JMLR: Workshop and Conference Proceedings vol 40:1–14, 2015

Minimax Fixed-Design Linear Regression

Peter L. Bartlett BARTLETT@CS.BERKELEY.EDU
University of California at Berkeley and Queensland University of Technology

Wouter M. Koolen WOUTER.KOOLEN@QUT.EDU.AU
Queensland University of Technology

Alan Malek MALEK@BERKELEY.EDU
University of California at Berkeley

Eiji Takimoto EIJI@INF.KYUSHU-U.AC.JP
Kyushu University

Manfred K. Warmuth MANFRED@UCSC.EDU
University of California, Santa Cruz

Abstract

We consider a linear regression game in which the covariates are known in advance: at each round, the learner predicts a real-value, the adversary reveals a label, and the learner incurs a squared error loss. The aim is to minimize the regret with respect to linear predictions. For a variety of constraints on the adversary's labels, we show that the minimax optimal strategy is linear, with a parameter choice that is reminiscent of ordinary least squares (and as easy to compute). The predictions depend on all covariates, past and future, with a particular weighting assigned to future covariates corresponding to the role that they play in the minimax regret. We study two families of label sequences: box constraints (under a covariate compatibility condition), and a weighted 2-norm constraint that emerges naturally from the analysis. The strategy is adaptive in the sense that it requires no knowledge of the constraint set. We obtain an explicit expression for the minimax regret for these games. For the case of uniform box constraints, we show that, with worst case covariate sequences, the regret is $O(d \log T)$, with no dependence on the scaling of the covariates.

Keywords: linear regression, online learning, minimax regret

1. Introduction

Linear regression is a core prediction problem in machine learning and statistics. The goal is to find parameters $\boldsymbol{w} \in \mathbb{R}^d$ so that the linear predictor $\boldsymbol{x} \mapsto \boldsymbol{w}^\intercal \boldsymbol{x}$ has small square loss $\sum_t (\boldsymbol{w}^\intercal \boldsymbol{x}_t - y_t)^2$ on the data $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots$ of interest. In this paper, we study the online fixed-design variant of the problem; that is, the covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ are given to the learner in advance and on round $t = 1, \ldots, T$, the learner predicts $\hat{y}_t$ before seeing the correct real-valued label y_t and incurring square loss $(\hat{y}_t - y_t)^2$. See Figure 1. The goal of the learner is to perform almost as well as the best fixed linear predictor in hindsight; i.e., the learner wants to minimize the regret

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

We study the minimax regret,

$$\min_{\hat{y}_1} \max_{y_1} \cdots \min_{\hat{y}_T} \max_{y_T} \mathcal{R}_T \tag{1}$$

where the sequence $y_1, \ldots, y_T$ is constrained to some set. We consider two cases: individually bounded labels, and a sum constraint on label sequences.

A strategy is a mapping from sequences $y_1, \ldots, y_{t-1}$ of previous outcomes to predictions $\hat{y}_t$. The minimax strategy is the one that minimizes the worst case regret over outcome sequences. In general, computing minimax strategies is computationally intractable: one needs to choose the optimal $\hat{y}_t$ for every possible history $y_1, \ldots, y_{t-1}$. Therefore, it is of considerable interest when the minimax strategies are easily computable.

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

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

Figure 1: Fixed-design protocol

Outline and our contribution In Section 2, we investigate the minimax regret problem for fixed design linear regression (1) when the labels y_t are bounded. We derive conditions on the covariates that yield an explicit, tractable minimax strategy. Specifically, we find that the minimax strategy is a simple, linear predictor. After t rounds, define a summary statistic $\boldsymbol{s}_t \coloneqq \sum_{q=1}^{t} y_q \boldsymbol{x}_q$. The minimax strategy (we call it MM) predicts

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

where the $\boldsymbol{P}_t$ are particular problem-specific matrices, defined in terms of the covariates, that turn out to be crucial for the linear regression problem. They can be computed in advance, before any labels are seen. The algorithm is efficient, as computation of the $\boldsymbol{P}_t$ sequence is amortized $O(d^3)$ per round (in contrast to the exponential dependence on T of brute-force backward induction). For the case $|y_t| \le B$ for all t, we show that the minimax regret is exactly $B^2 \sum_{t=1}^{T} \boldsymbol{x}_t^{\intercal} \boldsymbol{P}_t \boldsymbol{x}_t$. Section 3 provides a direct regret bound of $O(B^2 d \log T)$ for the worst case covariate sequence. Interestingly, this bound is invariant to the scale of $\boldsymbol{x}_t$.

The minimax strategy is reminiscent of the classical ordinary least squares (OLS) method for a probabilistic linear regression model: given data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T)$, OLS predicts $\boldsymbol{x}^{\intercal} \widehat{\boldsymbol{w}}$, where

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

The minimax predictions have the same form: $\widehat{\boldsymbol{w}}_{t+1} = \boldsymbol{P}_{t+1} \boldsymbol{s}_t$, where $\boldsymbol{P}_{t+1}$ and $\boldsymbol{s}_t$ correspond to online analogs to the empirical precision matrix and covariance vector. $\boldsymbol{P}_t$ is defined by:

$$\boldsymbol{P}_t^{-1} = \sum_{q=1}^{t} \boldsymbol{x}_q \boldsymbol{x}_q^{\intercal} + \sum_{q=t+1}^{T} \frac{\boldsymbol{x}_q^{\intercal} \boldsymbol{P}_q \boldsymbol{x}_q}{1 + \boldsymbol{x}_q^{\intercal} \boldsymbol{P}_q \boldsymbol{x}_q} \boldsymbol{x}_q \boldsymbol{x}_q^{\intercal}. \tag{2}$$

We can view each term $\boldsymbol{x}_t^{\intercal} \boldsymbol{P}_t \boldsymbol{x}_t$ as round t's contribution to the regret, and hence the $\boldsymbol{P}_t$ can be interpreted as an inverse second moment matrix when the outer products $\boldsymbol{x}_q \boldsymbol{x}_q^{\intercal}$ for unseen data are weighted according to their contribution to the minimax regret. We emphasize that $\boldsymbol{P}_t$ is a product of the minimax analysis; its elegant form showcases the beauty of the minimax approach.

