Uniform regret bounds over $\mathbb{R}^d$ for the sequential linear regression problem with the square loss

Pierre Gaillard PIERRE.GAILLARD@INRIA.FR
INRIA, ENS, PSL Research University Paris, France

Sébastien Gerchinovitz SEBASTIEN.GERCHINOVITZ@MATH.UNIV-TOULOUSE.FR
Institut de mathématiques de Toulouse, Université Paul Sabatier, Toulouse

Malo Huard MALO.HUARD@MATH.U-PSUD.FR
Gilles Stoltz GILLES.STOLTZ@MATH.U-PSUD.FR
Laboratoire de mathématiques d'Orsay, Université Paris-Sud, CNRS, Université Paris-Saclay, 91 405 Orsay, France

Editors: Aurélien Garivier and Satyen Kale

Abstract

We consider the setting of online linear regression for arbitrary deterministic sequences, with the square loss. We are interested in the aim set by Bartlett et al. (2015): obtain regret bounds that hold uniformly over all competitor vectors. When the feature sequence is known at the beginning of the game, they provided closed-form regret bounds of $2dB^2 \ln T + \mathcal{O}_T(1)$, where T is the number of rounds and B is a bound on the observations. Instead, we derive bounds with an optimal constant of 1 in front of the $dB^2 \ln T$ term. In the case of sequentially revealed features, we also derive an asymptotic regret bound of $dB^2 \ln T$ for any individual sequence of features and bounded observations. All our algorithms are variants of the online non-linear ridge regression forecaster, either with a data-dependent regularization or with almost no regularization.

Keywords: Adversarial learning, regret bounds, linear regression, (non-linear) ridge regression

1. Introduction and setting

We consider the setting of online linear regression for arbitrary deterministic sequences with the square loss, which unfolds as follows. First, the environment chooses a sequence of observations $(y_t)_{t\geqslant 1}$ in $\mathbb{R}$ and a sequence of feature vectors $(\mathbf{x}_t)_{t\geqslant 1}$ in $\mathbb{R}^d$. The observation sequence $(y_t)_{t\geqslant 1}$ is initially hidden to the learner, while the sequence of feature vectors (see Bartlett et al., 2015) may be given in advance or be initially hidden as well, depending on the setting considered: "beforehand-known features" (also called the fixed-design setting) or "sequentially revealed features". At each forecasting instance $t \geqslant 1$, Nature reveals $\mathbf{x}_t$ (if it was not initially given), then the learner forms a prediction $\widehat{y}_t \in \mathbb{R}$. The observation $y_t \in \mathbb{R}$ is then revealed and instance $t+1$ starts. In all results of this paper, the observations y_t will be assumed to be bounded in $[-B, B]$ (but the forecaster will have no knowledge of B), while we will avoid as much as possible boundedness assumptions of the features $\mathbf{x}_t$. See Figure 1.

Sequentially revealed features	Beforehand-known features
Given: [No input]	**Given:** $\mathbf{x}_1, \dots, \mathbf{x}_T \in \mathbb{R}^d$
For $t = 1, 2, \dots, T$, the learner:	**For** $t = 1, 2, \dots, T$, the learner:
• observes $\mathbf{x}_t \in \mathbb{R}^d$	
• predicts $\widehat{y}_t \in \mathbb{R}$	• predicts $\widehat{y}_t \in \mathbb{R}$
• observes $y_t \in [-B, B]$	• observes $y_t \in [-B, B]$
• incurs $(\widehat{y}_t - y_t)^2 \in \mathbb{R}$	• incurs $(\widehat{y}_t - y_t)^2 \in \mathbb{R}$

Figure 1: The two online linear regression settings considered, introduced by Bartlett et al. (2015); the learner has no knowledge neither of B nor (in the left case) of T.

The goal of the learner is to perform on the long run (when T is large enough) almost as well as the best fixed linear predictor in hindsight. To do so, the learner minimizes her cumulative regret,

$$\mathcal{R}_T(\mathbf{u}) = \sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 - \sum_{t=1}^{T} (y_t - \mathbf{u} \cdot \mathbf{x}_t)^2 \,,$$

either with respect to specific vectors $\mathbf{u} \in \mathbb{R}^d$ (e.g., in a compact subset) or uniformly over $\mathbb{R}^d$. In this article, and following Bartlett et al. (2015), we will be interested in

$$\sup_{\mathbf{u} \in \mathbb{R}^d} \mathcal{R}_T(\mathbf{u}) = \sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 - \inf_{\mathbf{u} \in \mathbb{R}^d} \sum_{t=1}^{T} (y_t - \mathbf{u} \cdot \mathbf{x}_t)^2 \,, \tag{1}$$

which we will refer to as the uniform regret over $\mathbb{R}^d$ (or simply, the uniform regret). The worst-case uniform regret corresponds to the largest uniform regret of a strategy, when considering all possible sequences of features $\mathbf{x}_t$ and (bounded) observations y_t; we will also refer to it as a twice uniform regret, see Section 4.1.

