Minimax Strategies for Online Linear Regression, Square Loss Prediction,and Time-Series Prediction

Alan Malek
University of California, Berkeley

August 15th, 2016

Joint work with:
Peter L. Bartlett, Wouter M. Koolen, Eiji Takimoto, Manfred Warmuth, Yasin Abbasi-Yadkori

Square loss protocol

Convex set $\mathcal{C}$, length T, and know loss functions ℓ.
For each round $t = 1, \ldots, T$,

- We play $\boldsymbol{a}_t \in \mathcal{C}$
- Nature reveals $y_t \in \mathcal{C}$
- We incur loss

$$\ell(\boldsymbol{a}_t, y_t) = \|\boldsymbol{a}_t - y_t\|^2$$

For some comparator class $\mathcal{A}$, the best comparator is

$$L_T^*(y_1^T) = \min_{\boldsymbol{a} \in \mathcal{A}} \sum_{t=1}^{T} \ell(\boldsymbol{a}, y_t).$$

Goal: find a strategy with minimum regret

$$\text{Regret} \;:=\; \sum_{t=1}^{T} \ell(\boldsymbol{a}_t, y_t) - L_T^*(y_1^T)$$

Section 1

What is Minimax?

What is minimax?

What is minimax?

- In Game Theory

$$V := \inf_{\mathcal{A} \in \text{Strategies}} \sup_{\mathcal{Y} \in \text{Strategies}} \text{Regret}(\mathcal{A}, Y)$$

What is minimax?

- In Game Theory

$$V := \inf_{\mathcal{A} \in \text{Strategies}} \sup_{\mathcal{Y} \in \text{Strategies}} \text{Regret}(\mathcal{A}, Y)$$

- What are strategies?

$$\mathcal{A} = \{g_t : (\boldsymbol{a}_1^{t-1}, \boldsymbol{s}_1^{t-1}) \to \boldsymbol{a}_t\}$$
$$\mathcal{Y} = \{f_t : (\boldsymbol{a}_1^{t}, \boldsymbol{s}_1^{t-1}) \to y_t\}$$

What is minimax?

- In Game Theory

$$V := \inf_{\mathcal{A} \in \text{Strategies}} \sup_{\mathcal{Y} \in \text{Strategies}} \text{Regret}(\mathcal{A}, Y)$$

- What are strategies?

$$\mathcal{A} = \{g_t : (\boldsymbol{a}_1^{t-1}, \boldsymbol{s}_1^{t-1}) \to \boldsymbol{a}_t\}$$
$$\mathcal{Y} = \{f_t : (\boldsymbol{a}_1^{t}, \boldsymbol{s}_1^{t-1}) \to y_t\}$$

- How can we solve for g_t and f_t?

Value of the game

$$\sum_{t=1}^{T} \ell(a_t, y_t) - L^*(y_1^T)$$

Value of the game

$$\min_{\boldsymbol{a}_T} \qquad \sum_{t=1}^{T} \ell(\boldsymbol{a}_t, y_t) - L^*(y_1^T)$$

Value of the game

$$\min_{a_T} \max_{y_T} \sum_{t=1}^{T} \ell(a_t, y_t) - L^*(y_1^T)$$

Value of the game

$$\min_{a_{T-1}} \max_{y_{T-1}} \min_{a_T} \max_{y_T} \sum_{t=1}^{T} \ell(a_t, y_t) - L^*(y_1^T)$$

Value of the game

$$V := \min_{a_1} \max_{y_1} \cdots \min_{a_{T-1}} \max_{y_{T-1}} \min_{a_T} \max_{y_T} \sum_{t=1}^{T} \ell(a_t, y_t) - L^*(y_1^T)$$

Value of the game

$$V := \min_{a_1} \max_{y_1} \cdots \min_{a_{T-1}} \max_{y_{T-1}} \min_{a_T} \max_{y_T} \sum_{t=1}^{T} \ell(a_t, y_t) - L^*(y_1^T)$$

- a_t can be a function of a_1^{t-1} and y_1^{t-1} only
- y_t can be a function of a_1^t and y_1^{t-1} only

Value of the game

$$V := \min_{a_1} \max_{y_1} \cdots \min_{a_{T-1}} \max_{y_{T-1}} \min_{a_T} \max_{y_T} \sum_{t=1}^{T} \ell(a_t, y_t) - L^*(y_1^T)$$

- a_t can be a function of a_1^{t-1} and y_1^{t-1} only
- y_t can be a function of a_1^{t} and y_1^{t-1} only
- Value is the regret when both players are optimal

Value of the game

$$V := \min_{a_1} \max_{y_1} \cdots \min_{a_{T-1}} \max_{y_{T-1}} \min_{a_T} \max_{y_T} \sum_{t=1}^{T} \ell(a_t, y_t) - L^*(y_1^T)$$

- a_t can be a function of a_1^{t-1} and y_1^{t-1} only
- y_t can be a function of a_1^t and y_1^{t-1} only
- Value is the regret when both players are optimal
- How can we compute them?

Backwards induction

$$\sum_{t=1}^{T} \ell(a_t, y_t) - L^*(y_1^T)$$

Backwards induction

$$\min_{\boldsymbol{a}_T} \max_{y_T} \sum_{t=1}^{T} \ell(\boldsymbol{a}_t, y_t) - L^*(y_1^T)$$

Backwards induction

$$\min_{\boldsymbol{a}_T} \max_{y_T} \sum_{t=1}^{T} \ell(\boldsymbol{a}_t, y_t) - L^*(y_1^T)$$

$$= \sum_{t=1}^{T-1} \ell(\boldsymbol{a}_t, y_t) + \min_{\boldsymbol{a}_T} \max_{y_T} \ell(\boldsymbol{a}_T, y_T) - L^*(y_1^T)$$

Backwards induction

$$\min_{a_T} \max_{y_T} \sum_{t=1}^{T} \ell(a_t, y_t) - L^*(y_1^T)$$

$$= \sum_{t=1}^{T-1} \ell(a_t, y_t) + \min_{a_T} \max_{y_T} \ell(a_T, y_T) - L^*(y_1^T)$$

$$= \sum_{t=1}^{T-1} \ell(a_t, y_t) + V_{T-1}(y_1^{T-1})$$

Backwards induction

$$\min_{a_{T-1}} \max_{y_{T-1}} \min_{a_T} \max_{y_T} \sum_{t=1}^{T} \ell(a_t, y_t) - L^*(y_1^T)$$

$$= \min_{a_{T-1}} \max_{y_{T-1}} \sum_{t=1}^{T-1} \ell(a_t, y_t) + V_{T-1}(y_1^{T-1})$$