These results apply to a range of minimax regret games, defined according to constraints on the labels of the form $|y_t| \le B_t$, under a condition that the covariates and the constraints are

compatible. Notice that a single strategy (MM) is minimax optimal for any of these constraint sets, so it is adaptive to any scaling of the labels. In Section 4, we investigate another way to stratify the complexity of the regression problem that is more natural. Instead of restricting the labels locally at each time step, we impose a global constraint on a sum of squared labels, each weighted by the relative hardness of the corresponding covariate in the sequence: we fix some $R \geq 0$ and require

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

We show that on this set of label sequences the strategy (MM) is also minimax optimal without any condition on the covariates. This suggests that the above quantity is a natural measure of the complexity of the constellation of labels and covariates for regression.

1.1. Related work

Linear regression is one of the classical problems in statistics and has been studied for over a century. The online version of linear regression is much more recent. Foster (1991) considered online linear regression with binary labels and ℓ_1-constrained parameters $\boldsymbol{w}$, and gave an $O(d \log(dT))$ regret bound for an ℓ_2-regularized follow-the-leader strategy. Cesa-Bianchi et al. (1996) considered ℓ_2-constrained parameters, and gave $O(\sqrt{T})$ regret bounds for a gradient descent algorithm with ℓ_2 regularization. Kivinen and Warmuth (1997) showed that an Exponentiated Gradient algorithm, based on relative entropy regularization, gives $O(\sqrt{T})$ regret. All of these results depend on the scale of the instances and labels. Vovk (1998) applied the Aggregating Algorithm (Vovk, 1990) to continuously many experts to arrive at an algorithm for online linear regression. This algorithm uses the inverse second moment matrix of past and current covariates, whereas the minimax strategy that we present uses the entire covariate sequence (see (2)). Vovk's algorithm was interpreted and re-analyzed in various ways (Forster, 1999; Azoury and Warmuth, 2001): it is minimax optimal for the last trial, and it satisfies a $O(\log T)$ scale-dependent regret bound. The scale dependence is perhaps not surprising when future instances are not available. The regret bound we obtain for the minimax strategy is $O(\log T)$ with no dependence on the scale of the covariates. Refined work on "last-step minimax" was done by Moroshko and Crammer (2014). We take the approach of Takimoto and Warmuth (2000) and Koolen et al. (2014), who studied minimax optimal strategies for prediction games with squared loss: rather than proposing an algorithm that explicitly involves regularization and proving a regret bound, we identify the optimal minimax strategy for square loss; the ideal regularization emerges.

1.2. The offline problem

Lemma 1 *Fix data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T)$. The loss of the best linear predictor in hindsight is*

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

where $\dagger$ denotes pseudo-inverse (any generalized inverse will do). It is minimized by

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

2. Minimax analysis for bounded labels

In this section we perform a minimax analysis of fixed-design linear regression with bounded labels y_t and give an exact expression for the minimax regret. As discussed in the introduction, the following problem-weighted inverse covariate matrices are central to the analysis and algorithm. They are defined recursively starting at T and going backwards:

$$\boldsymbol{P}_T = \left(\sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t^\intercal\right)^\dagger, \qquad \boldsymbol{P}_t = \boldsymbol{P}_{t+1} + \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}. \tag{3}$$

While not immediate, the definition of $\boldsymbol{P}_t$ here agrees with (2); see Lemma 11 in the appendix. In the proof, it becomes clear that the $\boldsymbol{P}_t$ arise exactly from solving the minimax problem.

Theorem 2 *Fix a constant $B > 0$ and a sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in \mathbb{R}^d$. Consider the following T-round game. On round $t \in \{1, \ldots, T\}$, the player first chooses $\hat{y}_t \in \mathbb{R}$, then the adversary chooses $y_t \in [-B, B]$ and the player incurs loss $(\hat{y}_t - y_t)^2$. The value of this game is*

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

Assume that the following covariate condition holds:

$$\sum_{q=1}^{t-1} |\boldsymbol{x}_q^\intercal \boldsymbol{P}_t \boldsymbol{x}_t| \;\le\; 1 \quad \textit{for all } 1 \le t \le T. \tag{4}$$

Then the value of the game is $B^2 \sum_{t=1}^{T} \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$, the optimal strategy is (MM): $\hat{y}_{t+1} = \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t$, where $\boldsymbol{s}_t = \sum_{q=1}^{t} y_q \boldsymbol{x}_q$, and the maximin probability distribution assigns $\Pr(y_{t+1} = \pm B) = 1/2 \pm \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t / (2B)$.

The proof shows that the minimax strategy optimizes the value-to-go, and therefore optimally exploits suboptimal play by the adversary.

