

Recall:

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

$$P_n = P_{n+1} x_{n+1} x_{n+1}^\top P_{n+1} + P_{n+1}.$$

Linear regression: Adversarial covariates

A covariance budget

Recall:

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

$$P_n = P_{n+1} x_{n+1} x_{n+1}^\top P_{n+1} + P_{n+1}.$$

Equivalently,

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

Linear regression: Adversarial covariates

A covariance budget

Recall:

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

$$P_n = P_{n+1} x_{n+1} x_{n+1}^\top P_{n+1} + P_{n+1}.$$

Equivalently,

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

Define

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

Linear regression

A reformulation

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

Linear regression

A reformulation

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

Theorem

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

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

Proof

Fix $\tilde{P}_0$, define $\tilde{P}_t$ by the forward iteration,

$$\tilde{P}_{t+1} = \tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t,$$

Proof

Fix $\tilde{P}_0$, define $\tilde{P}_t$ by the forward iteration,

$$\tilde{P}_{t+1} = \tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t,$$

then set $P_T = \tilde{P}_T$ and define P_t by the backwards iteration,

$$P_{t-1} = P_t + P_t x_t x_t^\top P_t.$$

Linear regression

Proof

Fix $\tilde{P}_0$, define $\tilde{P}_t$ by the forward iteration,

$$\tilde{P}_{t+1} = \tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t,$$

then set $P_T = \tilde{P}_T$ and define P_t by the backwards iteration,

$$P_{t-1} = P_t + P_t x_t x_t^\top P_t.$$

We'll show that $P_t = \tilde{P}_t$ for $t = T-1, \dots, 1$.

Linear regression

Proof

Suppose $P_{t+1} = \tilde{P}_{t+1}$. Then

$$P_t = P_{t+1} + P_{t+1} x_{t+1} x_{t+1}^\top P_{t+1}$$

Linear regression

Proof

Suppose $P_{t+1} = \tilde{P}_{t+1}$. Then

$$P_t = \tilde{P}_{t+1} + \tilde{P}_{t+1} x_{t+1} x_{t+1}^\top \tilde{P}_{t+1}$$

Proof

Suppose $P_{t+1} = \tilde{P}_{t+1}$. Then

$$
\begin{aligned}
P_t &= \tilde{P}_{t+1} + \tilde{P}_{t+1} x_{t+1} x_{t+1}^\top \tilde{P}_{t+1} \\
&= \tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \\
&\qquad + \left(\tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \right) x_{t+1} x_{t+1}^\top \left(\tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \right)
\end{aligned}
$$

Linear regression

Proof

Suppose $P_{t+1} = \tilde{P}_{t+1}$. Then

$$
\begin{aligned}
P_t &= \tilde{P}_{t+1} + \tilde{P}_{t+1} x_{t+1} x_{t+1}^\top \tilde{P}_{t+1} \\
&= \tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \\
&\qquad + \left(\tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \right) x_{t+1} x_{t+1}^\top \left(\tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \right) \\
&= \tilde{P}_t + \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \left(-\frac{a_t}{b_t^2} + 1 - 2a_t + a_t^2 \right)
\end{aligned}
$$

Linear regression

Proof

Suppose $P_{t+1} = \tilde{P}_{t+1}$. Then

$$\begin{aligned}
P_t &= \tilde{P}_{t+1} + \tilde{P}_{t+1} x_{t+1} x_{t+1}^\top \tilde{P}_{t+1} \\
&= \tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \\
&\qquad + \left(\tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \right) x_{t+1} x_{t+1}^\top \left(\tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \right) \\
&= \tilde{P}_t + \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \left(-\frac{a_t}{b_t^2} + 1 - 2a_t + a_t^2 \right) \\
&= \tilde{P}_t + \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \left((1 - a_t)^2 - \frac{a_t}{b_t^2} \right)
\end{aligned}$$