Backwards induction

$$\min_{\boldsymbol{a}_{T-1}} \max_{y_{T-1}} \min_{\boldsymbol{a}_T} \max_{y_T} \sum_{t=1}^{T} \ell(\boldsymbol{a}_t, y_t) - L^*(y_1^T)$$

$$= \min_{\boldsymbol{a}_{T-1}} \max_{y_{T-1}} \sum_{t=1}^{T-1} \ell(\boldsymbol{a}_t, y_t) + V_{T-1}(y_1^{T-1})$$

$$= \sum_{t=1}^{T-2} \ell(\boldsymbol{a}_t, y_t) + \min_{\boldsymbol{a}_{T-1}} \max_{y_{T-1}} \ell(\boldsymbol{a}_{T-1}, y_{T-1}) + V_{T-1}(y_1^{T-1})$$

Backwards induction

$$\min_{a_{T-1}} \max_{y_{T-1}} \min_{a_T} \max_{y_T} \sum_{t=1}^{T} \ell(a_t, y_t) - L^*(y_1^T)$$

$$\begin{aligned}
&= \min_{a_{T-1}} \max_{y_{T-1}} \sum_{t=1}^{T-1} \ell(a_t, y_t) + V_{T-1}(y_1^{T-1}) \\
&= \sum_{t=1}^{T-2} \ell(a_t, y_t) + \min_{a_{T-1}} \max_{y_{T-1}} \ell(a_{T-1}, y_{T-1}) + V_{T-1}(y_1^{T-1}) \\
&= \sum_{t=1}^{T-2} \ell(a_t, y_t) + V_{T-2}(y_1^{T-2})
\end{aligned}$$

Backwards induction

$$\min_{\boldsymbol{a}_1} \max_{y_1} \cdots \min_{\boldsymbol{a}_{T-1}} \max_{y_{T-1}} \min_{\boldsymbol{a}_T} \max_{y_T} \sum_{t=1}^{T} \ell(\boldsymbol{a}_t, y_t) - L^*(y_1^T)$$

$$= \min_{\boldsymbol{a}_1} \max_{y_1} \cdots \sum_{t=1}^{T-2} \ell(\boldsymbol{a}_t, y_t) + V_{T-2}(y_1^{T-2})$$

Backwards induction

$$\min_{a_1} \max_{y_1} \cdots \min_{a_{T-1}} \max_{y_{T-1}} \min_{a_T} \max_{y_T} \sum_{t=1}^{T} \ell(a_t, y_t) - L^*(y_1^T)$$

$$= \min_{a_1} \max_{y_1} \cdots \sum_{t=1}^{T-2} \ell(a_t, y_t) + V_{T-2}(y_1^{T-2})$$

$\cdots$

Value-to-go

Inductive definition:

$$V_T(y_1^T) := -L_T^*(y_1^T) \quad (1)$$

$$V_{t-1}(y_1^{t-1}) := \min_{a_t} \max_{y_t} \ell(a_t, y_t) + V_t(y_1^{t-1}, y_t) \quad (2)$$

Value-to-go

Inductive definition:

$$V_T(y_1^T) \;:=\; -L_T^*(y_1^T) \tag{1}$$

$$V_{t-1}(y_1^{t-1}) \;:=\; \min_{a_t} \max_{y_t} \ell(a_t, y_t) + V_t(y_1^{t-1}, y_t) \tag{2}$$

The minimax regret V equals value-to-go $V_0(\epsilon)$ (empty history).
The minimax strategy: after seeing $y_1, \ldots, y_{t-1}$,

- Compute $V_t(y_1, \ldots, y_t)$
- Choose a_t as the minimizer of

$$\max_{y_t} \ell(a_t, y_t) + V_t(y_1^{t-1}, y_t)$$

Value-to-go

Inductive definition:

$$V_T(y_1^T) := -L_T^*(y_1^T) \tag{1}$$

$$V_{t-1}(y_1^{t-1}) := \min_{a_t} \max_{y_t} \ell(a_t, y_t) + V_t(y_1^{t-1}, y_t) \tag{2}$$

The minimax regret V equals value-to-go $V_0(\epsilon)$ (empty history).
The minimax strategy: after seeing $y_1, \ldots, y_{t-1}$,

- Compute $V_t(y_1, \ldots, y_t)$
- Choose a_t as the minimizer of

$$\max_{y_t} \ell(a_t, y_t) + V_t(y_1^{t-1}, y_t)$$

Problem: this is expensive (usually exponentially so).

Outline

- What is minimax?
- Two square loss games
- Fixed-design online linear regression
- Time series prediction

Section 2

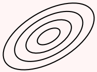
Square loss game

Square loss protocol (with Koolen and Bartlett)

Convex set $\mathcal{C}$, length T, and know loss functions ℓ.
For each round $t = 1, \ldots, T$,

- We play $\boldsymbol{a}_t \in \mathcal{C}$
- Nature reveals $y_t \in \mathcal{C}$
- We incur loss

matrix $\boldsymbol{W}$ weights prediction errors



$$\ell(\boldsymbol{a}_t, y_t) \;:=\; \|\boldsymbol{a}_t - y_t\|^2_{\boldsymbol{W}} \;=\; (\boldsymbol{a}_t - y_t)^\intercal \boldsymbol{W}^{-1}(\boldsymbol{a}_t - y_t)$$

Our goal is to minimize regret w.r.t. best fixed action $\boldsymbol{a}$ in hindsight

$$\text{Regret} \;:=\; \sum_{t=1}^{T} \ell(\boldsymbol{a}_t, y_t) - \min_{\boldsymbol{a}} \sum_{t=1}^{T} \ell(\boldsymbol{a}, y_t)$$

Solving the minimax strategy

Value-to-go:

$$V_T(y_1^T) := -L_T^*(y_1^T) \quad (1)$$

$$V_{t-1}(y_1^{t-1}) := \min_{a_t} \max_{y_t} \ell(a_t, y_t) + V_t(y_1^{t-1}, y_t) \quad (2)$$

- Using sufficient statistics

$$s_t = \sum_{\tau=1}^{t} y_\tau \qquad \text{and} \qquad {\sigma^2}_t = \sum_{\tau=1}^{t} y_\tau^\intercal W^{-1} y_\tau$$

- First, we need $L_T^*(y_1^T)$:

$$L_T^* = \inf_{a \in \mathbb{R}^d} \sum_{t=1}^{T} \|a - y_t\|_W^2 = {\sigma^2}_T - \frac{1}{T} s_T^\intercal W^{-1} s_t$$

and the minimizer is the mean outcome $a^* = \frac{1}{T} \sum_{t=1}^{T} y_t$.