Proof We can define the value of the game recursively, via

$$V\left(\boldsymbol{s}_T, \sigma_T^2, T\right) \coloneqq - \min_{\boldsymbol{w} \in \mathbb{R}^d} \left(\sum_{t=1}^{T} \left(\boldsymbol{w}^\intercal \boldsymbol{x}_t - y_t\right)^2\right),$$
$$V\left(\boldsymbol{s}_t, \sigma_t^2, t\right) \coloneqq \min_{\hat{y}_{t+1}} \max_{y_{t+1}} \left(\left(\hat{y}_{t+1} - y_{t+1}\right)^2 + V\left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}, \sigma_t^2 + y_{t+1}^2, t+1\right)\right),$$

where the state $(\boldsymbol{s}_t, \sigma_t^2)$ after the tth round is defined as

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

(and $\boldsymbol{s}_0 = \boldsymbol{0}$, $\sigma_0^2 = 0$). We show by induction that

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

where the γ_t coefficients are recursively defined as

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

This implies that the value of the game is $V(\mathbf{0}, 0, 0) = \gamma_0 = B^2 \sum_{t=1}^T \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$, as desired. Lemma 1 establishes the base case $V\left(\boldsymbol{s}_T, \sigma_T^2, T\right) = \boldsymbol{s}_T^\intercal \boldsymbol{P}_T \boldsymbol{s}_T - \sigma_T^2$. Now, assuming the induction hypothesis

$$V(\boldsymbol{s}_{t+1}, \sigma_{t+1}^2, t+1) = \boldsymbol{s}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_{t+1} - \sigma_{t+1}^2 + \gamma_{t+1},$$

we have

$$\begin{aligned}
V\left(\boldsymbol{s}_t, \sigma_t^2, t\right) &= \min_{\hat{y}_{t+1}} \max_{y_{t+1}} \left(\hat{y}_{t+1} - y_{t+1}\right)^2 + V\left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}, \sigma_t^2 + y_{t+1}^2, t+1\right) \\
&= \min_{\hat{y}_{t+1}} \max_{y_{t+1}} \left(\hat{y}_{t+1} - y_{t+1}\right)^2 + \left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}\right)^\intercal \boldsymbol{P}_{t+1} \left(\boldsymbol{s}_t + y_{t+1}\boldsymbol{x}_{t+1}\right) \\
&\qquad - \left(\sigma_t^2 + y_{t+1}^2\right) + \gamma_{t+1} \\
&= \min_{\hat{y}_{t+1}} \max_{y_{t+1}} \left(\hat{y}_{t+1}^2 - 2\hat{y}_{t+1}y_{t+1} + 2y_{t+1}\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t + y_{t+1}^2 \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\right. \\
&\qquad \left. + \boldsymbol{s}_t^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \sigma_t^2 + \gamma_{t+1}\right. \\
&= \min_{\hat{y}_{t+1}} \hat{y}_{t+1}^2 + \left(\max_{y_{t+1}} 2\left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \hat{y}_{t+1}\right) y_{t+1} + \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1} y_{t+1}^2\right) \\
&\qquad + \boldsymbol{s}_t^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \sigma_t^2 + \gamma_{t+1}.
\end{aligned}$$

The inner maximization is a quadratic in $y_{t+1} \in [-B, B]$ with a non-negative second derivative, so it is maximized by an extreme $y_{t+1} \in \{-B, B\}$, giving

$$\begin{aligned}
V\left(\boldsymbol{s}_t, \sigma_t^2, t\right) = \min_{\hat{y}_{t+1}} &\left(\hat{y}_{t+1}^2 + 2B\left|\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \hat{y}_{t+1}\right|\right) \\
&+ \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}B^2 + \boldsymbol{s}_t^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \sigma_t^2 + \gamma_{t+1}.
\end{aligned}$$

The minimization over $\hat{y}_{t+1}$ is of a convex function, which is minimized when 0 is in the subgradient, so that

$$\hat{y}_{t+1} = \begin{cases} -B & \text{if } \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t < -B, \\ B & \text{if } \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t > B, \\ \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t & \text{otherwise.} \end{cases} \tag{5}$$

Under the assumption (4) of the theorem, only the last case occurs:

$$\left|\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t\right| = \left|\sum_{q=1}^t \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_q y_q\right| \le \sum_{q=1}^t \left|\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_q\right| |y_q| \le B,$$

so we have $\hat{y}_{t+1} = \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t$. Plugging this solution in, we find

$$\begin{aligned}
V\left(\boldsymbol{s}_t, \sigma_t^2, t\right) &= \boldsymbol{s}_t^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t + B^2 \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1} + \boldsymbol{s}_t^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \sigma_t^2 + \gamma_{t+1} \\
&= \boldsymbol{s}_t^\intercal \left(\boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} + \boldsymbol{P}_{t+1}\right) \boldsymbol{s}_t - \sigma_t^2 + \gamma_{t+1} + B^2 \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1},
\end{aligned}$$

verifying the recursion for $\boldsymbol{P}_t$ and γ_t. From the perspective of the adversary, we need to solve

$$\begin{aligned} & \max_{p\in[0,1]} \min_{\hat{y}_{t+1}} \hat{y}_{t+1}^2 + 2\left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - \hat{y}_{t+1}\right)(2p-1)B + \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}B^2 \\ = & \max_{p\in[0,1]} B(2p-1)^2 + 2\left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t - B(2p-1)\right)(2p-1)B + \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}B^2. \end{aligned}$$

because the minimizer of $\hat{y}_{t+1}$ is the mean $B(2p-1)$. Setting the p-derivative to zero results in worst-case probability $1/2 \pm \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t/(2B)$ on $\pm B$. ■

Condition (4) can be easily tested; it does not involve the labels y_t. It can be viewed as forbidding outlier covariates: an x_t that is large relative to the others will cause the condition to fail, leading to clipping in (5). The condition appears to be restrictive: it is satisfied if the covariates are approximately orthonormal, which essentially corresponds to playing d interleaved independent one-dimensional regression problems, but we do not know of other problem instances that satisfy the condition.