Linear regression

Proof

Suppose $P_{t+1} = \tilde{P}_{t+1}$. Then

$$\begin{aligned}
P_t &= \tilde{P}_{t+1} + \tilde{P}_{t+1} x_{t+1} x_{t+1}^\top \tilde{P}_{t+1} \\
&= \tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \\
&\qquad + \left(\tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \right) x_{t+1} x_{t+1}^\top \left(\tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \right) \\
&= \tilde{P}_t + \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \left(-\frac{a_t}{b_t^2} + 1 - 2a_t + a_t^2 \right) \\
&= \tilde{P}_t + \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \left((1 - a_t)^2 - \frac{a_t}{b_t^2} \right) \\
&= \tilde{P}_t,
\end{aligned}$$

Linear regression

Proof

Suppose $P_{t+1} = \tilde{P}_{t+1}$. Then

$$
\begin{aligned}
P_t &= \tilde{P}_{t+1} + \tilde{P}_{t+1} x_{t+1} x_{t+1}^\top \tilde{P}_{t+1} \\
&= \tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \\
&\qquad + \left(\tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \right) x_{t+1} x_{t+1}^\top \left(\tilde{P}_t - \frac{a_t}{b_t^2} \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \right) \\
&= \tilde{P}_t + \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \left(-\frac{a_t}{b_t^2} + 1 - 2a_t + a_t^2 \right) \\
&= \tilde{P}_t + \tilde{P}_t x_{t+1} x_{t+1}^\top \tilde{P}_t \left((1 - a_t)^2 - \frac{a_t}{b_t^2} \right) \\
&= \tilde{P}_t,
\end{aligned}
$$

where we have used $(1 - a_t)^2 = \frac{a_t}{b_t^2}$.

Linear regression

Proof

To see that $(1 - a_t)^2 = \frac{a_t}{b_t^2}$:

Proof

To see that $(1 - a_t)^2 = \frac{a_t}{b_t^2}$:

$$b_t^2 (1 - a_t)^2 = b_t^2 \left(1 - \frac{\sqrt{4b_t^2 + 1} - 1}{\sqrt{4b_t^2 + 1} + 1} \right)^2$$

Proof

To see that $(1 - a_t)^2 = \frac{a_t}{b_t^2}$:

$$b_t^2(1 - a_t)^2 = b_t^2 \left(1 - \frac{\sqrt{4b_t^2 + 1} - 1}{\sqrt{4b_t^2 + 1} + 1}\right)^2$$
$$= \left(\frac{2b_t}{\sqrt{4b_t^2 + 1} + 1}\right)^2$$

Proof

To see that $(1 - a_t)^2 = \frac{a_t}{b_t^2}$:

$$b_t^2(1 - a_t)^2 = b_t^2 \left(1 - \frac{\sqrt{4b_t^2 + 1} - 1}{\sqrt{4b_t^2 + 1} + 1}\right)^2$$
$$= \left(\frac{2b_t}{\sqrt{4b_t^2 + 1} + 1}\right)^2$$
$$= \frac{\sqrt{4b_t^2 + 1} - 1}{\sqrt{4b_t^2 + 1} + 1}$$

Proof

To see that $(1 - a_t)^2 = \frac{a_t}{b_t^2}$:

$$\begin{aligned} b_t^2(1-a_t)^2 &= b_t^2\left(1 - \frac{\sqrt{4b_t^2+1}-1}{\sqrt{4b_t^2+1}+1}\right)^2 \\ &= \left(\frac{2b_t}{\sqrt{4b_t^2+1}+1}\right)^2 \\ &= \frac{\sqrt{4b_t^2+1}-1}{\sqrt{4b_t^2+1}+1} \\ &= a_t, \end{aligned}$$

Proof

To see that $(1 - a_t)^2 = \frac{a_t}{b_t^2}$:

$$\begin{aligned}
b_t^2(1 - a_t)^2 &= b_t^2 \left(1 - \frac{\sqrt{4b_t^2 + 1} - 1}{\sqrt{4b_t^2 + 1} + 1}\right)^2 \\
&= \left(\frac{2b_t}{\sqrt{4b_t^2 + 1} + 1}\right)^2 \\
&= \frac{\sqrt{4b_t^2 + 1} - 1}{\sqrt{4b_t^2 + 1} + 1} \\
&= a_t,
\end{aligned}$$

because

$$\left(\sqrt{4b_t^2 + 1} - 1\right)\left(\sqrt{4b_t^2 + 1} + 1\right) = 4b_t^2.$$

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.