Calculating the value function for $\mathcal{C} = \triangle$

Value-to-go:

$$V_T(y_1^T) := -L_T^*(y_1^T) \tag{1}$$

$$V_{t-1}(y_1^{t-1}) := \min_{\boldsymbol{a}_t} \max_{y_t} \ell(\boldsymbol{a}_t, y_t) + V_t(y_1^{t-1}, y_t) \tag{2}$$

- Base case: $V_T(y_1^T) = -L_T^* = \frac{1}{T}\boldsymbol{s}_T^\intercal \boldsymbol{W}^{-1}\boldsymbol{s}_T - \sigma^2 T$

Calculating the value function for $\mathcal{C} = \triangle$

Value-to-go:

$$V_T(y_1^T) \;:=\; -L_T^*(y_1^T) \tag{1}$$

$$V_{t-1}(y_1^{t-1}) \;:=\; \min_{\boldsymbol{a}_t} \max_{y_t} \ell(\boldsymbol{a}_t, y_t) + V_t(y_1^{t-1}, y_t) \tag{2}$$

- Base case: $V_T(y_1^T) = -L_T^* = \frac{1}{T} \boldsymbol{s}_T^\intercal \boldsymbol{W}^{-1} \boldsymbol{s}_T - \sigma^2{}_T$
- "Guess":

$$V_t(\boldsymbol{s}_t, \sigma^2{}_t) \;=\; \alpha_t \boldsymbol{s}_t^\intercal \boldsymbol{W}^{-1} \boldsymbol{s}_t - \sigma^2{}_t + \boldsymbol{b}^\intercal \boldsymbol{s}_t + \gamma_t,$$

Calculating the value function for $\mathcal{C} = \triangle$

Value-to-go:

$$V_T(y_1^T) := -L_T^*(y_1^T) \tag{1}$$

$$V_{t-1}(y_1^{t-1}) := \min_{\boldsymbol{a}_t} \max_{y_t} \ell(\boldsymbol{a}_t, y_t) + V_t(y_1^{t-1}, y_t) \tag{2}$$

- Base case: $V_T(y_1^T) = -L_T^* = \frac{1}{T} \boldsymbol{s}_T^\intercal \boldsymbol{W}^{-1} \boldsymbol{s}_T - \sigma^2{}_T$
- "Guess":

$$V_t(\boldsymbol{s}_t, \sigma^2{}_t) = \alpha_t \boldsymbol{s}_t^\intercal \boldsymbol{W}^{-1} \boldsymbol{s}_t - \sigma^2{}_t + \boldsymbol{b}^\intercal \boldsymbol{s}_t + \gamma_t,$$

- Base case: $\alpha_T = \frac{1}{T}$, $\gamma_t = 0$

Calculating the value function for $\mathcal{C} = \triangle$

Value-to-go:

$$V_T(y_1^T) := -L_T^*(y_1^T) \tag{1}$$
$$V_{t-1}(y_1^{t-1}) := \min_{\boldsymbol{a}_t} \max_{y_t} \ell(\boldsymbol{a}_t, y_t) + V_t(y_1^{t-1}, y_t) \tag{2}$$

- Base case: $V_T(y_1^T) = -L_T^* = \frac{1}{T}\boldsymbol{s}_T^\intercal \boldsymbol{W}^{-1}\boldsymbol{s}_T - \sigma^2{}_T$
- "Guess":

$$V_t(\boldsymbol{s}_t, \sigma^2{}_t) = \alpha_t \boldsymbol{s}_t^\intercal \boldsymbol{W}^{-1}\boldsymbol{s}_t - \sigma^2{}_t + \boldsymbol{b}^\intercal \boldsymbol{s}_t + \gamma_t,$$

- Base case: $\alpha_T = \frac{1}{T}$, $\gamma_t = 0$
- Induction:

$$V_t(\boldsymbol{s}_t, \sigma^2{}_t) = \min_{\boldsymbol{a}\in\triangle} \max_{y\in\triangle} \ell(\boldsymbol{a}, y) + V_{t+1}(\boldsymbol{s}_t + y, \sigma^2{}_t + y^\intercal \boldsymbol{W}^{-1} y)$$

$$
\begin{aligned}
V_t(\boldsymbol{s}_t, {\sigma^2}_t) = \min_{\boldsymbol{a} \in \triangle} \max_{y \in \triangle} & \|\boldsymbol{a} - y\|^2_{\boldsymbol{W}} + \alpha_t (\boldsymbol{s}_t + y)^\intercal \boldsymbol{W}^{-1} (\boldsymbol{s}_t + y) \\
& - ({\sigma^2}_t + y^\intercal \boldsymbol{W}^{-1} y) + \gamma_t + \boldsymbol{b}^\intercal (\boldsymbol{s}_t + y)
\end{aligned}
$$

$$
\begin{aligned}
V_t(\boldsymbol{s}_t, {\sigma^2}_t) &= \min_{\boldsymbol{a} \in \triangle} \max_{y \in \triangle} \|\boldsymbol{a} - y\|^2_{\boldsymbol{W}} + \alpha_t (\boldsymbol{s}_t + y)^\intercal \boldsymbol{W}^{-1} (\boldsymbol{s}_t + y) \\
&\quad - ({\sigma^2}_t + y^\intercal \boldsymbol{W}^{-1} y) + \gamma_t + \boldsymbol{b}^\intercal (\boldsymbol{s}_t + y) \\
&= \min_{\boldsymbol{a} \in \triangle} \max_{y \in \triangle} \|\boldsymbol{a} - y\|^2_{\boldsymbol{W}} + (\alpha_t - 1) y^\intercal \boldsymbol{W}^{-1} y + \boldsymbol{b}'^\intercal y (+c)
\end{aligned}
$$

$$
\begin{aligned}
V_t(\boldsymbol{s}_t, {\sigma^2}_t) &= \min_{\boldsymbol{a}\in\triangle} \max_{y\in\triangle} \|\boldsymbol{a} - y\|^2_{\boldsymbol{W}} + \alpha_t(\boldsymbol{s}_t + y)^\intercal \boldsymbol{W}^{-1}(\boldsymbol{s}_t + y) \\
&\quad - ({\sigma^2}_t + y^\intercal \boldsymbol{W}^{-1} y) + \gamma_t + \boldsymbol{b}^\intercal(\boldsymbol{s}_t + y) \\
&= \min_{\boldsymbol{a}\in\triangle} \max_{y\in\triangle} \|\boldsymbol{a} - y\|^2_{\boldsymbol{W}} + (\alpha_t - 1) y^\intercal \boldsymbol{W}^{-1} y + \boldsymbol{b}'^\intercal y (+c) \\
&= \min_{\boldsymbol{a}\in\triangle} \max_{k} \|\boldsymbol{a} - \boldsymbol{e}_k\|^2_{\boldsymbol{W}} + (\alpha_t - 1) \boldsymbol{e}_k^\intercal \boldsymbol{W}^{-1} \boldsymbol{e}_k + \boldsymbol{b}'^\intercal \boldsymbol{e}_k
\end{aligned}
$$