The condition arises because of the uniform constraint on the labels. There are, however, many other constraint sets for which the same strategy is still minimax optimal, but the corresponding conditions are milder. In particular, it is clear that the proof extends immediately to the case in which the adversary is constrained to choose label sequences from

$$\mathcal{Y}_B := \{(y_1, \ldots, y_T) : |y_t| \le B_t\}, \tag{6}$$

provided that the $B = (B_1, \ldots, B_T)$ are compatible with the data by satisfying

$$B_t \ge \sum_{q=1}^{t-1} |\boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_q|\, B_q. \tag{7}$$

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 $\pm B_{t+1}$. Condition (4) is a special case of these compatibility constraints (7) corresponding to $B_1 = \cdots = B_T$.

3. Regret bound for worst-case covariates

The previous section establishes that under the covariate condition (4), the minimax regret is equal to $B^2 \sum_{t=1}^T \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$. It is also clear that this remains an upper bound on the minimax regret even when (4) does not hold; the proof computes the regret for the choice $\hat{y}_{t+1} = \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1}\boldsymbol{s}_t$, regardless of whether that choice satisfies the optimality condition (5). We now investigate how the bound behaves. The regret factors into the contribution B^2 for the range of the labels, and the contribution $\sum_{t=1}^T \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$ for the constellation of the covariates. Intriguingly, but rather reasonably, it is invariant under invertible linear transformations of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ (as is Condition (4)). We now focus on the simpler question of how the regret scales with the time horizon T and the dimension d. That is, we maximize the regret with respect to the covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ unconstrained in $\mathbb{R}^d$. We show

$$\max_{\boldsymbol{x}_1, \ldots \boldsymbol{x}_T} \sum_{t=1}^T \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t \;=\; O(d \log T). \tag{8}$$

(Note that this is fixed-design linear regression; the learner sees the chosen covariates in advance.) We proceed in two steps. First, we treat the one dimensional case, in which we establish a $O(\log T)$ bound. Then we reduce the multidimensional case to d independent one-dimensional sub-cases.

3.1. Minimax regret in one dimension

As a precursor to the general vector regret case, we first bound the regret in the scalar case, i.e. $d = 1$ dimension (we will write non-bold C and x). The techniques needed are interesting and naturally generalize to the vector case.

The left-hand side of (8) is unwieldy because the C_t recursion runs backwards, and hence C_t involves all x_q for future $q > t$. To make the expression manageable, we break this cycle. We define

$$\phi_t(B, C) \coloneqq \max \left\{ \sum_{q=1}^{t} x_q^2 C_q : \sum_{q=1}^{t} x_q^2 = B,\, C_t = C,\, \forall t \le t\; C_{t-1} = C_t + x_t^2 C_t^2 \right\},$$

so that the regret of the worst case covariate sequence equals $\max_B \phi_T(B, 1/B)$. We first argue that we can eliminate one parameter from ϕ_t.

Lemma 3 *For all t and $c > 0$, $\phi_t(B, C) = \phi_t(cB, C/c)$.*

Proof We expand and reparameterize

$$\begin{aligned}
\phi_t(B, C) &= \max \left\{ \sum_{q=1}^{t} x_q^2 C_q : \sum_{q=1}^{t} x_q^2 = B,\, C_t = C,\, C_{t-1} = C_t + x_t^2 C_t^2 \right\} \\
&= \max \left\{ \sum_{q=1}^{t} c x_q^2 \frac{C_q}{c} : \sum_{q=1}^{t} c x_q^2 = cB,\, \frac{C_t}{c} = \frac{C}{c},\, \frac{C_{t-1}}{c} = \frac{C_t}{c} + c x_t^2 \left(\frac{C_t}{c}\right)^2 \right\} \\
&= \max \left\{ \sum_{q=1}^{t} x_q^2 C_q : \sum_{q=1}^{t} x_q^2 = cB,\, C_t = \frac{C}{c},\, C_{t-1} = C_t + x_i^2 A_t^2 \right\} = \phi_t\left(cB, \frac{C}{c}\right).
\end{aligned}$$

■

So from now on we restrict attention to the renormalized

$$\phi_t(B) := \phi_t(B, 1).$$

and we want to bound $\phi_T(1)$. Next we exhibit a recurrence relation for ϕ_t.

Lemma 4 *For all $B \ge 0$,*

$$\phi_1(B) = B, \qquad \phi_t(B) = \max\left\{\alpha + \phi_{t-1}\big((\alpha+1)(B-\alpha)\big) : 0 \le \alpha \le B\right\}.$$

Furthermore, for $B \le 1$, the argument does not explode: $\left\{(\alpha+1)(B-\alpha) : 0 \le \alpha \le B\right\} = [0, B]$.

Proof From the definition, $\phi_1(B) = B$.

$$\begin{aligned}
\phi_t(B) &= \max\left\{\sum_{q=1}^{t-1} x_q^2 C_q + x_t^2 : 0 \le x_t^2 \le B,\ \sum_{q=1}^{t-1} x_q^2 \le B - x_t^2,\ C_{t-1} = 1 + x_t^2\right\} \\
&= \max\left\{x_t^2 + \max\left\{\sum_{q=1}^{t-1} x_q^2 C_q : \sum_{q=1}^{t-1} x_q^2 \le B - x_t^2,\ C_{t-1} = 1 + x_t^2\right\} : 0 \le x_t^2 \le B\right\} \\
&= \max\big\{\alpha + \phi_{t-1}(B - \alpha, 1 + \alpha) : 0 \le \alpha \le B\big\} \\
&= \max\left\{\alpha + \phi_{t-1}\big((B - \alpha)(1 + \alpha)\big) : 0 \le \alpha \le B\right\}.
\end{aligned}$$

To see that $(B - \alpha)(1 + \alpha)$ (the argument passed on to ϕ_{t-1}) always stays in $[0, B]$, notice that it is a decreasing function of α, which equals B at $\alpha = 0$ and 0 at $\alpha = B$. ■

We are now ready to prove our main bound on ϕ_T.