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, under the forward iteration,
$$P_T^{-1} = \sum_{q=1}^{T} x_q x_q^\top.$$

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, under the forward iteration,
$$P_T^{-1} = \sum_{q=1}^{T} x_q x_q^\top .$$
2. $P_t^{-1} \succeq \sum_{q=1}^{t} x_q x_q^\top$.

Linear regression

Proof idea

Suppose that $P_t^{-1} - \sum_{q=1}^{t} x_q x_q^\top \succeq 0$.

Linear regression

Proof idea

Suppose that $P_t^{-1} - \sum_{q=1}^{t} x_q x_q^\top \succeq 0$. Write

$$P_t^{-1} - \sum_{q=1}^{t} x_q x_q^\top = \sum_{i=1}^{m} \lambda_i v_i v_i^\top,$$

with orthonormal $v_1, \ldots, v_m$ and $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$.

Linear regression

Proof idea

Suppose that $P_t^{-1} - \sum_{q=1}^{t} x_q x_q^\top \succeq 0$. Write

$$P_t^{-1} - \sum_{q=1}^{t} x_q x_q^\top = \sum_{i=1}^{m} \lambda_i v_i v_i^\top,$$

with orthonormal $v_1, \ldots, v_m$ and $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$. Then it's easy to show that we can choose $x_{t+1} = \beta v_k$ (for some $\beta \geq 0$) such that

$$P_{t+1}^{-1} - \sum_{q=1}^{t+1} x_q x_q^\top = \sum_{i=1}^{m-1} \lambda_i v_i v_i^\top.$$

Linear regression

Proof idea

Suppose that $P_t^{-1} - \sum_{q=1}^{t} x_q x_q^\top \succeq 0$. Write

$$P_t^{-1} - \sum_{q=1}^{t} x_q x_q^\top = \sum_{i=1}^{m} \lambda_i v_i v_i^\top,$$

with orthonormal $v_1, \ldots, v_m$ and $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$. Then it's easy to show that we can choose $x_{t+1} = \beta v_k$ (for some $\beta \geq 0$) such that

$$P_{t+1}^{-1} - \sum_{q=1}^{t+1} x_q x_q^\top = \sum_{i=1}^{m-1} \lambda_i v_i v_i^\top.$$

Actually we can choose β as any smaller value to ensure that the rank does not drop in one step, so we can "complete the sequence" in any number of steps, provided that it is at least m.

Proof idea

To see the reverse implication, once we have computed the P_ts by the forward iteration, we can write the equivalent expression

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

Linear regression

Proof idea

To see the reverse implication, once we have computed the P_ts by the forward iteration, we can write the equivalent expression

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

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, under the forward iteration,
$$P_T^{-1} = \sum_{q=1}^{T} x_q x_q^\top.$$
2. $P_t^{-1} \succeq \sum_{q=1}^{t} x_q x_q^\top$.

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, under the forward iteration,
$$P_T^{-1} = \sum_{q=1}^{T} x_q x_q^\top.$$
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 $x_1, \ldots, x_T$.

Linear regression

Legal covariate sequences

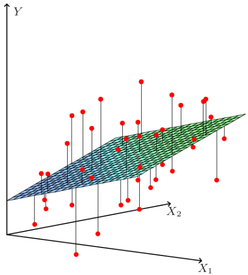
For any $t \geq 0$, any $x_1, \ldots, x_t$ and any P_t, the following two conditions are equivalent.