$$
\begin{aligned}
V_t(\boldsymbol{s}_t, {\sigma^2}_t) &= \min_{\boldsymbol{a}\in\triangle}\max_{y\in\triangle} \|\boldsymbol{a} - y\|_{\boldsymbol{W}}^2 + \alpha_t(\boldsymbol{s}_t + y)^\intercal \boldsymbol{W}^{-1}(\boldsymbol{s}_t + y) \\
&\quad - ({\sigma^2}_t + y^\intercal \boldsymbol{W}^{-1} y) + \gamma_t + \boldsymbol{b}^\intercal(\boldsymbol{s}_t + y) \\
&= \min_{\boldsymbol{a}\in\triangle}\max_{y\in\triangle} \|\boldsymbol{a} - y\|_{\boldsymbol{W}}^2 + (\alpha_t - 1) y^\intercal \boldsymbol{W}^{-1} y + \boldsymbol{b}'^\intercal y (+c) \\
&= \min_{\boldsymbol{a}\in\triangle}\max_{k} \|\boldsymbol{a} - \boldsymbol{e}_k\|_{\boldsymbol{W}}^2 + (\alpha_t - 1)\boldsymbol{e}_k^\intercal \boldsymbol{W}^{-1}\boldsymbol{e}_k + \boldsymbol{b}'^\intercal \boldsymbol{e}_k \\
&= \max_{\boldsymbol{p}}\min_{\boldsymbol{a}\in\triangle} \mathbb{E}_{k\sim\boldsymbol{p}}\Big[\|\boldsymbol{a} - \boldsymbol{e}_k\|_{\boldsymbol{W}}^2 + (\alpha_t - 1)\boldsymbol{e}_k^\intercal \boldsymbol{W}^{-1}\boldsymbol{e}_k + \boldsymbol{b}'^\intercal \boldsymbol{e}_k\Big]
\end{aligned}
$$

$$
\begin{aligned}
V_t(\boldsymbol{s}_t, \sigma^2{}_t) &= \min_{\boldsymbol{a}\in\triangle} \max_{y\in\triangle} \|\boldsymbol{a} - y\|^2_{\boldsymbol{W}} + \alpha_t(\boldsymbol{s}_t + y)^\intercal \boldsymbol{W}^{-1}(\boldsymbol{s}_t + y) \\
&\quad - (\sigma^2{}_t + y^\intercal \boldsymbol{W}^{-1} y) + \gamma_t + \boldsymbol{b}^\intercal(\boldsymbol{s}_t + y) \\
&= \min_{\boldsymbol{a}\in\triangle} \max_{y\in\triangle} \|\boldsymbol{a} - y\|^2_{\boldsymbol{W}} + (\alpha_t - 1) y^\intercal \boldsymbol{W}^{-1} y + \boldsymbol{b}'^\intercal y (+c) \\
&= \min_{\boldsymbol{a}\in\triangle} \max_{k} \|\boldsymbol{a} - \boldsymbol{e}_k\|^2_{\boldsymbol{W}} + (\alpha_t - 1)\boldsymbol{e}_k^\intercal \boldsymbol{W}^{-1} \boldsymbol{e}_k + \boldsymbol{b}'^\intercal \boldsymbol{e}_k \\
&= \max_{\boldsymbol{p}} \min_{\boldsymbol{a}\in\triangle} \mathbb{E}_{k\sim\boldsymbol{p}}\left[\|\boldsymbol{a} - \boldsymbol{e}_k\|^2_{\boldsymbol{W}} + (\alpha_t - 1)\boldsymbol{e}_k^\intercal \boldsymbol{W}^{-1} \boldsymbol{e}_k + \boldsymbol{b}'^\intercal \boldsymbol{e}_k\right] \\
&= \max_{\boldsymbol{p}} -\boldsymbol{p}^\intercal \boldsymbol{W}^{-1} \boldsymbol{p} + \left(\alpha_t \operatorname{diag}(\boldsymbol{W}^{-1}) + \boldsymbol{b}'\right)^\intercal \boldsymbol{p}
\end{aligned}
$$

$$
\begin{aligned}
V_t(\boldsymbol{s}_t, {\sigma^2}_t) &= \min_{\boldsymbol{a}\in\triangle} \max_{y\in\triangle} \|\boldsymbol{a} - y\|^2_{\boldsymbol{W}} + \alpha_t(\boldsymbol{s}_t + y)^\intercal \boldsymbol{W}^{-1}(\boldsymbol{s}_t + y) \\
&\quad - ({\sigma^2}_t + y^\intercal \boldsymbol{W}^{-1} y) + \gamma_t + \boldsymbol{b}^\intercal(\boldsymbol{s}_t + y) \\
&= \min_{\boldsymbol{a}\in\triangle} \max_{y\in\triangle} \|\boldsymbol{a} - y\|^2_{\boldsymbol{W}} + (\alpha_t - 1) y^\intercal \boldsymbol{W}^{-1} y + \boldsymbol{b}'^\intercal y (+c) \\
&= \min_{\boldsymbol{a}\in\triangle} \max_{k} \|\boldsymbol{a} - \boldsymbol{e}_k\|^2_{\boldsymbol{W}} + (\alpha_t - 1)\boldsymbol{e}_k^\intercal \boldsymbol{W}^{-1} \boldsymbol{e}_k + \boldsymbol{b}'^\intercal \boldsymbol{e}_k \\
&= \max_{\boldsymbol{p}} \min_{\boldsymbol{a}\in\triangle} \mathbb{E}_{k\sim\boldsymbol{p}}\Big[\|\boldsymbol{a} - \boldsymbol{e}_k\|^2_{\boldsymbol{W}} + (\alpha_t - 1)\boldsymbol{e}_k^\intercal \boldsymbol{W}^{-1} \boldsymbol{e}_k + \boldsymbol{b}'^\intercal \boldsymbol{e}_k\Big] \\
&= \max_{\boldsymbol{p}} -\boldsymbol{p}^\intercal \boldsymbol{W}^{-1} \boldsymbol{p} + \big(\alpha_t \operatorname{diag}(\boldsymbol{W}^{-1}) + \boldsymbol{b}'\big)^\intercal \boldsymbol{p}
\end{aligned}
$$

Easy to solve via Lagrange multipliers.

