Efficient Minimax Strategies for Online Prediction

Peter Bartlett

Computer Science and Statistics
University of California at Berkeley

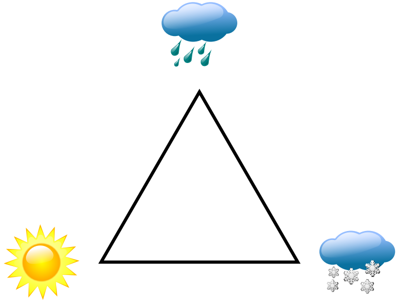
Mathematical Sciences
Queensland University of Technology

Joint work with Fares Hedayati, Wouter Koolen and Alan Malek

Online Prediction

A repeated game:

At round t:

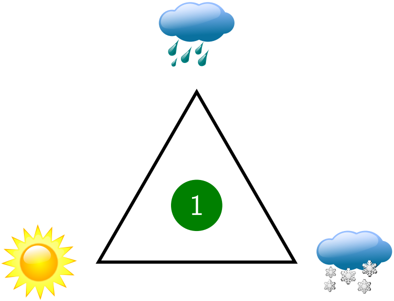


Online Prediction

A repeated game:

At round t:

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

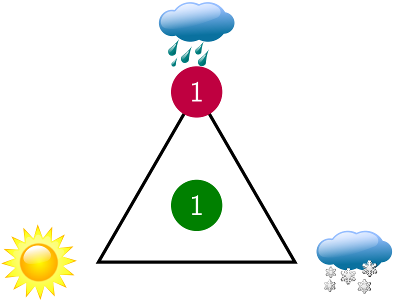


Online Prediction

A repeated game:

At round t:

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



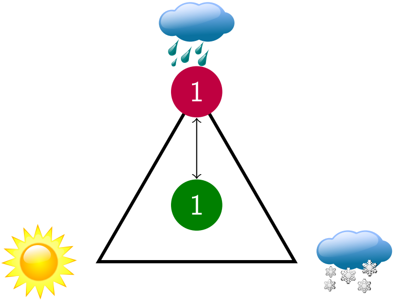
Online Prediction

A repeated game:

At round t:

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

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



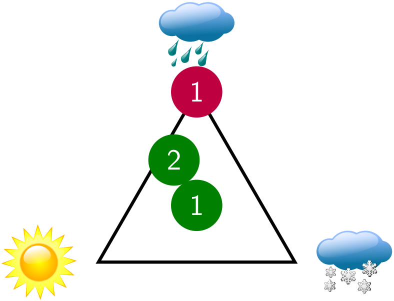
Brier loss

Online Prediction

A repeated game:

At round t:

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

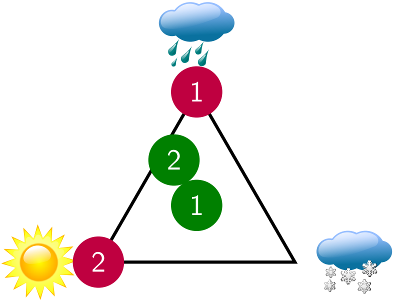


Online Prediction

A repeated game:

At round t:

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

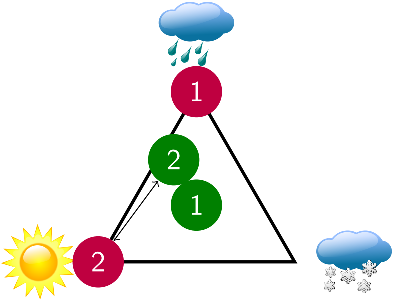


Online Prediction

A repeated game:

At round t:

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

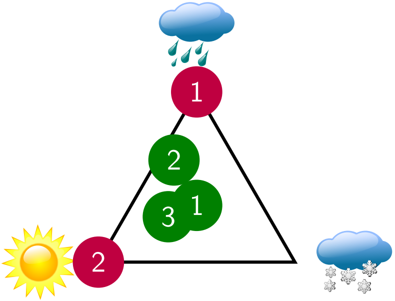


Online Prediction

A repeated game:

At round t:

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

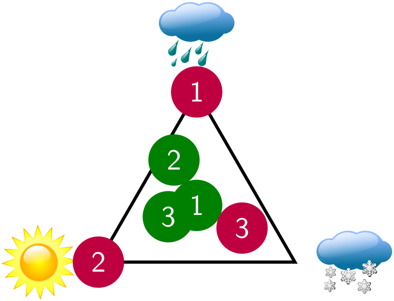


Online Prediction

A repeated game:

At round t:

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

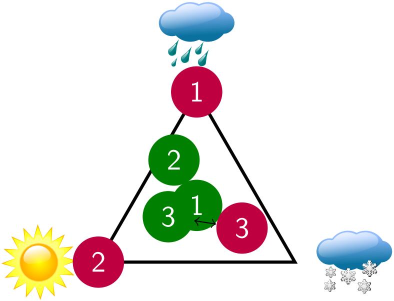


Online Prediction

A repeated game:

At round t:

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



Online Prediction

A repeated game:

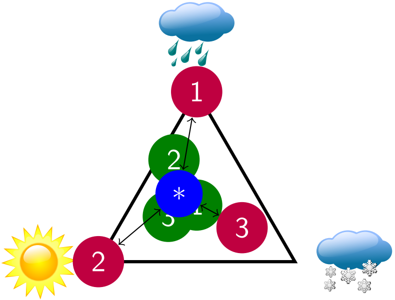
At round t:

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

Player's aim:

Minimize *regret:*

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



Online Prediction

A repeated game:

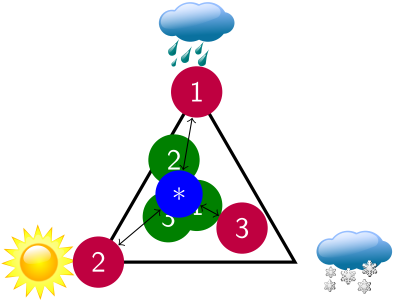
At round t:

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

Player's aim:

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

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



Online Prediction Games: Why

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

Online Prediction Games: Why

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

Online Prediction Games: Why

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

Online Prediction Games: Why

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

Online Prediction Games

The value of the game: Minimax Regret

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

Online Prediction Games

The value of the game: Minimax Regret

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

Strategy:

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

Online Prediction Games

The value of the game: Minimax Regret

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

Strategy:

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

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

Online Prediction Games

The value of the game: Minimax Regret

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

Minimax Optimal Strategy:

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

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

Online Prediction Games

Questions

Online Prediction Games

Questions

- Minimax regret?

Online Prediction Games

Questions

- Minimax regret?

Online Prediction Games

Questions

- Minimax regret?
- Optimal player's strategy?

Online Prediction Games

Questions

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

Online Prediction Games

Questions

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

Online Prediction Games

Questions

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

Online Prediction Games

Questions

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

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

1. $\frac{1}{2}\|a - y\|_2^2$,
   $a, y \in \mathbb{R}^d$.

Online Prediction Games

Questions

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

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

1. $\frac{1}{2}\|a - y\|_2^2$, $a, y \in \mathbb{R}^d$.
2. $\frac{1}{2}(a - y)^\top W(a - y)$, $W \succeq 0$.

Online Prediction Games

Questions

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

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

1. $\frac{1}{2}\|a - y\|_2^2$,
   $a, y \in \mathbb{R}^d$.
2. $\frac{1}{2}(a - y)^\top W(a - y)$,
   $W \succeq 0$.
3. $-\log a(y)$,
   $a \in \{p_\theta : \theta \in \Theta\}$.

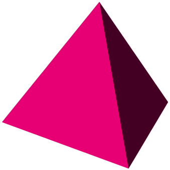
Online Prediction Games

Questions

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

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

1. $\frac{1}{2}\|a - y\|_2^2$,
   $a, y \in \mathbb{R}^d$.
2. $\frac{1}{2}(a - y)^\top W(a - y)$,
   $W \succeq 0$.
3. $-\log a(y)$,
   $a \in \{p_\theta : \theta \in \Theta\}$.



wikimedia.org

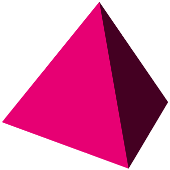
Online Prediction Games

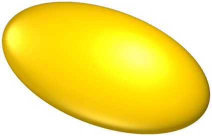
Questions

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

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

1. $\frac{1}{2}\|a - y\|_2^2$, $a, y \in \mathbb{R}^d$.
2. $\frac{1}{2}(a - y)^\top W(a - y)$, $W \succeq 0$.
3. $-\log a(y)$, $a \in \{p_\theta : \theta \in \Theta\}$.





wikimedia.org
mfo.de

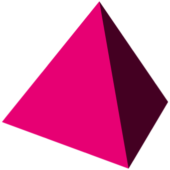
Online Prediction Games

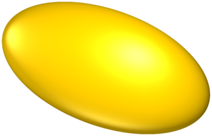
Questions

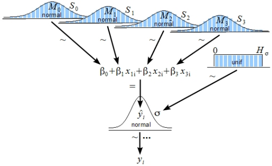
- Minimax regret?
- Optimal player's strategy?
- Efficiently computable?
- Optimal adversary's strategy?
- How do they depend on ℓ, $\mathcal{Y}$, $\mathcal{A}$?