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

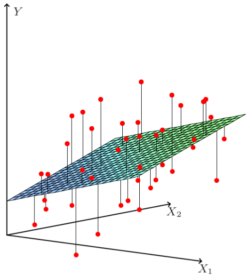
Adversarial covariates

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

Linear regression: Adversarial covariates; horizon-free

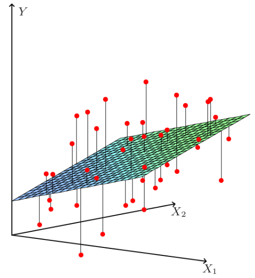


Linear regression: Adversarial covariates; horizon-free



Protocol

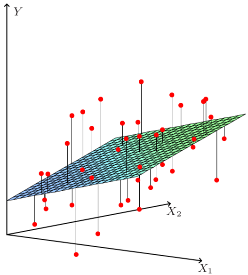
Linear regression: Adversarial covariates; horizon-free



Protocol

Given:

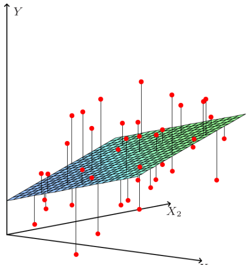
Linear regression: Adversarial covariates; horizon-free



Protocol

Given: $\mathcal{Z} \subset \bigcup_{T \geq 1} (\mathbb{R}^p \times \mathbb{R})^T$.

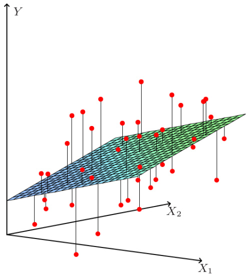
Linear regression: Adversarial covariates; horizon-free



Protocol

Given: $\mathcal{Z} \subset \bigcup_{T \geq 1} (\mathbb{R}^p \times \mathbb{R})^T$.
For $t = 1, 2, 3, \ldots$

Linear regression: Adversarial covariates; horizon-free

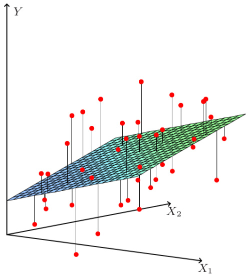


Protocol

Given: $\mathcal{Z} \subset \bigcup_{T \geq 1} (\mathbb{R}^p \times \mathbb{R})^T$.
For $t = 1, 2, 3, \ldots$

- Adversary reveals $x_t \in \mathbb{R}^p$

Linear regression: Adversarial covariates; horizon-free

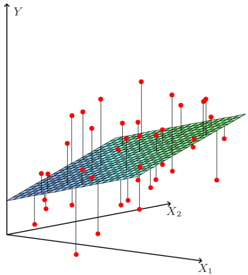


Protocol

Given: $\mathcal{Z} \subset \bigcup_{T \geq 1} (\mathbb{R}^p \times \mathbb{R})^T$.
For $t = 1, 2, 3, \ldots$

- Adversary reveals $x_t \in \mathbb{R}^p$
- Learner predicts $\hat{y}_t \in \mathbb{R}$

Linear regression: Adversarial covariates; horizon-free

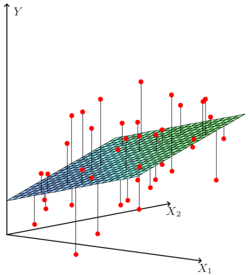


Protocol

Given: $\mathcal{Z} \subset \bigcup_{T\geq 1}(\mathbb{R}^p \times \mathbb{R})^T$.
For $t = 1, 2, 3, \ldots$

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

Linear regression: Adversarial covariates; horizon-free

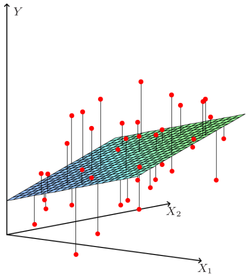


Protocol

Given: $\mathcal{Z} \subset \bigcup_{T \geq 1} (\mathbb{R}^p \times \mathbb{R})^T$.
For $t = 1, 2, 3, \ldots$