Theorem 5 *Fix any function f such that $f(0) \ge 0$ and $e^{-f(T)/2} + f(T) \le f(T+1)$ for all $T \ge 0$. Then*

$$\phi_T(1) \ \le\ f(T).$$

In particular, we have

$$\phi_T(1) \le 1 + 2\ln\left(1 + \frac{T}{2}\right).$$

Proof We prove by induction on T the stronger statement

$$\phi_T(B) \ \le\ \min\left\{-\ln(1 - B), f(T)\right\}$$

The base case $T = 0$ is safe since $\phi_0(B) = 0$ whereas both $-\ln(1 - B)$ and $f(0)$ are positive, the latter by assumption. We proceed with the induction step. Using the definition of ϕ_{T+1} and the induction hypothesis we conclude

$$\begin{aligned}
\phi_{T+1}(B) \ &=\ \max_{0 \le x \le B} \frac{\sqrt{(1 + B)^2 - 4x} - (1 - B)}{2} + \phi_T(x) \\
&\le\ \max_{0 \le x \le B} \frac{\sqrt{(1 + B)^2 - 4x} - (1 - B)}{2} + \min\left\{-\ln(1 - x), f(T)\right\} \qquad (9)
\end{aligned}$$

We now argue that the maximand in (9) changes from increasing to decreasing at x equal to $\hat{x} := 1 - e^{-f(T)}$ ($\hat{x} \in [0, 1]$, possibly $\hat{x} > B$), which is the x for which the minimum in (9) changes from its left to its right argument. For $x < \hat{x}$, we need to show that the derivative of the maximand toward x is positive, i.e.

$$\frac{-1}{\sqrt{(1 + B)^2 - 4x}} + \frac{1}{1 - x} \ \ge\ 0 \qquad \text{that is} \qquad (1 + x)^2 \ \le\ (1 + B)^2$$

which holds for all $x \le B$ of interest. On the other hand, for $x > \hat{x}$, the maximand is obviously decreasing in x. Put together this means that, for $B \le \hat{x}$, (9) is maximized at $x = B$. We find

$$\phi_{T+1}(B) \ \le\ \min\left\{-\ln(1 - B), f(T)\right\}$$

and the induction step is completed in this case because the assumption on f implies that it is increasing. On the other hand for $B \geq \hat{x}$ (9) is maximized at $x = \hat{x}$ and so

$$(9) \;=\; \frac{1}{2}\left(\sqrt{(1+B)^2 - 4\hat{x}} - (1-B)\right) + f(T)$$

The right hand side is increasing in B, and by substituting $B = 1$ we find

$$\phi_{T+1}(B) \;\leq\; \sqrt{1-\hat{x}} + f(T) \;=\; e^{-f(T)/2} + f(T) \;\leq\; f(T+1)$$

where the last step uses the assumption on f. Bounding the $\min$ in (9) by its left argument and choosing $x = B$ establishes $\phi_{T+1}(B) \leq -\ln(1-B)$ as before, completing the induction step.

Finally, choose $f(T) \;=\; 1 + 2\ln\left(1 + \frac{T}{2}\right)$. This $f(T)$ is valid since

$$f(T+1) - f(T) \;=\; -2\ln\left(1 - \frac{1}{T+3}\right) \;\geq\; \frac{2}{T+3} \;>\; e^{-1/2}\frac{2}{T+2} \;=\; e^{-f(T)/2}.$$

■

3.2. Vector case

We now turn to dimension $d > 1$. We start out with the analogous reparameterization by a separate budget $\boldsymbol{B}$ on $\sum_t \boldsymbol{x}\boldsymbol{x}^\intercal$ and starting point $\boldsymbol{P}$ for $\boldsymbol{P}_T$. We define

$$\phi_t(\boldsymbol{B}, \boldsymbol{P}) \coloneqq \max\left\{\sum_{q=1}^{t} \boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q : \sum_{q=1}^{t} \boldsymbol{x}_q\boldsymbol{x}_q^\intercal = \boldsymbol{B}, \boldsymbol{P}_t = \boldsymbol{P}, \boldsymbol{P}_{t-1} = \boldsymbol{P}_t + \boldsymbol{P}_t\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_t\right\},$$

and our goal becomes bounding $\max_{\boldsymbol{B}} \phi_T(\boldsymbol{B}, \boldsymbol{B}^\dagger)$. First, we show that ϕ_t has the following scale invariance property:

Lemma 6 *For any invertible symmetric matrix $\boldsymbol{W}$, $\phi_t(\boldsymbol{B}, \boldsymbol{P}) = \phi_t(\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{W}^{-\frac{1}{2}}, \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{W}^{\frac{1}{2}})$.*

Proof Let $\boldsymbol{x}'_t \coloneqq \boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{x}_t$, $\boldsymbol{P}'_t \coloneqq \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{P}_t\boldsymbol{W}^{\frac{1}{2}}$, $\boldsymbol{P}' \coloneqq \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{W}^{\frac{1}{2}}$ and $\boldsymbol{B}' \coloneqq \boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{W}^{-\frac{1}{2}}$. Then

$$\begin{aligned}
\phi_t(\boldsymbol{B}, \boldsymbol{P}) &= \max\left\{\sum_{q=1}^{t} \boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q : \sum_{q=1}^{t} \boldsymbol{x}_q\boldsymbol{x}_q^\intercal = \boldsymbol{B}, \boldsymbol{P}_t = \boldsymbol{P}, \boldsymbol{P}_{t-1} = \boldsymbol{P}_t + \boldsymbol{P}_t\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_t\right\} \\
&= \max\left\{\sum_{q=1}^{t} \boldsymbol{x}_q'^\intercal \boldsymbol{P}'_q \boldsymbol{x}'_q : \sum_{q=1}^{t} \boldsymbol{x}'_q\boldsymbol{x}_q'^\intercal = \boldsymbol{B}', \boldsymbol{P}'_n = \boldsymbol{P}', \boldsymbol{P}'_{t-1} = \boldsymbol{P}'_t + \boldsymbol{P}'_t\boldsymbol{x}'_t\boldsymbol{x}_t'^\intercal\boldsymbol{P}'_t\right\} \\
&= \phi_t(\boldsymbol{B}', \boldsymbol{P}').
\end{aligned}$$