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

1. $\frac{1}{2}\|a - y\|_2^2$,
   $a, y \in \mathbb{R}^d$.
2. $\frac{1}{2}(a - y)^\top W (a - y)$,
   $W \succeq 0$.
3. $-\log a(y)$,
   $a \in \{p_\theta : \theta \in \Theta\}$.







wikimedia.org
mfo.de
www.indiana.edu

Outline

Outline

- Computing minimax optimal strategies.

Outline

- Computing minimax optimal strategies.
- Prediction games with simple minimax optimal strategies.

Outline

- Computing minimax optimal strategies.
- Prediction games with simple minimax optimal strategies.
- Part 1: Log loss.

Outline

- Computing minimax optimal strategies.
- Prediction games with simple minimax optimal strategies.
- Part 1: Log loss.
  - Normalized maximum likelihood.

Outline

- Computing minimax optimal strategies.
- Prediction games with simple minimax optimal strategies.
- Part 1: Log loss.
  - Normalized maximum likelihood.
  - SNML: predicting like there's no tomorrow.

Outline

- Computing minimax optimal strategies.
- Prediction games with simple minimax optimal strategies.
- Part 1: Log loss.
    - Normalized maximum likelihood.
    - SNML: predicting like there's no tomorrow.
    - Bayesian strategies.

Outline

- Computing minimax optimal strategies.
- Prediction games with simple minimax optimal strategies.
- Part 1: Log loss.
    - Normalized maximum likelihood.
    - SNML: predicting like there's no tomorrow.
    - Bayesian strategies.
    - Optimality = exchangeability.

Outline

- Computing minimax optimal strategies.
- Prediction games with simple minimax optimal strategies.
- Part 1: Log loss.
  - Normalized maximum likelihood.
  - SNML: predicting like there's no tomorrow.
  - Bayesian strategies.
  - Optimality = exchangeability.
- Part 2: Euclidean loss.

Outline

- Computing minimax optimal strategies.
- Prediction games with simple minimax optimal strategies.
- Part 1: Log loss.
    - Normalized maximum likelihood.
    - SNML: predicting like there's no tomorrow.
    - Bayesian strategies.
    - Optimality = exchangeability.
- Part 2: Euclidean loss.
    - The role of the smallest ball.

Outline

- Computing minimax optimal strategies.
- Prediction games with simple minimax optimal strategies.
- Part 1: Log loss.
    - Normalized maximum likelihood.
    - SNML: predicting like there's no tomorrow.
    - Bayesian strategies.
    - Optimality = exchangeability.
- Part 2: Euclidean loss.
    - The role of the smallest ball.
    - The simplex and the ball.

Outline

- Computing minimax optimal strategies.
- Prediction games with simple minimax optimal strategies.
- Part 1: Log loss.
  - Normalized maximum likelihood.
  - SNML: predicting like there's no tomorrow.
  - Bayesian strategies.
  - Optimality = exchangeability.
- Part 2: Euclidean loss.
  - The role of the smallest ball.
  - The simplex and the ball.
  - Sub-game optimal strategies on ellipsoids.

Outline

- **Computing minimax optimal strategies.**
- Prediction games with simple minimax optimal strategies.
- Part 1: Log loss.
  - Normalized maximum likelihood.
  - SNML: predicting like there's no tomorrow.
  - Bayesian strategies.
  - Optimality = exchangeability.
- Part 2: Euclidean loss.
  - The role of the smallest ball.
  - The simplex and the ball.
  - Sub-game optimal strategies on ellipsoids.

Computing minimax optimal strategies

Computing minimax optimal strategies

The value of the game:

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

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

Computing minimax optimal strategies

The value of the game:

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

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:

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

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:

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

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

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 $V(y_1, \ldots, y_t)$,

Computing minimax optimal strategies

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

1. Compute $V(y_1, \ldots, y_t)$,
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 $V(y_1, \ldots, y_t)$,
2. Choose a_t as the minimizer of

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

Difficult!

Computing minimax optimal strategies

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

1. Compute $V(y_1, \ldots, y_t)$,
2. Choose a_t as the minimizer of

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

Difficult!

Efficient minimax optimal strategies

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

Games with simple minimax optimal strategies

Prediction Game	**Efficient** optimal strategy?

Games with simple minimax optimal strategies

Prediction Game	Efficient optimal strategy?
Log loss	

- Log loss: $\ell(\hat{p}, y) = -\log \hat{p}(y)$. ($\hat{p}$ a density; $\mathcal{C}$ a probability model.)

Games with simple minimax optimal strategies

Prediction Game	**Efficient** optimal strategy?
Log loss	

- Log loss: $\ell(\hat{p}, y) = -\log \hat{p}(y)$. ($\hat{p}$ a density; $\mathcal{C}$ a probability model.)
- Minimax optimal strategy: normalized maximum likelihood. [Shtarkov, 1987]

Games with simple minimax optimal strategies

Prediction Game	**Efficient** optimal strategy?
Log loss	some cases

- Log loss: $\ell(\hat{p}, y) = -\log \hat{p}(y)$. ($\hat{p}$ a density; $\mathcal{C}$ a probability model.)
- Minimax optimal strategy: normalized maximum likelihood. [Shtarkov, 1987]
- Computation difficult in general. Efficient special cases:
  - Multinomials [Kontkanen, Myllymäki, 2005]

Games with simple minimax optimal strategies

Prediction Game	**Efficient** optimal strategy?
Log loss	some cases

This talk:

- Log loss: $\ell(\hat{p}, y) = -\log \hat{p}(y)$. ($\hat{p}$ a density; $\mathcal{C}$ a probability model.)
- Minimax optimal strategy: normalized maximum likelihood. [Shtarkov, 1987]
- When are simpler strategies optimal?

Games with simple minimax optimal strategies

Prediction Game	**Efficient** optimal strategy?
Log loss	some cases ✓

This talk:

- Log loss: $\ell(\hat{p}, y) = -\log \hat{p}(y)$. ($\hat{p}$ a density; $\mathcal{C}$ a probability model.)
- Minimax optimal strategy: normalized maximum likelihood. [Shtarkov, 1987]
- When are simpler strategies optimal?
    - Sequential NML.
    - Bayesian prediction.

Games with simple minimax optimal strategies

Prediction Game	**Efficient** optimal strategy?
Log loss	some cases ✓
Absolute loss, binary	

- $\mathcal{Y} = \{0,1\}$, $\mathcal{A} = [0,1]$, $\ell(a,y) = |a-y|$. (Also $\mathcal{C} \subset$ *static* experts.)

Games with simple minimax optimal strategies

Prediction Game	**Efficient** optimal strategy?
Log loss	some cases ✓
Absolute loss, binary	

- $\mathcal{Y} = \{0,1\}$, $\mathcal{A} = [0,1]$, $\ell(a,y) = |a - y|$. (Also $\mathcal{C} \subset$ *static* experts.)
- Minimax optimal strategy: compare expected minimal cumulative loss for random futures.

[Cover, 1967], [Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, Warmuth, 1997], [Cesa-Bianchi, Shamir, 2011], [Koolen, 2011], [Gravin, Peres, Sivan, 2014]

Games with simple minimax optimal strategies

Prediction Game	**Efficient** optimal strategy?
Log loss	some cases ✓
Absolute loss, binary	can be approximated

- $\mathcal{Y} = \{0, 1\}$, $\mathcal{A} = [0, 1]$, $\ell(a, y) = |a - y|$. (Also $\mathcal{C} \subset$ *static* experts.)
- Minimax optimal strategy: compare expected minimal cumulative loss for random futures.

[Cover, 1967], [Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, Warmuth, 1997],

[Cesa-Bianchi, Shamir, 2011], [Koolen, 2011], [Gravin, Peres, Sivan, 2014]

Games with simple minimax optimal strategies

Prediction Game	**Efficient** optimal strategy?
Log loss	some cases ✓
Absolute loss, binary	can be approximated
Experts, bounded loss	

- $\mathcal{Y} = \Delta$, linear loss, best cumulative loss is bounded.

Games with simple minimax optimal strategies

Prediction Game	**Efficient** optimal strategy?
Log loss	some cases ✓
Absolute loss, binary	can be approximated
Experts, bounded loss	

- $\mathcal{Y} = \Delta$, linear loss, best cumulative loss is bounded.
- Minimax optimal strategy: estimate survival probability.

