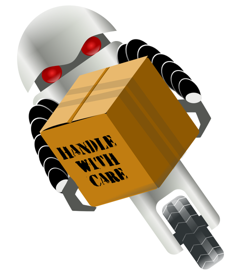
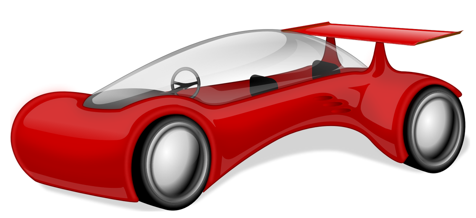
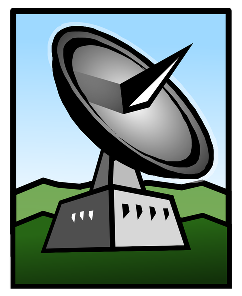
Minimax Time Series Prediction

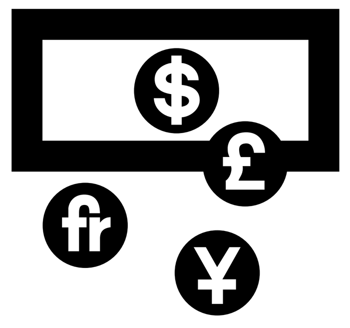
Wouter M. Koolen Alan Malek Peter Bartlett Yasin Abbasi-Yadkori