■

This scale invariance property allows us to focus on the renormalized $\phi_t(\boldsymbol{B}) \;\coloneqq\; \phi_t(\boldsymbol{B}, \boldsymbol{I})$ and the goal becomes to bound $\phi_T(\boldsymbol{I})$. We now derive a recursion for $\phi_t(\boldsymbol{B})$.

$$\begin{aligned}\phi_t(\boldsymbol{B}) &= \max\left\{\sum_{q=1}^{t}\boldsymbol{x}_q^\intercal\boldsymbol{P}_q\boldsymbol{x}_q : \sum_{q=1}^{t}\boldsymbol{x}_q\boldsymbol{x}_q^\intercal = \boldsymbol{B}, \boldsymbol{P}_t = \boldsymbol{I}, \boldsymbol{P}_{t-1} = \boldsymbol{P}_t + \boldsymbol{P}_t\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\boldsymbol{P}_t\right\}\\ &= \max\left\{\boldsymbol{x}^\intercal\boldsymbol{x} + \phi_{t-1}(\boldsymbol{B} - \boldsymbol{x}\boldsymbol{x}^\intercal, \boldsymbol{I} + \boldsymbol{x}\boldsymbol{x}^\intercal) : \boldsymbol{x}\boldsymbol{x}^\intercal \preceq \boldsymbol{B}\right\}\\ &= \max\left\{\boldsymbol{x}^\intercal\boldsymbol{x} + \phi_{t-1}\big((\boldsymbol{I} + \boldsymbol{x}\boldsymbol{x}^\intercal)^{\frac{1}{2}}(\boldsymbol{B} - \boldsymbol{x}\boldsymbol{x}^\intercal)(\boldsymbol{I} + \boldsymbol{x}\boldsymbol{x}^\intercal)^{\frac{1}{2}}\big) : \boldsymbol{x}\boldsymbol{x}^\intercal \preceq \boldsymbol{B}\right\}.\end{aligned}$$

Optimizing over rank-1 updates is difficult. Instead, we consider the simple relaxation obtained by replacing the rank 1 outer product $\boldsymbol{x}\boldsymbol{x}^\intercal$ by a positive semi-definite matrix $\boldsymbol{X}$:

$$\begin{aligned}\psi_0(\boldsymbol{B}) &\coloneqq \phi_0(\boldsymbol{B}) = 0\\ \psi_t(\boldsymbol{B}) &\coloneqq \max\left\{\mathrm{tr}(\boldsymbol{X}) + \psi_{t-1}\big((\boldsymbol{I} + \boldsymbol{X})^{\frac{1}{2}}(\boldsymbol{B} - \boldsymbol{X})(\boldsymbol{I} + \boldsymbol{X})^{\frac{1}{2}}\big) : \boldsymbol{X} \preceq \boldsymbol{B}\right\}.\end{aligned}$$

Obviously $\phi_t(\boldsymbol{B}) \le \psi_t(\boldsymbol{B})$ as ψ_t is a maximum over a larger set of a larger function. As codified in the following lemma, this relaxation makes the calculation of ψ_t much easier.

Lemma 7 *For any $B \ge 0$, $\psi_t(B\boldsymbol{I}) = \sum_{i=1}^{d}\phi_t(B)$, where $\phi_t(B)$ is the one-dimensional regret bound.*

Proof In the base case $t = 0$ both sides are zero. For the inductive hypothesis, assume that $\psi_{t-1}(B'\boldsymbol{I}) = \sum_{i=1}^{d}\phi_{t-1}(B')$. Let us denote the eigenvalues of $\boldsymbol{X}$ by $\alpha_1, \ldots, \alpha_d$. Then

$$\begin{aligned}\psi_t(B\boldsymbol{I}) &= \max\left\{\mathrm{tr}(\boldsymbol{X}) + \psi_{t-1}\big((\boldsymbol{I} + \boldsymbol{X})^{\frac{1}{2}}(B\boldsymbol{I} - \boldsymbol{X})(\boldsymbol{I} + \boldsymbol{X})^{\frac{1}{2}}\big) : \boldsymbol{X} \preceq B\boldsymbol{I}\right\}\\ &= \max\left\{\sum_{i=1}^{d}\alpha_i + \sum_{i=1}^{d}\phi_{t-1}\big((1+\alpha_i)(B-\alpha_i)\big) : 0 \le \alpha_i \le B\,\forall i\right\} = \sum_{i=1}^{d}\phi_t(B).\end{aligned}$$

■

With this factorization, a regret bound is immediate:

Theorem 8

$$\max_{\boldsymbol{x}_1,\ldots\boldsymbol{x}_T}\sum_{t=1}^{T}\boldsymbol{x}_t^\intercal\boldsymbol{P}_t\boldsymbol{x}_t = \phi_T(\boldsymbol{I}) \le d\left(1 + 2\ln\left(1 + \frac{T}{2}\right)\right).$$

4. Minimax analysis for problem-weighted 2-norm bounded label sequences

In this section we investigate another way of budgeting that is suggested by the problem. Namely, for some $R \ge 0$, we consider the set

$$\mathcal{Y}_R \coloneqq \left\{y_1, \ldots, y_T \in \mathbb{R} : \sum_{t=1}^{T} y_t^2\boldsymbol{x}_t^\intercal\boldsymbol{P}_t\boldsymbol{x}_t = R\right\} \tag{10}$$

of label sequences with a certain weighted 2-norm, where the weights are related to the hardness of the covariates. We analyze the minimax fixed design linear regression problem (1) on $\mathcal{Y}_R$, and show that the minimax strategy is again the simple linear strategy (MM). Recall that this strategy predicts

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

This is surprising for two reasons. First, this predictor does not incorporate knowledge of R. Second, there is no easy relation between R and the maximum label magnitude $B_{\max} := \max_t |y_t|$. As the minimax regret bound of Section 3 deteriorates with B^2, one might conjecture that the performance also degenerates. However, to the contrary, we show that the regret of the predictor (MM) now *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.