[Abernethy, Warmuth, Yellin, 2008]

Games with simple minimax optimal strategies

Prediction Game	**Efficient** optimal strategy?
Log loss	some cases ✓
Absolute loss, binary	can be approximated
Experts, bounded loss	can be approximated

- $\mathcal{Y} = \Delta$, linear loss, best cumulative loss is bounded.
- Minimax optimal strategy: estimate survival probability.

[Abernethy, Warmuth, Yellin, 2008]

Games with simple minimax optimal strategies

Prediction Game	**Efficient** optimal strategy?
Log loss	some cases ✓
Absolute loss, binary	can be approximated
Experts, bounded loss	can be approximated
Quadratic loss	

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

Games with simple minimax optimal strategies

Prediction Game	**Efficient** optimal strategy?
Log loss	some cases ✓
Absolute loss, binary	can be approximated
Experts, bounded loss	can be approximated
Quadratic loss	unit ball

- $\ell(a, y) = \frac{1}{2}\|a - y\|^2$,
- $\mathcal{Y}$ =unit ball.

[Takimoto, Warmuth, 2000]

Games with simple minimax optimal strategies

Prediction Game	**Efficient** optimal strategy?
Log loss	some cases ✓
Absolute loss, binary	can be approximated
Experts, bounded loss	can be approximated
Quadratic loss	unit ball
Quadratic/Mahalanobis loss	

This talk:

- $\ell(a, y) = \frac{1}{2}(a - y)^\top W(a - y)$, for $W \succeq 0$.
- $\mathcal{Y}$ =compact set, $\mathcal{A} \supseteq \mathrm{co}(\mathcal{Y})$.

Games with simple minimax optimal strategies

Prediction Game	**Efficient** optimal strategy?
Log loss	some cases ✓
Absolute loss, binary	can be approximated
Experts, bounded loss	can be approximated
Quadratic loss	unit ball
Quadratic/Mahalanobis loss	✓

This talk:

- $\ell(a, y) = \frac{1}{2}(a - y)^\top W(a - y)$, for $W \succeq 0$.
- $\mathcal{Y}$ =compact set, $\mathcal{A} \supseteq \mathrm{co}(\mathcal{Y})$.
- Efficient minimax optimal strategy.

Outline

- Computing minimax optimal strategies.
- Prediction games with simple minimax optimal strategies.
- **Part 1: Log loss.**
    - Normalized maximum likelihood.
    - SNML: predicting like there's no tomorrow.
    - Bayesian strategies.
    - Optimality = exchangeability.
- Part 2: Euclidean loss.
    - The role of the smallest ball.
    - The simplex and the ball.
    - Sub-game optimal strategies on ellipsoids.

Online density estimation with log loss

Online density estimation with log loss

Log loss

$$\ell(\hat{p}, y) = -\log \hat{p}(y).$$

Online density estimation with log loss

Comparison class

Parametric family of densities: $\mathcal{C} = \{p_\theta : \theta \in \Theta\}$, where $p_\theta : \mathcal{Y} \to \mathbb{R}^+$ is a parameterized probability density with respect to a reference measure λ on $\mathcal{Y}$.

Log loss

$$\ell(\hat{p}, y) = -\log \hat{p}(y).$$

Online density estimation with log loss

Comparison class

Parametric family of densities: $\mathcal{C} = \{p_\theta : \theta \in \Theta\}$, where $p_\theta : \mathcal{Y} \to \mathbb{R}^+$ is a parameterized probability density with respect to a reference measure λ on $\mathcal{Y}$.

Log loss

$$\ell(\hat{p}, y) = -\log \hat{p}(y).$$

Regret

Online density estimation with log loss

Comparison class

Parametric family of densities: $\mathcal{C} = \{p_\theta : \theta \in \Theta\}$, where $p_\theta : \mathcal{Y} \to \mathbb{R}^+$ is a parameterized probability density with respect to a reference measure λ on $\mathcal{Y}$.

Log loss

$$\ell(\hat{p}, y) = -\log \hat{p}(y).$$

Regret

$$R(y_1^n, \hat{p}) =$$

Online density estimation with log loss

Comparison class

Parametric family of densities: $\mathcal{C} = \{p_\theta : \theta \in \Theta\}$, where $p_\theta : \mathcal{Y} \to \mathbb{R}^+$ is a parameterized probability density with respect to a reference measure λ on $\mathcal{Y}$.

Log loss

$$\ell(\hat{p}, y) = -\log \hat{p}(y).$$

Regret

$$R(y_1^n, \hat{p}) = \sum_{t=1}^{n} \ell(\hat{p}_t, y_t) -$$

Online density estimation with log loss

Comparison class

Parametric family of densities: $\mathcal{C} = \{p_\theta : \theta \in \Theta\}$, where $p_\theta : \mathcal{Y} \to \mathbb{R}^+$ is a parameterized probability density with respect to a reference measure λ on $\mathcal{Y}$.

Log loss

$$\ell(\hat{p}, y) = -\log \hat{p}(y).$$

Regret

$$R(y_1^n, \hat{p}) = \sum_{t=1}^{n} \ell(\hat{p}_t, y_t) - \inf_{p \in \mathcal{C}} \sum_{t=1}^{n} \ell(p, y_t).$$

Online density estimation with log loss

Strategies are joint densities

Online density estimation with log loss

Strategies are joint densities

- A strategy $\hat{p}$ is a mapping from histories $y_1^t = (y_1, \ldots, y_t)$ to densities $\hat{p}(\cdot|y_1^t)$ on $\mathcal{Y}$.

Online density estimation with log loss

Strategies are joint densities

- A strategy $\hat{p}$ is a mapping from histories $y_1^t = (y_1, \ldots, y_t)$ to densities $\hat{p}(\cdot | y_1^t)$ on $\mathcal{Y}$.
- Every strategy is a joint density:

$$\hat{p}(y_1, \ldots, y_n) =$$

Online density estimation with log loss

Strategies are joint densities

- A strategy $\hat{p}$ is a mapping from histories $y_1^t = (y_1, \ldots, y_t)$ to densities $\hat{p}(\cdot|y_1^t)$ on $\mathcal{Y}$.
- Every strategy is a joint density:

$$\hat{p}(y_1, \ldots, y_n) = \hat{p}(y_1)$$

Online density estimation with log loss

Strategies are joint densities

- A strategy $\hat{p}$ is a mapping from histories $y_1^t = (y_1, \ldots, y_t)$ to densities $\hat{p}(\cdot|y_1^t)$ on $\mathcal{Y}$.
- Every strategy is a joint density:

$$\hat{p}(y_1, \ldots, y_n) = \hat{p}(y_1)\hat{p}(y_2|y_1)$$

Online density estimation with log loss

Strategies are joint densities

- A strategy $\hat{p}$ is a mapping from histories $y_1^t = (y_1, \ldots, y_t)$ to densities $\hat{p}(\cdot|y_1^t)$ on $\mathcal{Y}$.
- Every strategy is a joint density:

$$\hat{p}(y_1, \ldots, y_n) = \hat{p}(y_1)\hat{p}(y_2|y_1)\cdots\hat{p}(y_n|y_1^{n-1}).$$

Online density estimation with log loss

Strategies are joint densities

- A strategy $\hat{p}$ is a mapping from histories $y_1^t = (y_1, \ldots, y_t)$ to densities $\hat{p}(\cdot|y_1^t)$ on $\mathcal{Y}$.
- Every strategy is a joint density:

$$\hat{p}(y_1, \ldots, y_n) = \hat{p}(y_1)\hat{p}(y_2|y_1) \cdots \hat{p}(y_n|y_1^{n-1}).$$

- Regret wrt comparison $\mathcal{C} = \{p_\theta\}$ is log likelihood ratio:

$$R(y_1^n, \hat{p}) = \sum_{t=1}^{n} \ell(\hat{p}_t, y_t) - \inf_{p \in \mathcal{C}} \sum_{t=1}^{n} \ell(p, y_t)$$

Online density estimation with log loss

Strategies are joint densities

- A strategy $\hat{p}$ is a mapping from histories $y_1^t = (y_1, \ldots, y_t)$ to densities $\hat{p}(\cdot|y_1^t)$ on $\mathcal{Y}$.
- Every strategy is a joint density:
$$\hat{p}(y_1, \ldots, y_n) = \hat{p}(y_1)\hat{p}(y_2|y_1)\cdots\hat{p}(y_n|y_1^{n-1}).$$
- Regret wrt comparison $\mathcal{C} = \{p_\theta\}$ is log likelihood ratio:
$$\begin{aligned} R(y_1^n, \hat{p}) &= \sum_{t=1}^{n} \ell(\hat{p}_t, y_t) - \inf_{p \in \mathcal{C}} \sum_{t=1}^{n} \ell(p, y_t) \\ &= \sup_{\theta \in \Theta} \log p_\theta(y_1^n) - \log \hat{p}(y_1^n). \end{aligned}$$

Online density estimation with log loss

Strategies are joint densities

- A strategy $\hat{p}$ is a mapping from histories $y_1^t = (y_1, \ldots, y_t)$ to densities $\hat{p}(\cdot|y_1^t)$ on $\mathcal{Y}$.
- Every strategy is a joint density:
$$\hat{p}(y_1, \ldots, y_n) = \hat{p}(y_1)\hat{p}(y_2|y_1) \cdots \hat{p}(y_n|y_1^{n-1}).$$
- Regret wrt comparison $\mathcal{C} = \{p_\theta\}$ is log likelihood ratio:
$$\begin{aligned} R(y_1^n, \hat{p}) &= \sum_{t=1}^n \ell(\hat{p}_t, y_t) - \inf_{p \in \mathcal{C}} \sum_{t=1}^n \ell(p, y_t) \\ &= \sup_{\theta \in \Theta} \log p_\theta(y_1^n) - \log \hat{p}(y_1^n). \end{aligned}$$

Here, $p_\theta(y_1^n) = \prod_{t=1}^n p_\theta(y_t)$.