Time series everywhere

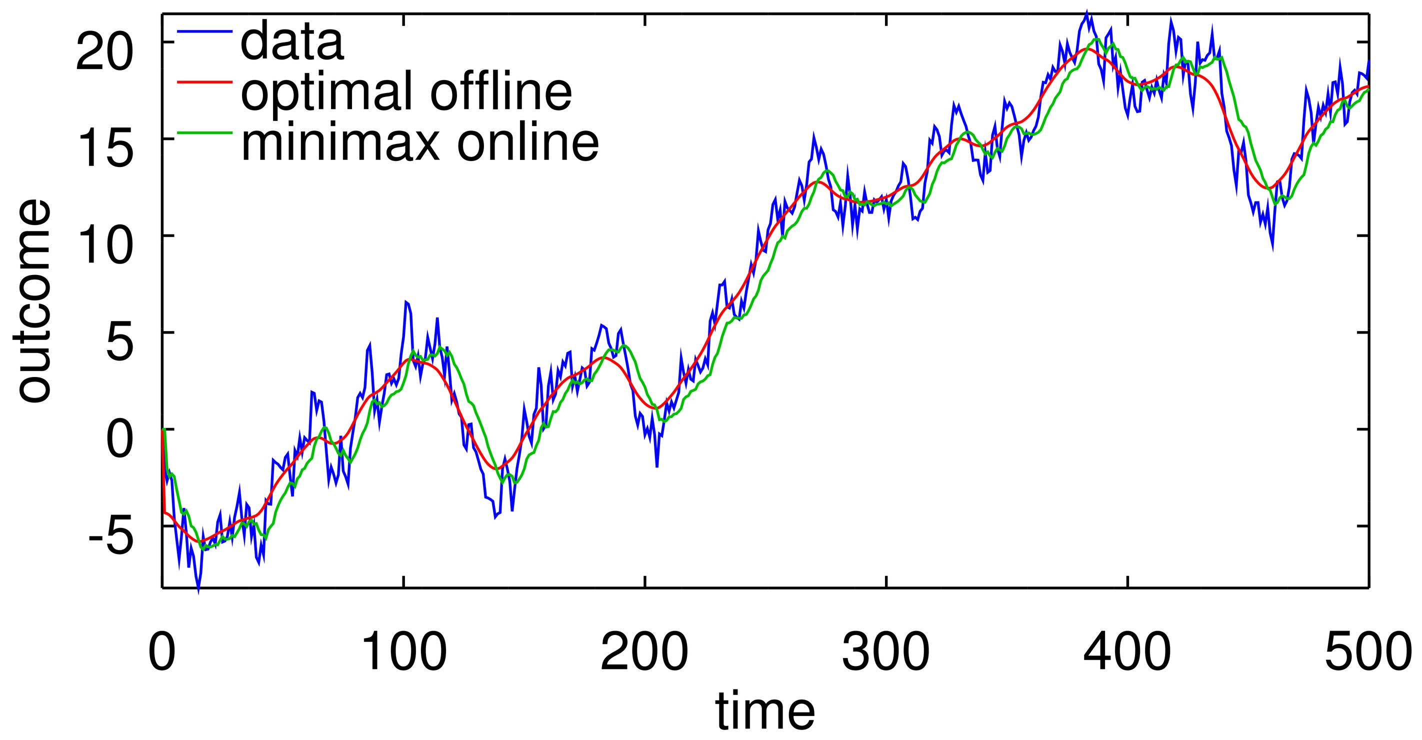

Pose estimation


Self-driving


Tracking


Finance

What we do: Online Prediction



Model: Time series Game

For some convex set $\mathcal{C}$, each round $t = 1, \ldots, T$

- We play $\boldsymbol{a}_t \in \mathcal{C}$
- Nature reveals $\boldsymbol{x}_t \in \mathcal{C}$
- We incur loss $\|\boldsymbol{a}_t - \boldsymbol{x}_t\|^2$

Fix horizon T and regularization scalar $\lambda_T > 0$. Regret is:

$$\underbrace{\sum_{t=1}^T \|\boldsymbol{a}_t - \boldsymbol{x}_t\|^2}_{\text{Our loss}} - \min_{\hat{\boldsymbol{a}}_1,\ldots,\hat{\boldsymbol{a}}_T} \left\{ \underbrace{\sum_{t=1}^T \|\hat{\boldsymbol{a}}_t - \boldsymbol{x}_t\|^2}_{\text{Loss of Comparator}} + \underbrace{\lambda_T \sum_{t=1}^{T+1} \|\hat{\boldsymbol{a}}_t - \hat{\boldsymbol{a}}_{t-1}\|^2}_{\text{Comparator Complexity}} \right\}$$

Objective: Minimax Regret

If we assume a perfect adversary, how well can we do? Value is

$$V \;:=\; \min_{\boldsymbol{a}_1} \max_{\boldsymbol{x}_1 : \|\boldsymbol{x}_1\| \le 1} \ldots \min_{\boldsymbol{a}_T} \max_{\boldsymbol{x}_T : \|\boldsymbol{x}_T\| \le 1} \text{Regret}$$

We play to minimize the worst-case regret; e.g. find $\boldsymbol{a}_t$ that guarantees we achieve the game's value.

In general

Let $\boldsymbol{X}_t = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_t]$ and $\hat{\boldsymbol{A}} = [\hat{\boldsymbol{a}}_1 \cdots \hat{\boldsymbol{a}}_T]$. For $\boldsymbol{v}_t \in \mathbb{R}^t$ and $\boldsymbol{K} \succeq \boldsymbol{0}$,

Data domain	$\|\boldsymbol{X}_t \boldsymbol{v}_t\| \le 1$	e.g. $\|\boldsymbol{x}_t\| \le 1$
Complexity	$\operatorname{tr}(\boldsymbol{K} \hat{\boldsymbol{A}}^\intercal \hat{\boldsymbol{A}})$	e.g. $\sum_{t=1}^{T+1} \|\hat{\boldsymbol{a}}_t - \hat{\boldsymbol{a}}_{t-1}\|^2$

Spectrum of games (in particular: higher order differences)

Offline Problem

Theorem 1 *For any complexity matrix $\boldsymbol{K} \succeq \boldsymbol{0}$, regularization scalar $\lambda_T \ge 0$, and $d \times T$ data matrix $\boldsymbol{X}_T = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_T]$ the problem*

$$L^* \;:=\; \min_{\hat{\boldsymbol{a}}_1,\ldots,\hat{\boldsymbol{a}}_T} \sum_{t=1}^T \|\hat{\boldsymbol{a}}_t - \boldsymbol{x}_t\|^2 + \lambda_T \operatorname{tr}(\boldsymbol{K} \hat{\boldsymbol{A}}^\intercal \hat{\boldsymbol{A}})$$

has linear (in $\boldsymbol{X}_T$) minimizer and quadratic (in $\boldsymbol{X}_T$) value given by

$$\hat{\boldsymbol{A}} = \boldsymbol{X}_T(\boldsymbol{I} + \lambda_T \boldsymbol{K})^{-1} \quad \textit{and} \quad L^* = \operatorname{tr}\left(\boldsymbol{X}_T(\boldsymbol{I} - (\boldsymbol{I} + \lambda_T \boldsymbol{K})^{-1})\boldsymbol{X}_T^\intercal\right).$$

Backward Induction Solution

We solve for the **value-to-go** V from each state $\boldsymbol{X}_t = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_t]$. We have $V(\boldsymbol{X}_T) \;:=\; -L^*$ and

$$V(\boldsymbol{X}_{t-1}) \;:=\; \min_{\boldsymbol{a}_t} \max_{\boldsymbol{x}_t : \|\boldsymbol{X}_t \boldsymbol{v}_t\| \le 1} \|\boldsymbol{a}_t - \boldsymbol{x}_t\|^2 + V(\boldsymbol{X}_t).$$

The minimax regret V equals value-to-go $V(\epsilon)$ from empty history.