Simplex game (e.g. Brier game)

Theorem

Let $\mathcal{C} = \triangle$. For $\boldsymbol{W}$ satisfying an alignment condition, the value-to-go is

$$V_t(\boldsymbol{s}_t, {\sigma^2}_t) = \alpha_t \boldsymbol{s}_t^\intercal \boldsymbol{W}^{-1} \boldsymbol{s}_t - {\sigma^2}_t + \underbrace{(1 - t\alpha_t)\operatorname{diag}(\boldsymbol{W}^{-1})^\intercal}_{\boldsymbol{b}_t} \boldsymbol{s}_t + \mathit{const}$$

with coefficients

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

The minimax and maximin strategies are

$$\boldsymbol{a}_t = \boldsymbol{p}_t = \frac{\boldsymbol{s}_t}{t} t\alpha_{t+1} + \boldsymbol{c}(1 - t\alpha_{t+1})$$

which is data mean $\frac{\boldsymbol{s}_t}{t}$ shrunk towards center $\boldsymbol{c} = \frac{\boldsymbol{W}\mathbf{1}}{\mathbf{1}^\intercal \boldsymbol{W}\mathbf{1}} + \left(\boldsymbol{W} - \frac{\boldsymbol{W}\mathbf{1}\mathbf{1}^\intercal \boldsymbol{W}}{\mathbf{1}^\intercal \boldsymbol{W}\mathbf{1}}\right)\operatorname{diag}(\boldsymbol{W}^{-1})$.

Ball game

Theorem

Let $\mathcal{C} = \bigcirc$. For any positive definite $\boldsymbol{W}$ the value-to-go is

$$V_t(\boldsymbol{s}_t, {\sigma^2}_t) = \boldsymbol{s}_t^\intercal \boldsymbol{A}_t \boldsymbol{s} - {\sigma^2}_t + \textit{const.}$$

For round t, the minimax strategy plays

$$\boldsymbol{a}^* \; = \; \left(\lambda_{\max}(\boldsymbol{A}_t)\boldsymbol{I} - (\boldsymbol{A}_t - \boldsymbol{W}^{-1})\right)^{-1} \boldsymbol{A}_t \boldsymbol{s}$$

with coefficients $\boldsymbol{A}_T = \frac{1}{T}\boldsymbol{W}^{-1}$ and

$$\boldsymbol{A}_{t-1} = \boldsymbol{A}_t \left(\boldsymbol{W}^{-1} + \lambda_{\max}(\boldsymbol{A}_t)\boldsymbol{I} - \boldsymbol{A}_t\right)^{-1} \boldsymbol{A}_t + \boldsymbol{A}_t.$$

Regret bounds

- $\text{Regret}_{\text{Brier}} \propto \sum_{t=1}^{T} \alpha_t$.
- $\text{Regret}_{\text{Ball}} = \lambda_{\max}(\boldsymbol{W}^{-1}) \sum_{t=1}^{T} \alpha_t$.
- [1] show that $\sum_{t=1}^{T} \alpha_t = \log(T) - \log\log(T) + O(\frac{\log(T)}{\log\log(T)})$.
- Compare with $O(\log(T))$ of Follow the Leader.

E. Takimoto, M. Warmuth
The minimax strategy for Gaussian density estimation
In *COLT '00*

Section 3

Online Linear regression

Online linear regression (with Bartlett, Koolen, Takimoto, Warmuth)

Fix a covariate sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ (fixed design) and length T.
For each round $t = 1, \ldots, T$,

- We play $\boldsymbol{a}_t \in \mathbb{R}$
- Nature reveals $y_t \in [-B_t, B_t]$
- We incur loss

$$\ell(\boldsymbol{a}_t, y_t) = (\boldsymbol{a}_t - y_t)^2$$

- Minimax Regret is

$$\min_{\boldsymbol{a}_1} \max_{y_1} \cdots \min_{\boldsymbol{a}_T} \max_{y_T} \sum_{t=1}^{T} (\boldsymbol{a}_t - y_t)^2 - \min_{\theta \in \mathbb{R}^d} \sum_{t=1}^{T} (\theta^\intercal \boldsymbol{x}_t - y_t)^2$$

Online linear regression (with Bartlett, Koolen, Takimoto, Warmuth)

Fix a covariate sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ (fixed design) and length T.
For each round $t = 1, \ldots, T$,

- We play $\boldsymbol{a}_t \in \mathbb{R}$
- Nature reveals $y_t \in [-B_t, B_t]$
- We incur loss

$$\ell(\boldsymbol{a}_t, y_t) \;=\; (\boldsymbol{a}_t - y_t)^2$$

- Minimax Regret is

$$\min_{\boldsymbol{a}_1} \max_{y_1} \cdots \min_{\boldsymbol{a}_T} \max_{y_T} \; \underbrace{\sum_{t=1}^{T} (\boldsymbol{a}_t - y_t)^2}_{\text{algorithm}} - \min_{\theta \in \mathbb{R}^d} \sum_{t=1}^{T} (\theta^\intercal \boldsymbol{x}_t - y_t)^2$$

Online linear regression (with Bartlett, Koolen, Takimoto, Warmuth)

Fix a covariate sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ (fixed design) and length T.
For each round $t = 1, \ldots, T$,

- We play $\boldsymbol{a}_t \in \mathbb{R}$
- Nature reveals $y_t \in [-B_t, B_t]$
- We incur loss

$$\ell(\boldsymbol{a}_t, y_t) \;=\; (\boldsymbol{a}_t - y_t)^2$$

- Minimax Regret is

$$\min_{\boldsymbol{a}_1} \max_{y_1} \cdots \min_{\boldsymbol{a}_T} \max_{y_T} \; \underbrace{\sum_{t=1}^{T} (\boldsymbol{a}_t - y_t)^2}_{\text{algorithm}} - \underbrace{\min_{\theta \in \mathbb{R}^d} \sum_{t=1}^{T} (\theta^\intercal \boldsymbol{x}_t - y_t)^2}_{\text{best linear predictor}}$$

Solving the value-to-go

Value-to-go:

$$V_T(y_1^T) := -L_T^*(y_1^T) \quad (1)$$

$$V_{t-1}(y_1^{t-1}) := \min_{\boldsymbol{a}_t} \max_{y_t} \ell(\boldsymbol{a}_t, y_t) + V_t(y_1^{t-1}, y_t) \quad (2)$$