Online density estimation with log loss

Many interpretations of prediction with log loss

Online density estimation with log loss

Many interpretations of prediction with log loss

- Sequential probability prediction.

Online density estimation with log loss

Many interpretations of prediction with log loss

- Sequential probability prediction.
- Sequential lossless data compression.

Online density estimation with log loss

Many interpretations of prediction with log loss

- Sequential probability prediction.
- Sequential lossless data compression.
- Repeated gambling/investment.

Online density estimation with log loss

Many interpretations of prediction with log loss

- Sequential probability prediction.
- Sequential lossless data compression.
- Repeated gambling/investment.

Long history in several communities.

[Kelly, 1956], [Solomonoff, 1964], [Kolmogorov, 1965], [Cover, 1974], [Rissanen, 1976, 1987, 1996], [Shtarkov, 1987], [Feder, Merhav and Gutman, 1992], [Freund, 1996], [Xie and Barron, 2000], [Cesa-Bianchi and Lugosi, 2001, 2006], [Grünwald, 2007]

Normalized maximum likelihood

Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1^n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n).$$

Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1^n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n).$$

NML is optimal [Shtarkov, 1987]

Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1^n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n).$$

NML is optimal [Shtarkov, 1987]

1. NML equalizes regret: for any sequence y_1^n, regret is

$$\log \int_{Y^n} \sup_{\theta \in \Theta} p_\theta(z^n)\, d\lambda^n(z^n).$$

Normalized maximum likelihood

NML

$$p^{(n)}_{nml}(y_1^n) \propto \sup_{\theta\in\Theta} p_\theta(y_1^n).$$

NML is optimal [Shtarkov, 1987]

1. NML equalizes regret: for any sequence y_1^n, regret is

$$\log \int_{Y^n} \sup_{\theta\in\Theta} p_\theta(z^n)\, d\lambda^n(z^n).$$

2. Any strategy that does not equalize regret has strictly worse maximum regret.

Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1^n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1^n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

- To predict, we compute conditional distributions, marginalize.

Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1^n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

$$p_{nml}^{(n)}(y_t | y_1^{t-1}) = \frac{\int_{\mathcal{Y}^{n-t}} \sup_{\theta\in\Theta} p_\theta(y_1^t z_{t+1}^n)\, d\lambda^{n-t}(z_{t+1}^n)}{\int_{\mathcal{Y}^{n-t+1}} \sup_{\theta\in\Theta} p_\theta(y_1^{t-1} z_t^n)\, d\lambda^{n-t+1}(z_t^n)}$$

- To predict, we compute conditional distributions, marginalize.

Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1^n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

$$p_{nml}^{(n)}(y_t | y_1^{t-1}) = \frac{\int_{\mathcal{Y}^{n-t}} \sup_{\theta \in \Theta} p_\theta(y_1^t z_{t+1}^n)\, d\lambda^{n-t}(z_{t+1}^n)}{\int_{\mathcal{Y}^{n-t+1}} \sup_{\theta \in \Theta} p_\theta(y_1^{t-1} z_t^n)\, d\lambda^{n-t+1}(z_t^n)}$$

- To predict, we compute conditional distributions, marginalize.

Normalized maximum likelihood

NML

$$p^{(n)}_{nml}(y_1^n) \propto \sup_{\theta\in\Theta} p_\theta(y_1^n)$$

$$p^{(n)}_{nml}(y_t|y_1^{t-1}) = \frac{\int_{\mathcal{Y}^{n-t}} \sup_{\theta\in\Theta} p_\theta(y_1^t z_{t+1}^n)\, d\lambda^{n-t}(z_{t+1}^n)}{\int_{\mathcal{Y}^{n-t+1}} \sup_{\theta\in\Theta} p_\theta(y_1^{t-1} z_t^n)\, d\lambda^{n-t+1}(z_t^n)}$$

- To predict, we compute conditional distributions, marginalize.
- All that conditioning is computationally expensive!

Normalized maximum likelihood

NML

$$p^{(n)}_{nml}(y_1^n) \propto \sup_{\theta\in\Theta} p_\theta(y_1^n)$$

$$p^{(n)}_{nml}(y_t|y_1^{t-1}) = \frac{\int_{\mathcal{Y}^{n-t}} \sup_{\theta\in\Theta} p_\theta(y_1^t z_{t+1}^n)\, d\lambda^{n-t}(z_{t+1}^n)}{\int_{\mathcal{Y}^{n-t+1}} \sup_{\theta\in\Theta} p_\theta(y_1^{t-1} z_t^n)\, d\lambda^{n-t+1}(z_t^n)}$$

- To predict, we compute conditional distributions, marginalize.
- All that conditioning is computationally expensive!
- When is a computationally cheaper strategy optimal?

Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1^n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

$$p_{nml}^{(n)}(y_t | y_1^{t-1}) = \frac{\int_{\mathcal{Y}^{n-t}} \sup_{\theta \in \Theta} p_\theta(y_1^t z_{t+1}^n)\, d\lambda^{n-t}(z_{t+1}^n)}{\int_{\mathcal{Y}^{n-t+1}} \sup_{\theta \in \Theta} p_\theta(y_1^{t-1} z_t^n)\, d\lambda^{n-t+1}(z_t^n)}$$

- To predict, we compute conditional distributions, marginalize.
- All that conditioning is computationally expensive!
- When is a computationally cheaper strategy optimal?
    - Horizon-independent NML?

Normalized maximum likelihood

NML

$$p^{(n)}_{nml}(y_1^n) \propto \sup_{\theta\in\Theta} p_\theta(y_1^n)$$

$$p^{(n)}_{nml}(y_t|y_1^{t-1}) = \frac{\int_{\mathcal{Y}^{n-t}} \sup_{\theta\in\Theta} p_\theta(y_1^t z_{t+1}^n)\, d\lambda^{n-t}(z_{t+1}^n)}{\int_{\mathcal{Y}^{n-t+1}} \sup_{\theta\in\Theta} p_\theta(y_1^{t-1} z_t^n)\, d\lambda^{n-t+1}(z_t^n)}$$

- To predict, we compute conditional distributions, marginalize.
- All that conditioning is computationally expensive!
- When is a computationally cheaper strategy optimal?
  - Horizon-independent NML?
  - Bayesian prediction?

Outline

- Computing minimax optimal strategies.
- Prediction games with simple minimax optimal strategies.
- Part 1: Log loss.
  - Normalized maximum likelihood.
  - **SNML: predicting like there's no tomorrow.**
  - Bayesian strategies.
  - Optimality = exchangeability.
- Part 2: Euclidean loss.
  - The role of the smallest ball.
  - The simplex and the ball.
  - Sub-game optimal strategies on ellipsoids.

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood

- Pretend that this is the last prediction we'll ever make.

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood

$$p_{snml}(y_t | y_1^{t-1}) = p_{nml}^{(t)}(y_t | y_1^{t-1})$$

- Pretend that this is the last prediction we'll ever make.

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood

$$p_{snml}(y_t|y_1^{t-1}) = p_{nml}^{(t)}(y_t|y_1^{t-1}) \propto \sup_{\theta\in\Theta} p_\theta(y_1^t)$$

- Pretend that this is the last prediction we'll ever make.

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood

$$p_{snml}(y_t|y_1^{t-1}) = p_{nml}^{(t)}(y_t|y_1^{t-1}) \propto \sup_{\theta\in\Theta} p_\theta(y_1^t)$$

- Pretend that this is the last prediction we'll ever make.
- Simpler conditional calculation.

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood

$$p_{snml}(y_t | y_1^{t-1}) = p_{nml}^{(t)}(y_t | y_1^{t-1}) \propto \sup_{\theta \in \Theta} p_\theta(y_1^t)$$

- Pretend that this is the last prediction we'll ever make.
- Simpler conditional calculation.
- Known to have asymptotically optimal regret.