- Adversary reveals $x_t \in \mathbb{R}^p$
- Learner predicts $\hat{y}_t \in \mathbb{R}$
- Adversary reveals $y_t \in \mathbb{R}$ $\qquad (x_1^T, y_1^T \in \mathcal{Z})$

Linear regression: Adversarial covariates; horizon-free

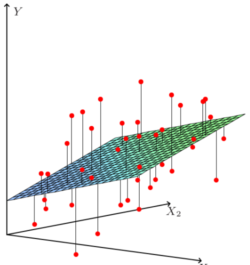


Protocol

Given: $\mathcal{Z} \subset \bigcup_{T \geq 1} (\mathbb{R}^p \times \mathbb{R})^T$.
For $t = 1, 2, 3, \ldots$

- Adversary reveals $x_t \in \mathbb{R}^p$
- Learner predicts $\hat{y}_t \in \mathbb{R}$
- Adversary reveals $y_t \in \mathbb{R}$ $\quad (x_1^T, y_1^T \in \mathcal{Z})$
- Learner incurs loss $(\hat{y}_t - y_t)^2$.

Linear regression: Adversarial covariates; horizon-free



Protocol

Given: $\mathcal{Z} \subset \bigcup_{T\geq 1}(\mathbb{R}^p \times \mathbb{R})^T$.
For $t = 1, 2, 3, \ldots$

- Adversary reveals $x_t \in \mathbb{R}^p$
- Learner predicts $\hat{y}_t \in \mathbb{R}$
- Adversary reveals $y_t \in \mathbb{R}$ $\qquad (x_1^T, y_1^T \in \mathcal{Z})$
- Learner incurs loss $(\hat{y}_t - y_t)^2$.

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

Linear regression

Constraints on y_ts

1. *Box constraints*: $\mathcal{B}(B) := \{y_1^T : |y_t| \leq B_t\}$, for $B_1, B_2, \ldots > 0$.

Linear regression

Constraints on y_ts

1. *Box constraints*: $\mathcal{B}(B) := \{y_1^T : |y_t| \leq B_t\}$, for $B_1, B_2, \ldots > 0$.
2. *Ellipsoidal constraints*:

$$\mathcal{E}(x_1^T, R) := \left\{ y_1^T : \sum_{t=1}^{T} y_t^2 x_t^\top P_t x_t \leq R \right\}.$$

Linear regression

Constraints on x_ts

1. *Compatibility constraints*:

$$\mathcal{X}(B) = \left\{ x_1^T : B_t \geq \sum_{s=1}^{t-1} \left| x_t^\top P_t x_s \right| B_s \text{ for } 2 \leq t \leq T \right\}.$$

Linear regression

Constraints on x_ts

1. *Compatibility constraints*:

$$\mathcal{X}(B) = \left\{ x_1^T : B_t \geq \sum_{s=1}^{t-1} \left| x_t^\top P_t x_s \right| B_s \text{ for } 2 \leq t \leq T \right\}.$$

2. *Covariance constraints*:

$$\overline{\mathcal{X}}(\Sigma) = \left\{ x_1^T : \text{for } P_0, \ldots, P_T \text{ defined by } x_1^T,\ P_0^{-1} = \Sigma \right\}.$$

Linear regression

Theorem

For all positive semidefinite Σ; $B_1, B_2, \ldots > 0$
the forward strategy s^* is horizon-independent minimax optimal,

Linear regression

Theorem

For all positive semidefinite Σ; $B_1, B_2, \ldots > 0$
the forward strategy s^* is horizon-independent minimax optimal,

$$\sup_T \sup_{x_1^T \in \mathcal{X}} \left(\sup_{y_1^T \in \mathcal{Y}(x_1^T)} R_T(s^*, x_1^T, y_1^T) - \min_s \sup_{y_1^T \in \mathcal{Y}(x_1^T)} R_T(s, x_1^T, y_1^T) \right) = 0.$$

with respect to the following $(\mathcal{X}, \mathcal{Y}(x_1^t))$:

$$(\mathcal{X}(B_1^T) \cap \overline{\mathcal{X}}(\Sigma), \mathcal{B}(B_1^T)),$$