- Define

$$\boldsymbol{s}_t = \sum_{\tau=1}^{t} y_\tau \boldsymbol{x}_\tau, \qquad \sigma^2{}_t = \sum_{\tau=1}^{t} y_\tau^2, \qquad \boldsymbol{P}_T = \left(\sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t^\intercal\right)^{-1}$$

- Base case is ordinary least squares: $\theta^* = \boldsymbol{P}_T \boldsymbol{s}_T$ and

$$V_T(y_1^T) = -L_T^*(y_1^T) = \min_{\theta \in \mathbb{R}^d} \sum_{t=1}^{T} (\theta^\intercal \boldsymbol{x}_t - y_t)^2$$

$$= \boldsymbol{s}_T^\intercal \boldsymbol{P}_T \boldsymbol{s}_T - \sigma^2{}_T.$$

- Induction Hypothesis

$$V_t(\boldsymbol{s}_t, \sigma_t^2) = \boldsymbol{s}_t^\intercal \boldsymbol{P}_t \boldsymbol{s}_t - {\sigma^2}_t + \gamma_t,$$

- Induction Hypothesis

$$V_t(\boldsymbol{s}_t, \sigma_t^2) \;=\; \boldsymbol{s}_t^\intercal \boldsymbol{P}_t \boldsymbol{s}_t - \sigma^2{}_t + \gamma_t,$$

- Backwards induction:

$$\begin{aligned}
V_t\left(\boldsymbol{s}_t, \sigma_t^2\right) &:= \min_{\boldsymbol{a}_{t+1}} \max_{y_{t+1}} \left(\boldsymbol{a}_{t+1} - y_{t+1}\right)^2 \\
&\qquad + \left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}\right)^\intercal \boldsymbol{P}_{t+1} \left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}\right) \\
&\qquad - \left(\sigma^2{}_t + y_{t+1}^2\right) + \gamma_{t+1} \\
&= \min_{\boldsymbol{a}_{t+1}} \max_{y_{t+1}} \boldsymbol{a}_{t+1}^2 + 2y_{t+1}\left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \boldsymbol{a}_{t+1}\right) \\
&\qquad + \left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\right) y_{t+1}^2 + \text{const}
\end{aligned}$$

- Induction Hypothesis

$$V_t(\boldsymbol{s}_t, \sigma_t^2) \; = \; \boldsymbol{s}_t^\intercal \boldsymbol{P}_t \boldsymbol{s}_t - {\sigma^2}_t + \gamma_t,$$

- Backwards induction:

$$\begin{aligned} V_t\left(\boldsymbol{s}_t, \sigma_t^2\right) &:= \min_{\boldsymbol{a}_{t+1}} \max_{y_{t+1}} \left(\boldsymbol{a}_{t+1} - y_{t+1}\right)^2 \\ &\quad + \left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}\right)^\intercal \boldsymbol{P}_{t+1} \left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}\right) \\ &\quad - \left({\sigma^2}_t + y_{t+1}^2\right) + \gamma_{t+1} \\ &= \min_{\boldsymbol{a}_{t+1}} \max_{y_{t+1}} \boldsymbol{a}_{t+1}^2 + 2y_{t+1}\left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \boldsymbol{a}_{t+1}\right) \\ &\quad + \left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\right) y_{t+1}^2 + \text{const} \end{aligned}$$

- Induction Hypothesis

$$V_t(\boldsymbol{s}_t, \sigma_t^2) \;=\; \boldsymbol{s}_t^\intercal \boldsymbol{P}_t \boldsymbol{s}_t - {\sigma^2}_t + \gamma_t,$$

- Backwards induction:

$$\begin{aligned} V_t\left(\boldsymbol{s}_t, \sigma_t^2\right) &:= \min_{\boldsymbol{a}_{t+1}} \max_{y_{t+1}} \left(\boldsymbol{a}_{t+1} - y_{t+1}\right)^2 \\ &\quad + \left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}\right)^\intercal \boldsymbol{P}_{t+1} \left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}\right) \\ &\quad - \left({\sigma^2}_t + y_{t+1}^2\right) + \gamma_{t+1} \\ &= \min_{\boldsymbol{a}_{t+1}} \max_{y_{t+1}} \boldsymbol{a}_{t+1}^2 + 2y_{t+1}\left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \boldsymbol{a}_{t+1}\right) \\ &\quad + \left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\right) y_{t+1}^2 + \text{const} \end{aligned}$$

- This is convex in y_{t+1} and hence $y_{t+1} = \pm B_{t+1}$, so

$$\begin{aligned} V_t\left(\boldsymbol{s}_t, \sigma_t^2\right) &= \min_{\boldsymbol{a}_{t+1}} \boldsymbol{a}_{t+1}^2 + 2B_{t+1}\left|\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \boldsymbol{a}_{t+1}\right| \\ &\quad + \left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\right) B^2 + \text{const} \end{aligned}$$

- We had

$$V_t\left(\boldsymbol{s}_t, \sigma_t^2\right) = \min_{\boldsymbol{a}_{t+1}} \boldsymbol{a}_{t+1}^2 + 2B_{t+1}\left|\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \boldsymbol{a}_{t+1}\right| + \left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\right)B^2 + \text{const}$$

- We had

$$V_t\left(\boldsymbol{s}_t, \sigma_t^2\right) = \min_{\boldsymbol{a}_{t+1}} \boldsymbol{a}_{t+1}^2 + 2B_{t+1}\left|\boldsymbol{x}_{t+1}^\top \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \boldsymbol{a}_{t+1}\right| + \left(\boldsymbol{x}_{t+1}^\top \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\right)B^2 + \text{const}$$

- If $|\boldsymbol{x}_{t+1}^\top \boldsymbol{P}_{t+1}\boldsymbol{s}_t| \leq B_{t+1}$, setting subgradient to 0 yields

$$\boldsymbol{a}_{t+1} = \boldsymbol{x}_{t+1}^\top \boldsymbol{P}_{t+1}\boldsymbol{s}_t$$

- We had

$$V_t\left(\boldsymbol{s}_t, \sigma_t^2\right) = \min_{\boldsymbol{a}_{t+1}} \boldsymbol{a}_{t+1}^2 + 2B_{t+1}\left|\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \boldsymbol{a}_{t+1}\right| + \left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\right) B^2 + \text{const}$$

- If $|\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t| \leq B_{t+1}$, setting subgradient to 0 yields

$$\boldsymbol{a}_{t+1} = \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t$$