[Takimoto and Warmuth, 2000], [Roos and Rissanen, 2008], [Kotłowski and Grünwald, 2011]

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood

$$p_{snml}(y_t | y_1^{t-1}) = p_{nml}^{(t)}(y_t | y_1^{t-1}) \propto \sup_{\theta \in \Theta} p_\theta(y_1^t)$$

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood

$$p_{snml}(y_t | y_1^{t-1}) = p_{nml}^{(t)}(y_t | y_1^{t-1}) \propto \sup_{\theta \in \Theta} p_\theta(y_1^t)$$

Theorem

Sequential NML is optimal iff p_{snml} is exchangeable.

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood

$$p_{snml}(y_t|y_1^{t-1}) = p_{nml}^{(t)}(y_t|y_1^{t-1}) \propto \sup_{\theta\in\Theta} p_\theta(y_1^t)$$

Theorem

Sequential NML is optimal iff p_{snml} is exchangeable.

- p_{snml} is exchangeable means
  $p_{snml}(y_1, y_2, y_3, y_4) = p_{snml}(y_1, y_2, y_4, y_3) = \cdots = p_{snml}(y_4, y_3, y_2, y_1)$.

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood

$$p_{snml}(y_t|y_1^{t-1}) = p_{nml}^{(t)}(y_t|y_1^{t-1}) \propto \sup_{\theta \in \Theta} p_\theta(y_1^t)$$

Theorem

Sequential NML is optimal iff p_{snml} is exchangeable.

Proof idea:

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood

$$p_{snml}(y_t|y_1^{t-1}) = p_{nml}^{(t)}(y_t|y_1^{t-1}) \propto \sup_{\theta\in\Theta} p_\theta(y_1^t)$$

Theorem

Sequential NML is optimal iff p_{snml} is exchangeable.

Proof idea:

- SNML's regret doesn't depend on last observation.

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood

$$p_{snml}(y_t|y_1^{t-1}) = p_{nml}^{(t)}(y_t|y_1^{t-1}) \propto \sup_{\theta\in\Theta} p_\theta(y_1^t)$$

Theorem

Sequential NML is optimal iff p_{snml} is exchangeable.

Proof idea:

- SNML's regret doesn't depend on last observation.
- $(\Leftarrow)$ Exchangeability implies regret is independent of observations. Hence SNML is an equalizer: same as NML.

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood

$$p_{snml}(y_t|y_1^{t-1}) = p_{nml}^{(t)}(y_t|y_1^{t-1}) \propto \sup_{\theta\in\Theta} p_\theta(y_1^t)$$

Theorem

Sequential NML is optimal iff p_{snml} is exchangeable.

Proof idea:

- SNML's regret doesn't depend on last observation.
- $(\Leftarrow)$ Exchangeability implies regret is independent of observations. Hence SNML is an equalizer: same as NML.
- $(\Rightarrow)$ $p_{nml}^{(n)}(y_1^n)$ is permutation-invariant.

Outline

- Computing minimax optimal strategies.
- Prediction games with simple minimax optimal strategies.
- Part 1: Log loss.
  - Normalized maximum likelihood.
  - SNML: predicting like there's no tomorrow.
  - **Bayesian strategies.**
  - Optimality = exchangeability.
- Part 2: Euclidean loss.
  - The role of the smallest ball.
  - The simplex and the ball.
  - Sub-game optimal strategies on ellipsoids.

Bayesian strategies

Bayesian strategies

Bayesian strategies

For prior π on Θ:

$$p_\pi(x_1^t) = \int_{\theta \in \Theta} p_\theta(x_1^t)\, d\pi(\theta)$$

Bayesian strategies

For prior π on Θ:

$$p_\pi(x_1^t) = \int_{\theta \in \Theta} p_\theta(x_1^t)\, d\pi(\theta)$$

- Sequential update to prior.

Bayesian strategies

Bayesian strategies

For prior π on Θ:

$$p_\pi(x_1^t) = \int_{\theta \in \Theta} p_\theta(x_1^t)\, d\pi(\theta)$$

$$p_\pi(\theta | x_1^t) \propto p_\pi(\theta | x_1^{t-1}) p_\theta(x_t).$$

- Sequential update to prior.

Bayesian strategies

For prior π on Θ:

$$p_\pi(x_1^t) = \int_{\theta\in\Theta} p_\theta(x_1^t)\, d\pi(\theta)$$

$$p_\pi(\theta|x_1^t) \propto p_\pi(\theta|x_1^{t-1}) p_\theta(x_t).$$

- Sequential update to prior.
- Jeffreys prior:

$$\pi(\theta) \propto \sqrt{|I(\theta)|},$$

Bayesian strategies

Bayesian strategies

For prior π on Θ:

$$p_\pi(x_1^t) = \int_{\theta\in\Theta} p_\theta(x_1^t)\, d\pi(\theta)$$

$$p_\pi(\theta|x_1^t) \propto p_\pi(\theta|x_1^{t-1}) p_\theta(x_t).$$

- Sequential update to prior.
- Jeffreys prior:

$$\pi(\theta) \propto \sqrt{|I(\theta)|},$$

- Attractive properties (e.g., invariant to parameterization).

Bayesian strategies

Bayesian strategies

For prior π on Θ:

$$p_\pi(x_1^t) = \int_{\theta\in\Theta} p_\theta(x_1^t)\, d\pi(\theta)$$

$$p_\pi(\theta|x_1^t) \propto p_\pi(\theta|x_1^{t-1}) p_\theta(x_t).$$

- Sequential update to prior.
- Jeffreys prior:

$$\pi(\theta) \propto \sqrt{|I(\theta)|},$$

- Attractive properties (e.g., invariant to parameterization).
- Asymptotically optimal regret for exponential families.

Sequential NML and Bayesian strategies

Optimality

Sequential NML and Bayesian strategies

Optimality

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

Sequential NML and Bayesian strategies

Optimality

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.

Sequential NML and Bayesian strategies

Optimality

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.

Sequential NML and Bayesian strategies

Optimality

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.
4. NML = Bayesian with Jeffreys prior.

Sequential NML and Bayesian strategies

Optimality

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML $=$ SNML.
2. p_{snml} exchangeable.
3. NML $=$ Bayesian.
4. NML $=$ Bayesian with Jeffreys prior.
5. SNML $=$ Bayesian.

Sequential NML and Bayesian strategies

Optimality

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.
4. NML = Bayesian with Jeffreys prior.
5. SNML = Bayesian.
6. SNML = Bayesian with Jeffreys prior.

Sequential NML and Bayesian strategies

Optimality

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.
4. NML = Bayesian with Jeffreys prior.
5. SNML = Bayesian.
6. SNML = Bayesian with Jeffreys prior.

- If we can ignore the time horizon and be optimal, that's the same as Bayesian prediction with Jeffreys prior.

Sequential NML and Bayesian strategies

Optimality

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML $=$ SNML.
2. p_{snml} exchangeable.
3. NML $=$ Bayesian.
4. NML $=$ Bayesian with Jeffreys prior.
5. SNML $=$ Bayesian.
6. SNML $=$ Bayesian with Jeffreys prior.

- If we can ignore the time horizon and be optimal, that's the same as Bayesian prediction with Jeffreys prior.
- If any Bayesian strategy is optimal, it uses Jeffreys prior.

Sequential NML and Bayesian strategies

Optimality

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.
4. NML = Bayesian with Jeffreys prior.
5. SNML = Bayesian.
6. SNML = Bayesian with Jeffreys prior.

- If we can ignore the time horizon and be optimal, that's the same as Bayesian prediction with Jeffreys prior.
- If any Bayesian strategy is optimal, it uses Jeffreys prior.
- Why?

Sequential NML and Bayesian strategies

Optimality

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.
4. NML = Bayesian with Jeffreys prior.
5. SNML = Bayesian.
6. SNML = Bayesian with Jeffreys prior.

- If we can ignore the time horizon and be optimal, that's the same as Bayesian prediction with Jeffreys prior.
- If any Bayesian strategy is optimal, it uses Jeffreys prior.
- Why? If NML=SNML, then we can consider long time horizons, so the asymptotics emerge.

Sequential NML and Bayesian strategies

Optimality

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.
4. NML = Bayesian with Jeffreys prior.
5. SNML = Bayesian.
6. SNML = Bayesian with Jeffreys prior.

- If we can ignore the time horizon and be optimal, that's the same as Bayesian prediction with Jeffreys prior.
- If any Bayesian strategy is optimal, it uses Jeffreys prior.
- Why? If NML=SNML, then we can consider long time horizons, so the asymptotics emerge. Asymptotic normality of the MLE implies Jeffreys prior is the only candidate.