The regret this algorithm incurs is better than the minimax regret with $B = B_{\max}$ under Condition (4). Still, it inherits the $B_{\max}^2 d \log T$ bound. In addition, minimax optimality for the family of constraints $\mathcal{Y}_R$ is stronger than the corresponding result for the family of box constraints $\mathcal{Y}_B$ defined in (6), in the sense that, given some budget R and a sequence of B_ts that satisfy the compatibility inequalities (7), we can rescale the B_ts so that $\mathcal{Y}_B$ is contained in $\mathcal{Y}_R$, but the minimax regret is the same in both cases.

We proceed in two steps. We characterize the worst-case regret of the simple linear predictor (MM) on the set $\mathcal{Y}_R$. Then we argue that the worst-case regret of any predictor is at least as large.

Lemma 9 *Let $\boldsymbol{P}_t$ be as defined in (3). For all $y_1, \ldots, y_T$, strategy (MM) has regret $\sum_{t=1}^T y_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t$:*

$$\mathcal{R}_T \;=\; \sum_{t=1}^{T} (\hat{y}_t - y_t)^2 - \min_{\boldsymbol{w} \in \mathbb{R}^d} \sum_{t=1}^{T} (\boldsymbol{w}^\intercal \boldsymbol{x}_t - y_t)^2 \;=\; \sum_{t=1}^{T} y_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t . \tag{11}$$

Proof The worst-case (over labels) slack in (11) can be recursively calculated by

$$\begin{aligned} F\left(\boldsymbol{s}_T, \sigma_T^2, T\right) &:= -\min_{\boldsymbol{w} \in \mathbb{R}^d} \sum_{t=1}^{T} \left(\boldsymbol{w}^\intercal \boldsymbol{x}_t - y_t\right)^2 - \sum_{t=1}^{T} y_t^2 \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t , \\ F\left(\boldsymbol{s}_t, \sigma_t^2, t\right) &:= \max_{y_{t+1}} \Big(\left(\hat{y}_{t+1} - y_{t+1}\right)^2 + F\left(\boldsymbol{s}_t + y_{t+1} \boldsymbol{x}_{t+1}, \sigma_t^2 + y_{t+1}^2, t+1\right) \Big) . \end{aligned}$$

Note that the max over $y_1, \ldots, y_T$ of the difference between left and right hand side of (11) is equal to $F(\mathbf{0}, 0, 0)$. We now show by induction that

$$F(\boldsymbol{s}_t, \sigma_t^2, t) = \boldsymbol{s}_t^\intercal \boldsymbol{P}_t \boldsymbol{s}_t - \sigma_t^2 - \sum_{q=1}^{t} y_q^2 \boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q .$$

Lemma 1 verifies the base case. To check the inductive step, we calculate

$$\begin{aligned} F\left(\boldsymbol{s}_t, \sigma_t^2, t\right) &= \max_{y_{t+1}} \Big(\left(\hat{y}_{t+1} - y_{t+1}\right)^2 + F\left(\boldsymbol{s}_t + y_{t+1} \boldsymbol{x}_{t+1}, \sigma_t^2 + y_{t+1}^2, t+1\right) \Big) \\ &= \hat{y}_{t+1}^2 + \max_{y_{t+1}} 2 \left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t - \hat{y}_{t+1}\right) y_{t+1} + \boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{x}_{t+1} y_{t+1}^2 \\ &\qquad + \boldsymbol{s}_t^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t - \sigma_t^2 - \sum_{q=1}^{t+1} y_q^2 \boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q \\ &= \hat{y}_{t+1}^2 + \boldsymbol{s}_t^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t - \sigma_t^2 - \sum_{q=1}^{t} y_q^2 \boldsymbol{x}_q \boldsymbol{P}_q \boldsymbol{x}_q + \max_{y_{t+1}} 2 \left(\boldsymbol{x}_{t+1}^\intercal \boldsymbol{P}_{t+1} \boldsymbol{s}_t - \hat{y}_{t+1}\right) y_{t+1} , \end{aligned}$$

where the max term on the right is zero by the choice of $\hat{y}_{t+1}$. Finally, we obtain

$$F\left(\boldsymbol{s}_t, \sigma_t^2, t\right) = \boldsymbol{s}_t^\intercal \left(\boldsymbol{P}_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\intercal\boldsymbol{P}_{t+1} + \boldsymbol{P}_{t+1}\right)\boldsymbol{s}_t - \sigma_t^2 - \sum_{q=1}^{t} y_q^2 \boldsymbol{x}_q \boldsymbol{P}_q \boldsymbol{x}_q$$

as desired. The theorem statement is immediate upon noting that $F(\mathbf{0}, 0, 0) = 0$. ■

Using this result, we can show that our predictor is minimax optimal.