Linear regression

Theorem

For all positive semidefinite Σ; $B_1, B_2, \ldots > 0$; and $R > 0$, the forward strategy s^* is horizon-independent minimax optimal,

$$\sup_T \sup_{x_1^T \in \mathcal{X}} \left(\sup_{y_1^T \in \mathcal{Y}(x_1^T)} R_T(s^*, x_1^T, y_1^T) - \min_s \sup_{y_1^T \in \mathcal{Y}(x_1^T)} R_T(s, x_1^T, y_1^T) \right) = 0.$$

with respect to the following $(\mathcal{X}, \mathcal{Y}(x_1^t))$:

$$(\mathcal{X}(B_1^T) \cap \overline{\mathcal{X}}(\Sigma), \mathcal{B}(B_1^T)), \qquad (\overline{\mathcal{X}}(\Sigma), \mathcal{E}(x_1^T, R)).$$

Theorem

For all positive semidefinite Σ; $B_1, B_2, \ldots > 0$; and $R > 0$, the forward strategy s^* is horizon-independent minimax optimal,

$$\sup_T \sup_{x_1^T \in \mathcal{X}} \left(\sup_{y_1^T \in \mathcal{Y}(x_1^T)} R_T(s^*, x_1^T, y_1^T) - \min_s \sup_{y_1^T \in \mathcal{Y}(x_1^T)} R_T(s, x_1^T, y_1^T) \right) = 0.$$

with respect to the following $(\mathcal{X}, \mathcal{Y}(x_1^t))$:

$$(\mathcal{X}(B_1^T) \cap \overline{\mathcal{X}}(\Sigma), \mathcal{B}(B_1^T)), \qquad (\overline{\mathcal{X}}(\Sigma), \mathcal{E}(x_1^T, R)).$$

That is, s^* performs as well as the best strategy that sees the covariate sequence.

The minimax strategy as regularized least squares

The minimax strategy predicts $\hat{y}_n = \hat{\theta}_n^\top x_n$, where $\hat{\theta}_n$ is the solution to

$$\min_{\theta} \sum_{t=1}^{n-1} (\theta^\top x_t - y_t)^2 + \theta^\top R_n \theta,$$

The minimax strategy as regularized least squares

The minimax strategy predicts $\hat{y}_n = \hat{\theta}_n^\top x_n$, where $\hat{\theta}_n$ is the solution to

$$\min_\theta \sum_{t=1}^{n-1} (\theta^\top x_t - y_t)^2 + \theta^\top R_n \theta,$$

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

The minimax strategy as regularized least squares

The minimax strategy predicts $\hat{y}_n = \hat{\theta}_n^\top x_n$, where $\hat{\theta}_n$ is the solution to

$$\min_{\theta} \sum_{t=1}^{n-1} (\theta^\top x_t - y_t)^2 + \theta^\top R_n \theta,$$

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

Indeed,

$$\hat{\theta}_n = \left(\sum_{t=1}^{n} x_t x_t^\top + R_n \right)^{-1} s_{n-1}$$

The minimax strategy as regularized least squares

The minimax strategy predicts $\hat{y}_n = \hat{\theta}_n^\top x_n$, where $\hat{\theta}_n$ is the solution to

$$\min_\theta \sum_{t=1}^{n-1} (\theta^\top x_t - y_t)^2 + \theta^\top R_n \theta,$$

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

Indeed,

$$\hat{\theta}_n = \left(\sum_{t=1}^{n} x_t x_t^\top + R_n \right)^{-1} s_{n-1} = P_n^{-1} s_{n-1}.$$

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.