Minimax Fixed-Design Linear Regression

Peter L. Bartlett Wouter M. Koolen Alan Malek Eiji Takimoto Manfred K. Warmuth

Scope and Contribution

Linear regression is one of the fundamental machine learning tasks.

We consider the online version of linear regression with fixed design (instances are revealed from the outset, labels are predicted sequentially). We show that the *exact* minimax strategy is *tractable*.

- *Ideal regularization* emerges from the problem
- Case study for incorporating *unlabeled data*
- Optimal strategy employs *intricate shrinkage*

Protocol

Given: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in \mathbb{R}^d$
For $t = 1, 2, \ldots, T$:

- Learner issues prediction $\hat{y}_t \in \mathbb{R}$
- Adversary reveals label $y_t \in \mathbb{R}$
- Learner incurs loss $(\hat{y}_t - y_t)^2$.

Offline Problem

The best *linear predictor* in hindsight:

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

is ordinary least squares

$$\theta = \left(\sum_{t=1}^T \boldsymbol{x}_t \boldsymbol{x}_t^\intercal\right)^{-1} \left(\sum_{t=1}^T y_t \boldsymbol{x}_t\right)$$

with loss

$$\sum_{t=1}^T y_t^2 - \left(\sum_{t=1}^T y_t \boldsymbol{x}_t\right)^\intercal \left(\sum_{t=1}^T \boldsymbol{x}_t \boldsymbol{x}_t^\intercal\right)^{-1} \left(\sum_{t=1}^T y_t \boldsymbol{x}_t\right).$$

Online Problem

The goal of the learner is to predict almost as well as the best linear predictor in hindsight. The overhead is measured by the **regret**

$$\mathcal{R}_T := \sum_{t=1}^T (\hat{y}_t - y_t)^2 - \min_{\theta \in \mathbb{R}^d} \sum_{t=1}^T (\theta^\intercal \boldsymbol{x}_t - y_t)^2.$$

We consider the minimax problem

$$\min_{\hat{y}_1} \max_{y_1} \cdots \min_{\hat{y}_T} \max_{y_T} \mathcal{R}_T$$

So, what is the optimal strategy to choose $\hat{y}_t$ given $y_t, \ldots, y_{t-1}$?

Popular approaches

$$\hat{y}_{t+1}^{\text{FTL}} := \boldsymbol{x}_{t+1}^\intercal \left(\sum_{q=1}^t \boldsymbol{x}_t \boldsymbol{x}_t^\intercal\right)^{-1} \sum_{q=1}^t y_t \boldsymbol{x}_t$$

$$\hat{y}_{t+1}^{\text{Ridge}} := \boldsymbol{x}_{t+1}^\intercal \left(\sum_{q=1}^t \boldsymbol{x}_t \boldsymbol{x}_t^\intercal + \lambda \boldsymbol{I}\right)^{-1} \sum_{q=1}^t y_t \boldsymbol{x}_t$$

$$\hat{y}_{t+1}^{\text{LSM}} := \boldsymbol{x}_{t+1}^\intercal \left(\sum_{q=1}^{t+1} \boldsymbol{x}_t \boldsymbol{x}_t^\intercal\right)^{-1} \sum_{q=1}^t y_t \boldsymbol{x}_t$$

Recurrence

Define recursively

$$\boldsymbol{P}_T = \left(\sum_{t=1}^T \boldsymbol{x}_t \boldsymbol{x}_t^\intercal\right)^{-1},$$

and

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

or, equivalently,

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

We can compute $\boldsymbol{P}_1 \cdots \boldsymbol{P}_T$ in $O(Td^2 + d^3)$ time.

The MM Strategy

After t rounds, define a summary statistic $\boldsymbol{s}_t := \sum_{q=1}^t y_q \boldsymbol{x}_q$. We define the MM strategy to predict

$$\hat{y}_{t+1} = \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t, \tag{MM}$$

Box-Constrained Labels

Consider the label sequence constraint

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

We show that (MM) is minimax for this set provided that the budgets $B = (B_1, \ldots, B_T)$ are compatible with the covariates by satisfying

$$B_t \geq \sum_{q=1}^{t-1} |\boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_q| B_q. \tag{1}$$

In this case, the minimax regret is

$$\sum_{t=1}^T B_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$$

and the maximin probability distribution for y_{t+1} puts weight $1/2 \pm \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t / (2B_{t+1})$ on $\pm B_{t+1}$.

Ellipse-Constrained Labels

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.

Analysis flavor

Recursion for value of minimax problem.

$$V_T\left(\boldsymbol{s}_T, \sigma_T^2\right) = -\min_{\theta \in \mathbb{R}^d} \left(\sum_{t=1}^T (\theta^\intercal \boldsymbol{x}_t - y_t)^2\right),$$

$$V_t\left(\boldsymbol{s}_t, \sigma_t^2\right) = \min_{\hat{y}_{t+1}} \max_{y_{t+1}} \Big((\hat{y}_{t+1} - y_{t+1})^2 + V_{t+1}\left(\boldsymbol{s}_t + y_{t+1} \boldsymbol{x}_{t+1}, \sigma_t^2 + y_{t+1}^2\right) \Big)$$

with the state $(\boldsymbol{s}_t, \sigma_t^2)$ after t rounds defined by

$$\boldsymbol{s}_t = \sum_{q=1}^t y_q \boldsymbol{x}_q, \qquad \sigma_t^2 = \sum_{q=1}^t y_q^2$$

(and $\boldsymbol{s}_0 = \boldsymbol{0}$, $\sigma_0^2 = 0$).

Crux: value stays quadratic

We show by induction that

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

with the γ_t coefficients recursively defined by

$$\gamma_T = 0, \quad \gamma_t = \gamma_{t+1} + B_{t+1}^2 \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{x}_{t+1}.$$

(where $|y_t| \leq B_t$) and hence the value equals

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

Clipping

The condition (1) is necessary to ensure that the label constraint $|y_t| \leq B_t$ on the adversary is *inactive* for the worst-case label.

If (1) is violated then the Adversary is *clipped* to $y_t = \pm B_t$ and the Learner benefits by clipping as well. This breaks the nice quadratic recursion.

Regret bound

For box-constrained label with $B_t = B$ we prove that

$$\mathcal{R}_T \leq O(B^2 d \ln T)$$

(independent of scale of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$).

Future directions

- Worst-case *ordering* of given set of covariates? In 1d increasing magnitude seems hardest. How does this generalize?
- Worst-case covariates? We conjecture composition of orthogonal 1d problems. Would improve regret to $O(B^2 d \ln(T/d))$.
- Gap between minimax and strategies like [?] with correct asymptotics. $O(\ln \ln T)$ difference?
- Worst case covariates with *adversarial* design? Is the minimax analysis tractable, perhaps under some reasonable conditions?