- Plugging in this $\boldsymbol{a}_{t+1}$, we get

$$V_t\left(\boldsymbol{s}_t, {\sigma^2}_t\right) = \boldsymbol{s}_t^\intercal \overbrace{\left(\boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} + \boldsymbol{P}_{t+1}\right)}^{:=\boldsymbol{P}_t} \boldsymbol{s}_t - {\sigma^2}_t + \underbrace{\gamma_{t+1} + B_{t+1}^2 \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}}_{:=\gamma_t}$$

- We had

$$V_t\left(\boldsymbol{s}_t, \sigma_t^2\right) = \min_{\boldsymbol{a}_{t+1}} \boldsymbol{a}_{t+1}^2 + 2B_{t+1}\left|\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \boldsymbol{a}_{t+1}\right| + \left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\right) B^2 + \text{const}$$

- If $\left|\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t\right| \leq B_{t+1}$, setting subgradient to 0 yields

$$\boldsymbol{a}_{t+1} = \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t$$

- Plugging in this $\boldsymbol{a}_{t+1}$, we get

$$V_t\left(\boldsymbol{s}_t, {\sigma^2}_t\right) = \boldsymbol{s}_t^\intercal \overbrace{\left(\boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} + \boldsymbol{P}_{t+1}\right)}^{:=\boldsymbol{P}_t} \boldsymbol{s}_t - {\sigma^2}_t + \underbrace{\gamma_{t+1} + B_{t+1}^2 \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}}_{:=\gamma_t}$$

- Value is

$$V_0(\mathbf{0}, 0) = \gamma_0 = \sum_{t=1}^{T} B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$$

In Summary

Theorem

The strategy

$$a_{t+1} \;=\; x_{t+1}^{\intercal} P_{t+1} s_t, \qquad \text{(MM)}$$

is minimax optimal and the value-to-go is

$$V_t(s_t, \sigma_t^2) \;=\; s_t^{\intercal} P_t s_t - {\sigma^2}_t + \gamma_t,$$

with coefficients

$$P_T = \left(\sum_{t=1}^{T} x_t x_t^{\intercal} \right)^{-1}, \qquad P_t = P_{t+1} + P_{t+1} x_{t+1} x_{t+1}^{\intercal} P_{t+1},$$

$$\gamma_T = 0, \qquad \gamma_t = \gamma_{t+1} + B_{t+1}^2 x_{t+1}^{\intercal} P_{t+1} x_{t+1},$$

provided the box constraints $|x_{t+1}^{\intercal} P_{t+1} s_t| \leq B_{t+1}$ *hold.*

Interpretation of $\boldsymbol{P}_t$

$$\boldsymbol{P}_t^{-1} = \underbrace{\sum_{\tau=1}^{t} \boldsymbol{x}_\tau \boldsymbol{x}_\tau^\intercal}_{\text{least squares}} + \underbrace{\sum_{\tau=t+1}^{T} \frac{\boldsymbol{x}_\tau^\intercal \boldsymbol{P}_\tau \boldsymbol{x}_\tau}{1 + \boldsymbol{x}_\tau^\intercal \boldsymbol{P}_\tau \boldsymbol{x}_\tau} \boldsymbol{x}_\tau \boldsymbol{x}_\tau^\intercal}_{\text{re-weighted future instances}} .$$

Interpretation of $\boldsymbol{P}_t$

$$\boldsymbol{P}_t^{-1} = \underbrace{\sum_{\tau=1}^{t} \boldsymbol{x}_\tau \boldsymbol{x}_\tau^\intercal}_{\text{least squares}} + \underbrace{\sum_{\tau=t+1}^{T} \frac{\boldsymbol{x}_\tau^\intercal \boldsymbol{P}_\tau \boldsymbol{x}_\tau}{1 + \boldsymbol{x}_\tau^\intercal \boldsymbol{P}_\tau \boldsymbol{x}_\tau} \boldsymbol{x}_\tau \boldsymbol{x}_\tau^\intercal}_{\text{re-weighted future instances}} .$$

Recall regret at round t: $B_t \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$

Section 4

Tracking

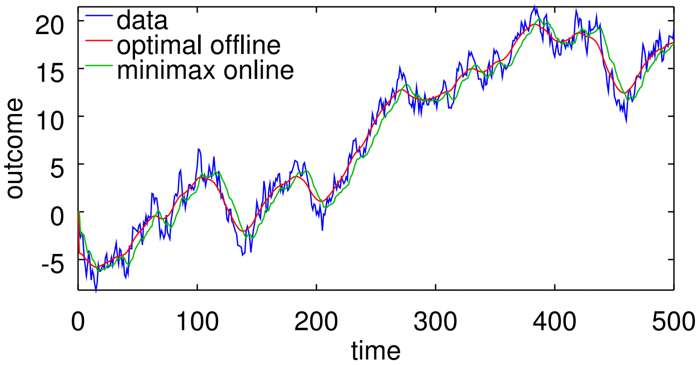
Time series prediction protocol (with Koolen, Bartlett, Abbasi-Yadkori)

Fix a convex set $\mathcal{C}$, length T, regularization parameter λ_T.
For each round $t = 1, \ldots, T$,

- We play $\boldsymbol{a}_t \in \mathcal{C}$
- Nature reveals $\boldsymbol{y}_t \in \mathcal{C}$
- We incur loss $\ell(\boldsymbol{a}_t, \boldsymbol{y}_t) \;:=\; \|\boldsymbol{a}_t - \boldsymbol{y}_t\|^2$
- Regret:

$$\underbrace{\sum_{t=1}^{T} \|\boldsymbol{a}_t - \boldsymbol{y}_t\|^2}_{\text{Our loss}} - \min_{\hat{\boldsymbol{a}}_1,\ldots,\hat{\boldsymbol{a}}_T} \left\{ \underbrace{\sum_{t=1}^{T} \|\hat{\boldsymbol{a}}_t - \boldsymbol{y}_t\|^2}_{\text{Loss of Comparator}} + \underbrace{\lambda_T \operatorname{tr}(\boldsymbol{K} \hat{\boldsymbol{A}}^\intercal \hat{\boldsymbol{A}})}_{\text{Comparator Complexity}} \right\}$$

  where $\hat{\boldsymbol{A}} = [\hat{\boldsymbol{a}}_1 \cdots \hat{\boldsymbol{a}}_T]$
- E.g. $\operatorname{tr}(\boldsymbol{K} \hat{\boldsymbol{A}}^\intercal \hat{\boldsymbol{A}}) = \sum_{t=1}^{T+1} \|\hat{\boldsymbol{a}}_t - \hat{\boldsymbol{a}}_{t-1}\|^2$



data
optimal offline
minimax online
outcome
time
20
15
10
5
0
-5
0
100
200
300
400
500