Theorem 10 *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ be fixed and let $\boldsymbol{P}_t$ be the corresponding prediction matrices. Then for every R, strategy* (MM) *is minimax optimal on the set of labelings $\mathcal{Y}_R$ as defined in* (10)*.*

Proof First, note that strategy (MM) suffers regret R on every y_t sequence in $\mathcal{Y}_R$. Now fix any predictor $\mathfrak{X}$ and consider the label sequence $0, \ldots, 0, \pm\sqrt{\frac{R}{\boldsymbol{x}_T^\intercal \boldsymbol{P}_T \boldsymbol{x}_T}}$, where the sign of the label in the last round is chosen to oppose the sign of the predictor's prediction. Our predictor (MM) predicts the first $T-1$ perfectly and is at least as good on the Tth round. So on *every* round, predictor $\mathfrak{X}$ incurs at least the loss of (MM), and hence its worst-case regret is at least R. Thus, (MM), which incurs regret exactly R, is minimax optimal. ■

5. Conclusion and open problems

We have studied online linear regression in the fixed-design case. We showed that the linear algorithm, $\hat{y}_t = \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{s}_{t-1}$, is minimax optimal for two families of label constraint sets: box-constrained label sequences, provided the covariates satisfy a compatibility condition, and label sequences with bounded problem-weighted ℓ_2 norm. We derived an exact characterization of the regret. Interestingly, it is independent of a rescaling of the $\boldsymbol{x}_t$ sequence. We have also shown a $O(B^2 d \log T)$ bound on the regret.

One question that is raised by Section 3.2 is whether the hardest multidimensional covariates are orthogonal, i.e. a composition of independent one dimensional problems. This would improve the regret bound to $B^2 d \log(T/d)$.

It would also be interesting to understand the gap between the minimax regret and that of strategies like the one in (Vovk, 1998) that have the correct worst case asymptotics.

Still keeping the set of covariates fixed, what order would the learner or adversary choose? Which ordering is most advantageous, and which is hardest? From the analysis for a single dimension in Section 3.1, it seems that processing the covariates in increasing order of magnitude is hardest. How does this generalize to the multivariate case? What is the worst case sequence of covariates when the sequence is not revealed to the learner a priori? Is the minimax analysis tractable, perhaps under some reasonable conditions?

Acknowledgments

We gratefully acknowledge the support of the NSF through grants CCF-1115788 and IIS-1118028, and of the Australian Research Council through an Australian Laureate Fellowship (FL110100281) and through the ARC Centre of Excellence for Mathematical and Statistical Frontiers.

References

Katy S. Azoury and Manfred K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

Nicolò Cesa-Bianchi, Philip M. Long, and Manfred K. Warmuth. Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *Neural Networks, IEEE Transactions on*, 7(3):604–619, 1996.

Jürgen Forster. On relative loss bounds in generalized linear regression. In *Fundamentals of Computation Theory*, pages 269–280. Springer, 1999.

Dean P. Foster. Prediction in the worst case. *Annals of Statistics*, 19(2):1084–1090, 1991.

Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

Wouter M. Koolen, Alan Malek, and Peter L. Bartlett. Efficient minimax strategies for square loss games. In *Advances in Neural Information Processing Systems*, pages 3230–3238, 2014.

Edward Moroshko and Koby Crammer. Weighted last-step min–max algorithm with improved sub-logarithmic regret. *Theoretical Computer Science*, 558:107–124, 2014.

Eiji Takimoto and Manfred K. Warmuth. The minimax strategy for Gaussian density estimation. In *13th COLT*, pages 100–106, 2000.

Volodimir G. Vovk. Aggregating strategies. In *Proc. Third Workshop on Computational Learning Theory*, pages 371–383. Morgan Kaufmann, 1990.

Volodya Vovk. Competitive on-line linear regression. *Advances in Neural Information Processing Systems*, pages 364–370, 1998.

Appendix A. Alternative $\boldsymbol{P}_t$ recurrence

Lemma 11 *For the $\boldsymbol{P}_t$ matrices defined in Theorem 2, we have*

$$\boldsymbol{P}_t^{-1} = \sum_{q=1}^{t} \boldsymbol{x}_q \boldsymbol{x}_q^\intercal + \sum_{q=t+1}^{T} \frac{\boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q}{1 + \boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q} \boldsymbol{x}_q \boldsymbol{x}_q^\intercal .$$

Proof The proof is by induction. We start with

$$\boldsymbol{P}_T^{-1} = \sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t^\intercal .$$

Suppose the equation in the lemma is true for $1 < t \leq T$. Then by the Sherman-Morrison formula,

$$\begin{aligned}
\boldsymbol{P}_{t-1}^{-1} &= (\boldsymbol{P}_t + \boldsymbol{P}_t \boldsymbol{x}_t \boldsymbol{x}_t^\intercal \boldsymbol{P}_t)^{-1} \\
&= \boldsymbol{P}_t^{-1} - \frac{\boldsymbol{x}_t \boldsymbol{x}_t^\intercal}{1 + \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t} \\
&= \sum_{q=1}^{t} \boldsymbol{x}_q \boldsymbol{x}_q^\intercal + \sum_{q=t+1}^{T} \frac{\boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q}{1 + \boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q} \boldsymbol{x}_q \boldsymbol{x}_q^\intercal - \frac{\boldsymbol{x}_t \boldsymbol{x}_t^\intercal}{1 + \boldsymbol{x}_t^\intercal \boldsymbol{P}_t \boldsymbol{x}_t} \\
&= \sum_{q=1}^{t-1} \boldsymbol{x}_q \boldsymbol{x}_q^\intercal + \sum_{q=t}^{T} \frac{\boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q}{1 + \boldsymbol{x}_q^\intercal \boldsymbol{P}_q \boldsymbol{x}_q} \boldsymbol{x}_q \boldsymbol{x}_q^\intercal.
\end{aligned}$$

■