Crux

Value-to-go V stays quadratic in $\boldsymbol{X}_t$ for all $t \le T$ and corresponding minimax strategy is linear in $\boldsymbol{X}_{t-1}$. Remains to compute coefficients.

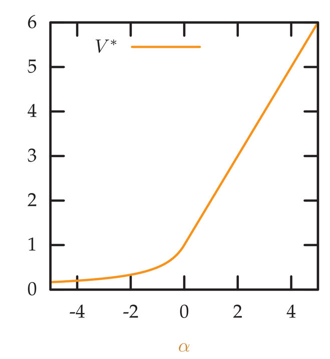
Single-shot squared loss game

Theorem 2 *If $\|\boldsymbol{b}\| \le 1$, then the minimax problem*

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

has value and minimizer

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



Non-trivial induction:

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

Minimax time series prediction

We precompute:

Input: $T, \boldsymbol{K}, \lambda_T, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_T$

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

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

Theorem 3 *Under a (typical)* no clipping *condition on $\boldsymbol{X}_T$,*

$$V(\boldsymbol{X}_t) \;=\; \operatorname{tr}\left(\boldsymbol{X}_t (\boldsymbol{R}_t - \boldsymbol{I}) \boldsymbol{X}_t^\intercal\right) + \sum_{s=t+1}^T \max\{c_s, 0\}$$

linear filter

$$\boldsymbol{a}_t \;=\; \boldsymbol{X}_{t-1} \begin{cases} \frac{\boldsymbol{b}_t}{1-c_t} & \text{if } c_t \le 0, \\ \boldsymbol{b}_t - c_t \boldsymbol{v}_t^{<t} & \text{if } c_t \ge 0. \end{cases}$$

Vanilla case Norm-bounded Data with Increment Squared Regularization

- Cheap scalar $O(T)$ preprocessing sweep.
- Predict in $O(d)$ time per round using $O(d)$ memory.
- Filter weights roughly decay exponentially backwards.
- Can upper and lower bound regret to get

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

constant λ_T overfits

But wait, there's more

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

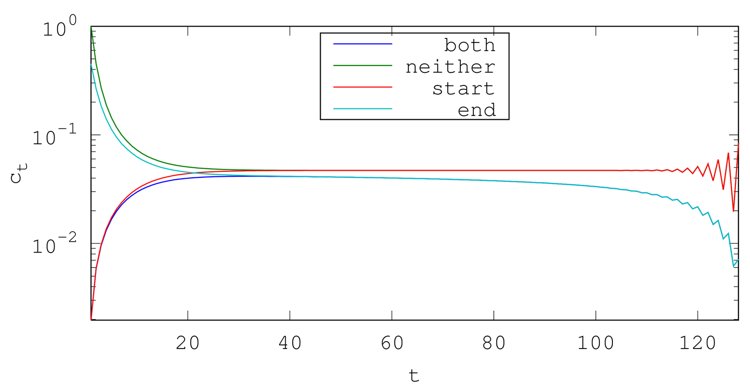


Figure 1: $\boldsymbol{v}_t = \boldsymbol{e}_t - \boldsymbol{e}_{t-1}$

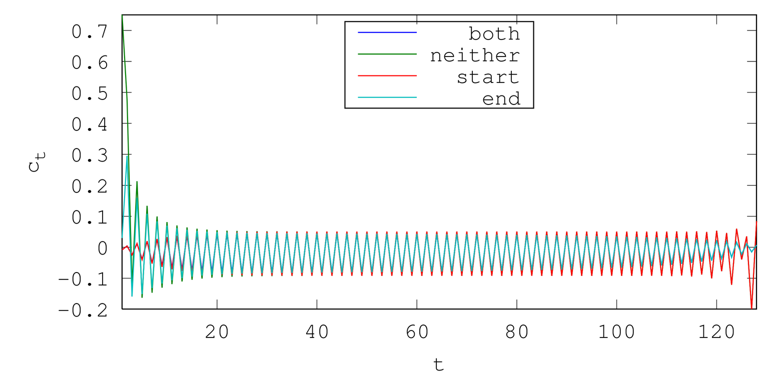


Figure 2: $\boldsymbol{v}_t = \boldsymbol{e}_t - 2\boldsymbol{e}_{t-1} + \boldsymbol{e}_{t-2}$