Solving the value-to-go (part 3)

Value-to-go:

$$V_T(y_1^T) \;:=\; -L_T^*(y_1^T) \tag{1}$$

$$V_{t-1}(y_1^{t-1}) \;:=\; \min_{\boldsymbol{a}_t} \max_{y_t} \ell(\boldsymbol{a}_t, y_t) + V_t(y_1^{t-1}, y_t) \tag{2}$$

Histories are $Y_t = [y_1 \cdots y_t]$.
Offline Problem: $\hat{\boldsymbol{A}} = Y_T(\boldsymbol{I} + \lambda_T \boldsymbol{K})^{-1}$ and value

$$V_T(Y_T) = -L^* = -\operatorname{tr}\left(Y_T(\boldsymbol{I} - (\boldsymbol{I} + \lambda_T \boldsymbol{K})^{-1}) Y_T^\intercal\right)$$

Behavior of backwards induction

Theorem

If $\|\boldsymbol{b}\| \leq 1$*, then the minimax problem*

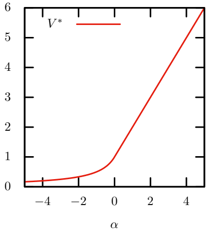
$$V^* := \min_{\boldsymbol{a}} \max_{y:\|y\|\leq 1} \|\boldsymbol{a} - y\|^2 + (\alpha - 1)\|y\|^2 + 2\boldsymbol{b}^\intercal y$$

has value and minimizer

$$V^* = \begin{cases} \frac{\|\boldsymbol{b}\|^2}{1-\alpha} & \text{if } \alpha \leq 0, \\ \|\boldsymbol{b}\|^2 + \alpha & \text{if } \alpha \geq 0, \end{cases} \quad \text{and} \quad \boldsymbol{a} = \begin{cases} \frac{\boldsymbol{b}}{1-\alpha} & \text{if } \alpha \leq 0, \\ \boldsymbol{b} & \text{if } \alpha \geq 0. \end{cases}$$

Non-trivial induction:

- Curvature of optimization can switch between rounds
- Yet can pre-compute beforehand



Minimax solution

Input: T, $\boldsymbol{K}$, λ_T

Using:

- single-shot game solution, and
- lots of matrix identities

Output: matrices $\boldsymbol{R}_t = \begin{pmatrix} \boldsymbol{A}_t & \boldsymbol{b}_t \\ \boldsymbol{b}_t^\intercal & c_t \end{pmatrix}$

strategy $$\boldsymbol{a}_t = X_{t-1} \begin{cases} \frac{\boldsymbol{b}_t}{1-c_t} & \text{if } c_t \leq 0, \\ \boldsymbol{b}_t - c_t \boldsymbol{e}_t & \text{if } c_t \geq 0. \end{cases}$$

Theorem

Under a (typical) no clipping condition on Y_T,

$$V_t(Y_t) = \operatorname{tr}\left(Y_t\left(\boldsymbol{R}_t - \boldsymbol{I}\right) Y_t^\intercal\right) + \sum_{\tau=t+1}^{T} \max\{c_\tau, 0\}$$

and, in the vanilla case (norm bounded data, increments penalized),

$$V = \Theta\left(\frac{T}{\sqrt{1+\lambda_T}}\right).$$

Section 5

Conclusion

- Minimax algorithms can be computationally efficient with enough structure, e.g.
  - Normalized Maximum likelihood that is Bayesian
  - Certain square losses
- Exploited the fact that saddle point problems with square loss are nice
- Can we characterize the class of functions that are closed w.r.t. the backwards induction operator?

Section 6

Extra slides

Ball game maximin

The maximin strategy plays two unit length vectors with

$$\Pr\left(y = \boldsymbol{a}_\perp \pm \sqrt{1 - \boldsymbol{a}_\perp^\intercal \boldsymbol{a}_\perp}\boldsymbol{v}_{\max}\right) = \frac{1}{2} \pm \frac{\boldsymbol{a}_\parallel^\intercal \boldsymbol{v}_{\max}}{2\sqrt{1 - \boldsymbol{a}_\perp^\intercal \boldsymbol{a}_\perp}},$$

where $\lambda_{\max}$ and $\boldsymbol{v}_{\max}$ correspond to the largest eigenvalue of $\boldsymbol{A}_{t+1}$ and $\boldsymbol{a}_\perp$ and $\boldsymbol{a}_\parallel$ are the components of $\boldsymbol{a}^*$ perpendicular and parallel to $\boldsymbol{v}_{\max}$.

Tracking: second order $\boldsymbol{K}$

- Computation: if $\boldsymbol{K}$ and $\boldsymbol{v}_t$ are banded then $\boldsymbol{R}_t^{-1}$ is sparse
- Here we *imposed* data bound $\|Y_t \boldsymbol{v}_t\| \leq 1$. In the paper we show that the minimax strategy guarantees an *adaptive* bound scaling with $\|Y_t \boldsymbol{v}_t\|$.
- A second order smoothness version of $\boldsymbol{K}$ gives complicated c_t

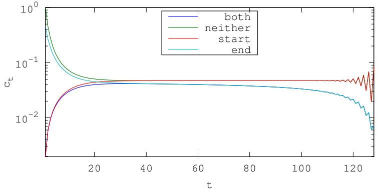


Figure: $\boldsymbol{v}_t = \boldsymbol{e}_t - \boldsymbol{e}_{t-1}$

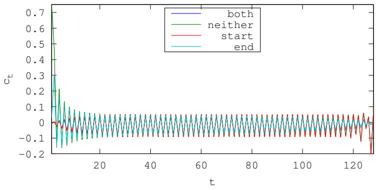


Figure: $\boldsymbol{v}_t = \boldsymbol{e}_t - 2\boldsymbol{e}_{t-1} + \boldsymbol{e}_{t-2}$

Ellipse

Fix a budget $R \geq 0$, and consider label sequences

$$\mathcal{Y}_R := \left\{ y_1, \ldots, y_T \in \mathbb{R} : \sum_{t=1}^{T} y_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t = R \right\}$$

We show that (MM) is minimax for this set.

In fact, the regret of (MM) *equals*

$$\mathcal{R}_T \;=\; \sum_{t=1}^{T} y_t^2 \boldsymbol{x}_t \boldsymbol{P}_t \boldsymbol{x}_t.$$

This means that this algorithm has two very special properties. First, it is a *strong equalizer* in the sense that it suffers the same regret on all 2^T sign-flips of the labels. And second, it is *adaptive* to the complexity R of the labels.