Online density estimation with log loss

Extensions

[B., Grünwald, Harremoës, Hedayati, Kotłowski, 2013]

Online density estimation with log loss

Extensions

[B., Grünwald, Harremoës, Hedayati, Kotłowski, 2013]

- One-dimensional exponential families:

$$p_\theta(y) = h(y) \exp\left(\theta y - A(\theta)\right).$$

Online density estimation with log loss

Extensions

[B., Grünwald, Harremoës, Hedayati, Kotłowski, 2013]

- One-dimensional exponential families:
$$p_\theta(y) = h(y) \exp\left(\theta y - A(\theta)\right).$$
- p_{SNML} is exchangeable (i.e., SNML optimal, Bayesian optimal) $\Leftrightarrow$

Online density estimation with log loss

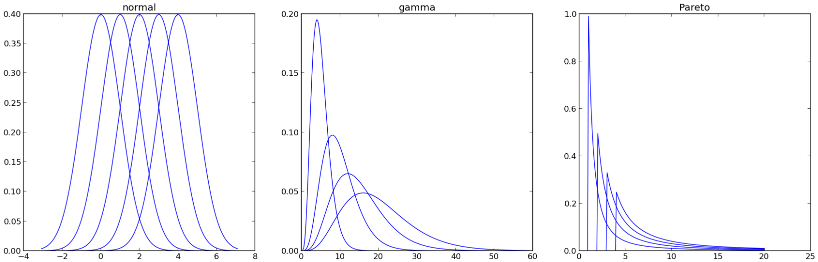
Extensions

[B., Grünwald, Harremoës, Hedayati, Kotłowski, 2013]

- One-dimensional exponential families:

$$p_\theta(y) = h(y) \exp\left(\theta y - A(\theta)\right).$$

- p_{SNML} is exchangeable (i.e., SNML optimal, Bayesian optimal) $\Leftrightarrow$
  1. Gaussian distributions with fixed variance $\sigma^2 > 0$,



Online density estimation with log loss

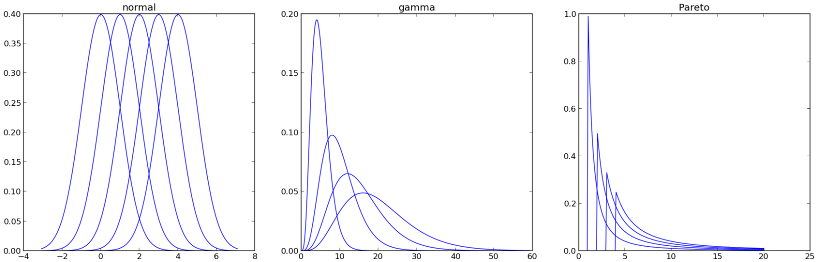
Extensions

[B., Grünwald, Harremoës, Hedayati, Kotłowski, 2013]

- One-dimensional exponential families:

$$p_\theta(y) = h(y) \exp\left(\theta y - A(\theta)\right).$$

- p_{SNML} is exchangeable (i.e., SNML optimal, Bayesian optimal) $\Leftrightarrow$
  1. Gaussian distributions with fixed variance $\sigma^2 > 0$,
  2. gamma distributions with fixed shape $k > 0$,

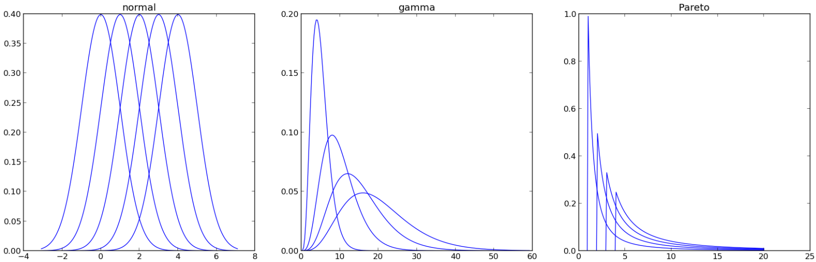


Online density estimation with log loss

Extensions

[B., Grünwald, Harremoës, Hedayati, Kotłowski, 2013]

- One-dimensional exponential families:
$$p_\theta(y) = h(y)\exp\left(\theta y - A(\theta)\right).$$
- p_{SNML} is exchangeable (i.e., SNML optimal, Bayesian optimal) $\Leftrightarrow$
  1. Gaussian distributions with fixed variance $\sigma^2 > 0$,
  2. gamma distributions with fixed shape $k > 0$,
  3. Tweedie exponential family of order $3/2$,

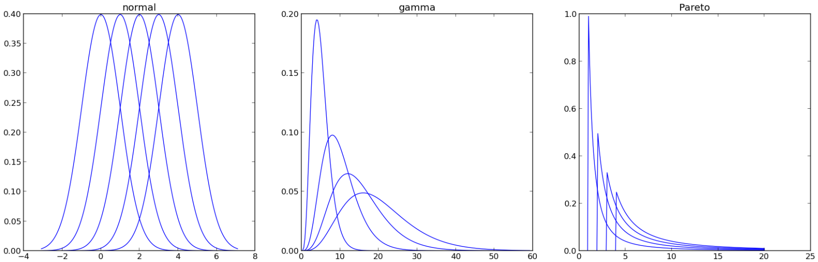


Online density estimation with log loss

Extensions

[B., Grünwald, Harremoës, Hedayati, Kotłowski, 2013]

- One-dimensional exponential families:
$$p_\theta(y) = h(y)\exp\left(\theta y - A(\theta)\right).$$
- p_{SNML} is exchangeable (i.e., SNML optimal, Bayesian optimal) $\Leftrightarrow$
  1. Gaussian distributions with fixed variance $\sigma^2 > 0$,
  2. gamma distributions with fixed shape $k > 0$,
  3. Tweedie exponential family of order $3/2$,
  4. Or smooth transformations
     (Pareto, Laplace, Rayleigh, Lévy, Nakagami)



Outline

- Computing minimax optimal strategies.
- Prediction games with simple minimax optimal strategies.
- Part 1: Log loss.
  - Normalized maximum likelihood.
  - SNML: predicting like there's no tomorrow.
  - Bayesian strategies.
  - Optimality = exchangeability.
- **Part 2: Euclidean loss.**
  - The role of the smallest ball.
  - The simplex and the ball.
  - Sub-game optimal strategies on ellipsoids.

Mahalanobis loss and squared Euclidean loss

The value of the game: minimax regret

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

Mahalanobis loss and squared Euclidean loss

The value of the game: minimax regret

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

$$\mathcal{A} = \mathbb{R}^d,$$

Mahalanobis loss and squared Euclidean loss

The value of the game: minimax regret

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

$$\mathcal{A} = \mathbb{R}^d, \qquad \mathcal{Y} \subset \mathbb{R}^d,$$

Mahalanobis loss and squared Euclidean loss

The value of the game: minimax regret

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

$$\mathcal{A} = \mathbb{R}^d, \qquad \mathcal{Y} \subset \mathbb{R}^d,$$

$$\ell_W(a, y) = \frac{1}{2}(a - y)^\top W (a - y), \qquad W \succeq 0.$$

Mahalanobis loss and squared Euclidean loss

The value of the game: minimax regret

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

$$\mathcal{A} = \mathbb{R}^d, \qquad \mathcal{Y} \subset \mathbb{R}^d,$$

$$\ell_W(a, y) = \frac{1}{2}(a - y)^\top W (a - y), \qquad W \succeq 0.$$

Mahalanobis loss $\rightarrow$ quadratic loss

Since $(a - y)^\top W (a - y) = \left\| W^{1/2}(a - y) \right\|^2$, we can work with $\ell(a, y) = \frac{1}{2}\|a - y\|^2$ and $W^{1/2}\mathcal{Y}$:

$$V_T(\mathcal{Y}, \ell_W) = V_T(W^{1/2}\mathcal{Y}, \ell).$$

Main result: the role of the smallest ball

Main result: the role of the smallest ball

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

Define the 'minimum radius' function:
$J_{\mathcal{Y}}(c) = \max_{y \in \mathcal{Y}} \|y - c\|$,
so the smallest ball containing $\mathcal{Y}$ is
$B_{\mathcal{Y}} = \{y \in \mathbb{R}^d : \|y - c\| \leq r\}$,
with $r = J_{\mathcal{Y}}(c) = \min_x J_{\mathcal{Y}}(x)$.

Main result: the role of the smallest ball

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

Define the 'minimum radius' function:
$J_{\mathcal{Y}}(c) = \max_{y \in \mathcal{Y}} \|y - c\|$,
so the smallest ball containing $\mathcal{Y}$ is
$B_{\mathcal{Y}} = \{y \in \mathbb{R}^d : \|y - c\| \leq r\}$,
with $r = J_{\mathcal{Y}}(c) = \min_x J_{\mathcal{Y}}(x)$.

Main Theorem

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

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

Main result: the role of the smallest ball

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

Define the 'minimum radius' function:
$J_{\mathcal{Y}}(c) = \max_{y \in \mathcal{Y}} \|y - c\|$,
so the smallest ball containing $\mathcal{Y}$ is
$B_{\mathcal{Y}} = \{y \in \mathbb{R}^d : \|y - c\| \leq r\}$,
with $r = J_{\mathcal{Y}}(c) = \min_x J_{\mathcal{Y}}(x)$.

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

The simplex case

Consider 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

Consider a set of $d+1$ affinely independent points in $\mathbb{R}^d$, all lying on the surface of the smallest ball.
Use sufficient 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

The simplex case

Consider a set of $d+1$ affinely independent points in $\mathbb{R}^d$, all lying on the surface of the smallest ball.
Use sufficient 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

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

Online prediction with quadratic loss

The simplex case

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

Value-to-go: quadratic in state

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

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

Online prediction with quadratic loss

The simplex case

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

Value-to-go: quadratic in state

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

Minimax strategy: affine in state

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

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

Online prediction with quadratic loss

The simplex case

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

Value-to-go: quadratic in state

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

Minimax strategy: affine in state

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

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

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

Online prediction with quadratic loss

The simplex case

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

Value-to-go: quadratic in state

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

Minimax strategy: affine in state

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

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

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

Online prediction with quadratic loss

The simplex case

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

Value-to-go: quadratic in state

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

Minimax strategy: affine in state

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

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

Maximin distribution: same mean.

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

Online prediction with quadratic loss

Value-to-go: quadratic in state

$$V(y_1, \ldots, y_n) = \frac{1}{2}\left(\alpha_n \|s_n\|^2 - \sigma_n^2 + r^2 \sum_{t=n+1}^{T} \alpha_t\right).$$

where:

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

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

Online prediction with quadratic loss

Value-to-go: quadratic in state

$$V(y_1, \ldots, y_n) = \frac{1}{2}\left(\alpha_n \|s_n\|^2 - \sigma_n^2 + r^2 \sum_{t=n+1}^{T} \alpha_t\right).$$

where:

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

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

Minimax regret for simplex

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

Online prediction with quadratic loss

Value-to-go: quadratic in state

$$V(y_1, \ldots, y_n) = \frac{1}{2}\left(\alpha_n \|s_n\|^2 - \sigma_n^2 + r^2 \sum_{t=n+1}^{T} \alpha_t\right).$$

where:

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

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

Minimax regret for simplex

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

Online prediction with quadratic loss

Value-to-go: quadratic in state

$$V(y_1, \ldots, y_n) = \frac{1}{2}\left(\alpha_n \|s_n\|^2 - \sigma_n^2 + r^2 \sum_{t=n+1}^{T} \alpha_t\right).$$

where:

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

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

Minimax regret for simplex

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

Online prediction with quadratic loss on the simplex

Proof idea

Online prediction with quadratic loss on the simplex

Proof idea

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

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

Online prediction with quadratic loss on the simplex

Proof idea

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

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

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

Online prediction with quadratic loss on the simplex

Proof idea

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

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

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

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

Online prediction with quadratic loss on the simplex

Proof idea

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

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

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

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

Online prediction with quadratic loss on the simplex

Proof idea

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

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

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

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

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

Online prediction with quadratic loss on the simplex

Proof idea

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

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

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

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

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

Online prediction with quadratic loss on the ball

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

Online prediction with quadratic loss on the ball

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

Use sufficient 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: $\mathcal{Y} = \{y : \|y - c\| \leq r\}$

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

Value-to-go: quadratic in state

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

Online prediction with quadratic loss on the ball

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

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

Value-to-go: quadratic in state

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

Minimax strategy: affine in state

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

Online prediction with quadratic loss on the ball

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

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

Value-to-go: quadratic in state

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

Minimax strategy: affine in state

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

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

Online prediction with quadratic loss on the ball

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

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

Value-to-go: quadratic in state

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

Minimax strategy: affine in state

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

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

Maximin distribution: same mean.

Online prediction with quadratic loss on the ball

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

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

Value-to-go: quadratic in state

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

Minimax strategy: affine in state

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

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

Maximin distribution: same mean.

Minimax regret for ball

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

Online prediction with quadratic loss on the ball

Proof idea

Online prediction with quadratic loss on the ball

Proof idea

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

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

Online prediction with quadratic loss on the ball

Proof idea

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

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

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

Online prediction with quadratic loss on the ball

Proof idea

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

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

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

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

Online prediction with quadratic loss on the ball

Proof idea

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

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

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

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

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

Online prediction with quadratic loss on the ball

Proof idea

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

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

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

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

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

Online prediction with quadratic loss

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

Online prediction with quadratic loss

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

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

Online prediction with quadratic loss

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

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

Online prediction with quadratic loss

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

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

Online prediction with quadratic loss

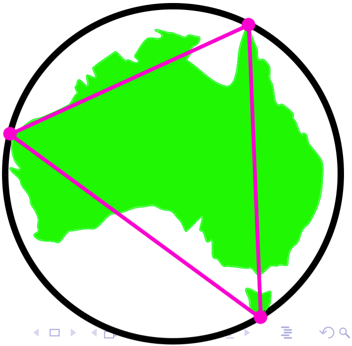
The general case: closed, bounded $\mathcal{Y} \subset \mathbb{R}^d$

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

From below

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

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



Online prediction with quadratic loss

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

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

From below

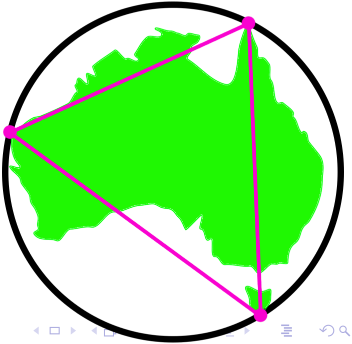
$\mathcal{Y} \supseteq S$, so

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

From above

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

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



Main result: the role of the smallest ball

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

Define the 'minimum radius' function:
$J_{\mathcal{Y}}(c) = \max_{y \in \mathcal{Y}} \|y - c\|$,
so the smallest ball containing $\mathcal{Y}$ is
$B_{\mathcal{Y}} = \{y \in \mathbb{R}^d : \|y - c\| \leq r\}$,
with $r = J_{\mathcal{Y}}(c) = \min_x J_{\mathcal{Y}}(x)$.

Main result: the role of the smallest ball

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

Define the 'minimum radius' function:
$J_{\mathcal{Y}}(c) = \max_{y \in \mathcal{Y}} \|y - c\|$,
so the smallest ball containing $\mathcal{Y}$ is
$B_{\mathcal{Y}} = \{y \in \mathbb{R}^d : \|y - c\| \leq r\}$,
with $r = J_{\mathcal{Y}}(c) = \min_x J_{\mathcal{Y}}(x)$.

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

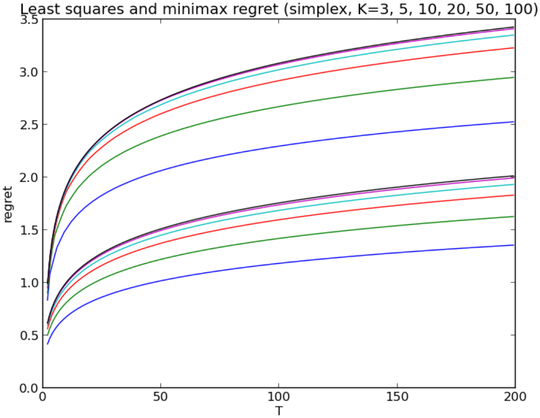
Minimax regret

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

Online prediction with quadratic loss

Minimax regret

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



Online prediction with quadratic loss

Sub-game optimal

- For any closed, bounded $\mathcal{Y}$, the minimax regret is achieved by the strategy for $B_{\mathcal{Y}}$.

Online prediction with quadratic loss

Sub-game optimal

- For any closed, bounded $\mathcal{Y}$, the minimax regret is achieved by the strategy for $B_{\mathcal{Y}}$.
- If $\mathcal{Y}$ is a simplex (with vertices on the surface of $B_{\mathcal{Y}}$), or if $\mathcal{Y} = B_{\mathcal{Y}}$, this strategy is *sub-game optimal*: given any history, it minimizes the worst case regret.

Online prediction with quadratic loss

Sub-game optimal

- For any closed, bounded $\mathcal{Y}$, the minimax regret is achieved by the strategy for $B_{\mathcal{Y}}$.
- If $\mathcal{Y}$ is a simplex (with vertices on the surface of $B_{\mathcal{Y}}$), or if $\mathcal{Y} = B_{\mathcal{Y}}$, this strategy is *sub-game optimal*: given any history, it minimizes the worst case regret.
- For arbitrary $\mathcal{Y}$, this minimax optimal strategy might not be sub-game optimal.

Online prediction with quadratic loss

Sub-game optimal

- For any closed, bounded $\mathcal{Y}$, the minimax regret is achieved by the strategy for $B_{\mathcal{Y}}$.
- If $\mathcal{Y}$ is a simplex (with vertices on the surface of $B_{\mathcal{Y}}$), or if $\mathcal{Y} = B_{\mathcal{Y}}$, this strategy is *sub-game optimal*: given any history, it minimizes the worst case regret.
- For arbitrary $\mathcal{Y}$, this minimax optimal strategy might not be sub-game optimal.
- For $\mathcal{Y}$ an *ellipsoid*, a more complex strategy has this property...

Online prediction with quadratic loss on an ellipsoid

The ellipsoid

$$\mathcal{Y} = \left\{ y \in \mathbb{R}^d : y^\top W y \leq 1 \right\}. \qquad (W \succ 0.)$$

Online prediction with quadratic loss on an ellipsoid

The ellipsoid

$$\mathcal{Y} = \left\{ y \in \mathbb{R}^d : y^\top W y \leq 1 \right\}. \qquad (W \succ 0.)$$

Use sufficient statistics: $s_n = \sum_{t=1}^n y_t, \qquad \sigma_n^2 = \sum_{t=1}^n \|y_t\|^2.$

Online prediction with quadratic loss on an ellipsoid

The ellipsoid

$$\mathcal{Y} = \left\{ y \in \mathbb{R}^d : y^\top W y \leq 1 \right\}. \qquad (W \succ 0.)$$

Use sufficient statistics: $s_n = \sum_{t=1}^n y_t, \qquad \sigma_n^2 = \sum_{t=1}^n \|y_t\|^2.$

Value-to-go: quadratic in state

$$V(y_1, \ldots, y_n) = \frac{1}{2} \left(s_n^\top A_n s_n - \sigma_n^2 + \lambda_{\max}(W^{-1}) \sum_{t=n+1}^{T} \alpha_t \right).$$

$$W^{-1} = \sum_i \nu_i u_i u_i^\top \qquad A_t = \sum_i \frac{\lambda_i^{(t)}}{\nu_i} u_i u_i^\top,$$

$$\lambda_i^{(T)} = \frac{\nu_i}{T}, \qquad \lambda_i^{(t)} = \frac{\lambda_i^{(t+1)}}{\nu_i + \lambda_{\max}^{(t+1)} - \lambda_i^{(t+1)}} \left(\nu_i + \lambda_i^{(t+1)} \right)$$

Online prediction with quadratic loss on an ellipsoid

Minimax strategy: linear in state

$$a^*_{n+1} = B_n s_n.$$

Online prediction with quadratic loss on an ellipsoid

Minimax strategy: linear in state

$$a^*_{n+1} = B_n s_n.$$

$$W^{-1} = \sum_i \nu_i u_i u_i^\top \qquad B_t = \sum_i \frac{\lambda_i^{(t+1)}}{\nu_i + \lambda_{\max}^{(t+1)} - \lambda_i^{(t+1)}} u_i u_i^\top.$$

$$\lambda_i^{(T)} = \frac{\nu_i}{T}, \qquad \lambda_i^{(t)} = \frac{\lambda_i^{(t+1)}}{\nu_i + \lambda_{\max}^{(t+1)} - \lambda_i^{(t+1)}} \left(\nu_i + \lambda_i^{(t+1)}\right)$$

$$\frac{\lambda_{\max}^{(t)}}{\nu_{\max}} = \alpha_t.$$

other directions: more shrinkage.

Online prediction with quadratic loss on an ellipsoid

Minimax strategy: linear in state

$$a^*_{n+1} = B_n s_n.$$

Maximin distribution: same mean, concentrated on two points along the major axis direction.

$$W^{-1} = \sum_i \nu_i u_i u_i^\top \qquad B_t = \sum_i \frac{\lambda_i^{(t+1)}}{\nu_i + \lambda_{\max}^{(t+1)} - \lambda_i^{(t+1)}} u_i u_i^\top.$$

$$\lambda_i^{(T)} = \frac{\nu_i}{T}, \qquad \lambda_i^{(t)} = \frac{\lambda_i^{(t+1)}}{\nu_i + \lambda_{\max}^{(t+1)} - \lambda_i^{(t+1)}} \left(\nu_i + \lambda_i^{(t+1)}\right)$$

$$\frac{\lambda_{\max}^{(t)}}{\nu_{\max}} = \alpha_t.$$

other directions: more shrinkage.

Online prediction with quadratic loss

Online prediction with quadratic loss

- Minimax regret depends on the radius of the smallest ball.

Online prediction with quadratic loss

- Minimax regret depends on the radius of the smallest ball.
- The minimax strategy is simple: shrink the sample average towards the center of the smallest ball.

Online prediction with quadratic loss

- Minimax regret depends on the radius of the smallest ball.
- The minimax strategy is simple: shrink the sample average towards the center of the smallest ball.
- For the simplex and the ball, the strategy is sub-game optimal.

Online prediction with quadratic loss

- Minimax regret depends on the radius of the smallest ball.
- The minimax strategy is simple: shrink the sample average towards the center of the smallest ball.
- For the simplex and the ball, the strategy is sub-game optimal.
- For arbitrary ellipsoids, the strategy involves the same shrinkage in the largest eigenvalue direction, more shrinkage in other directions. This strategy is also sub-game optimal.

Online prediction with quadratic loss

- Minimax regret depends on the radius of the smallest ball.
- The minimax strategy is simple: shrink the sample average towards the center of the smallest ball.
- For the simplex and the ball, the strategy is sub-game optimal.
- For arbitrary ellipsoids, the strategy involves the same shrinkage in the largest eigenvalue direction, more shrinkage in other directions. This strategy is also sub-game optimal.
- Sub-game optimal strategies for other cases (when the convex hull of the contact points between $\mathcal{Y}$ and the surface of the smallest ball is a proper subset of $\mathcal{Y}$)?

Online prediction with quadratic loss

- Minimax regret depends on the radius of the smallest ball.
- The minimax strategy is simple: shrink the sample average towards the center of the smallest ball.
- For the simplex and the ball, the strategy is sub-game optimal.
- For arbitrary ellipsoids, the strategy involves the same shrinkage in the largest eigenvalue direction, more shrinkage in other directions. This strategy is also sub-game optimal.
- Sub-game optimal strategies for other cases (when the convex hull of the contact points between $\mathcal{Y}$ and the surface of the smallest ball is a proper subset of $\mathcal{Y}$)?

Extensions:

Online prediction with quadratic loss

- Minimax regret depends on the radius of the smallest ball.
- The minimax strategy is simple: shrink the sample average towards the center of the smallest ball.
- For the simplex and the ball, the strategy is sub-game optimal.
- For arbitrary ellipsoids, the strategy involves the same shrinkage in the largest eigenvalue direction, more shrinkage in other directions. This strategy is also sub-game optimal.
- Sub-game optimal strategies for other cases (when the convex hull of the contact points between $\mathcal{Y}$ and the surface of the smallest ball is a proper subset of $\mathcal{Y}$)?

Extensions:

- Changing losses: $\ell_n(a, y) = (a - y)^\top W_n (a - y)$.

Online prediction with quadratic loss

- Minimax regret depends on the radius of the smallest ball.
- The minimax strategy is simple: shrink the sample average towards the center of the smallest ball.
- For the simplex and the ball, the strategy is sub-game optimal.
- For arbitrary ellipsoids, the strategy involves the same shrinkage in the largest eigenvalue direction, more shrinkage in other directions. This strategy is also sub-game optimal.
- Sub-game optimal strategies for other cases (when the convex hull of the contact points between $\mathcal{Y}$ and the surface of the smallest ball is a proper subset of $\mathcal{Y}$)?

Extensions:

- Changing losses: $\ell_n(a, y) = (a - y)^\top W_n (a - y)$.
- Linear regression: $\ell_n(\theta, y) = (\theta^\top x_n - y)^2$.

Online prediction with quadratic loss

- Minimax regret depends on the radius of the smallest ball.
- The minimax strategy is simple: shrink the sample average towards the center of the smallest ball.
- For the simplex and the ball, the strategy is sub-game optimal.
- For arbitrary ellipsoids, the strategy involves the same shrinkage in the largest eigenvalue direction, more shrinkage in other directions. This strategy is also sub-game optimal.
- Sub-game optimal strategies for other cases (when the convex hull of the contact points between $\mathcal{Y}$ and the surface of the smallest ball is a proper subset of $\mathcal{Y}$)?

Extensions:

- Changing losses: $\ell_n(a, y) = (a - y)^\top W_n (a - y)$.
- Linear regression: $\ell_n(\theta, y) = (\theta^\top x_n - y)^2$.
- Hilbert space.

Outline

- Computing minimax optimal strategies.
- Prediction games with simple minimax optimal strategies.
- Part 1: Log loss.
  - Normalized maximum likelihood.
  - SNML: predicting like there's no tomorrow.
  - Bayesian strategies.
  - Optimality = exchangeability.
- Part 2: Euclidean loss.
  - The role of the smallest ball.
  - The simplex and the ball.
  - Sub-game optimal strategies on ellipsoids.