Notation. Bounded sequences of real numbers $(a_t)_{t \geqslant 1}$, possibly depending on external quantities like the feature vectors $\mathbf{x}_1, \mathbf{x}_2, \dots$, are denoted by $a_T = \mathcal{O}_T(1)$. For a given positive function f, the piece of notation $a_T = \mathcal{O}_T\big(f(T)\big)$ then indicates that $a_T / f(T) = \mathcal{O}_T(1)$. Also, the notation $a_T = \Theta_T(1)$ is a short-hand for the facts that $a_T = \mathcal{O}_T(1)$ and $1/a_T = \mathcal{O}_T(1)$, i.e., for the fact that $(a_t)_{t \geqslant 1}$ is bounded from above and from below. We define a similar extension $a_T = \Theta_T\big(f(T)\big)$ meaning that $a_T / f(T) = \Theta_T(1)$.

Earlier works. Linear regression with batch stochastic data has been extensively studied by the statistics community. Our setting of online linear regression for arbitrary sequences is of more recent interest; it dates back to Foster (1991), who considered binary labels $y_t \in \{0, 1\}$ and vectors $\mathbf{u}$ with bounded ℓ_1–norm. We refer the interested reader to the monograph by Cesa-Bianchi and Lugosi (2006, Chapter 11) for a thorough introduction to this literature and to Bartlett et al. (2015) for an overview of the state of the art. Here, we will mostly highlight some key contributions. One is by Vovk (2001) and Azoury and Warmuth (2001): they designed the non-linear ridge regression recalled in Section 2.1, which achieves a regret of order $d \ln T$ only uniformly over vectors $\mathbf{u}$ with bounded ℓ_2–norm. Vovk (2001) also provided a matching minimax lower bound $dB^2 \ln T - \mathcal{O}_T(1)$

on the worst-case uniform regret over $\mathbb{R}^d$ of any forecaster, where B is a bound on the observations $|y_t|$ (see also the lower bound provided by Takimoto and Warmuth, 2000). More recently, Bartlett et al. (2015) computed the minimax regret for the problem with beforehand-known features and provided an algorithm that is optimal under some (stringent) conditions on the sequences $(\mathbf{x}_t)_{t\geqslant 1}$ and $(y_t)_{t\geqslant 1}$ of features and observations. The best closed-form (but general: for all sequences) uniform regret they could obtain for this algorithm was of order $2dB^2 \ln T$. This algorithm is scale invariant with respect to the sequence of features $(\mathbf{x}_t)_{t\geqslant 1}$. Their analysis emphasizes the importance of a data-dependent metric to regularize the algorithm, which is harder to construct when the features are only revealed sequentially. To that end, Malek and Bartlett (2018) show that, under quite intricate constraints on the features and observations, the backward algorithm of Bartlett et al. (2015) can also be computed in a forward (and thus legitimate) fashion in the case when the features are only revealed sequentially. It is thus also optimal; see, e.g., Lemma 39 and Theorem 46 therein.

Organization of the paper and contributions. We first recall and discuss the regret bound of the non-linear ridge regression algorithm (Section 2.1), whose proof will be a building block for our new analyses; we will show that perhaps surprisingly it enjoys a uniform regret bound $2dB^2 \ln T + \mathcal{O}_T(1)$. For the sake of completeness, we also state and re-prove (Section 2.2) the regret lower bound by Vovk (2001) (and Takimoto and Warmuth, 2000), as most of the discussions in this paper will be about the optimal constant in front of the $dB^2 \ln T$ regret bound; this optimal constant will be seen to equal 1. Our proof resorts to a general argument, namely, the van Trees inequality (see Gill and Levit, 1995), to lower bound the error made by any forecaster, while Vovk (2001) was heavily relying on the fact that in a Bayesian stochastic context, the optimal strategy can be determined. This new tool for the machine learning community could be of general interest to derive lower bounds in other settings. We also believe that our lower bound proof is enlightening for statisticians. It shows that the expectation of the regret is larger than a sum of quadratic estimation errors for a d–dimensional parameter. Each of these errors corresponds to an estimation based on a sample of respective length $t-1$, thus is larger than something of the order of d/t, which is the optimal parametric estimation rate. Hence the final $d(1+1/2+\ldots+1/T) \sim d \ln T$ regret lower bound.

We next show (Section 3) that in the case of beforehand-known features, the non-linear ridge regression algorithm and its analysis may make good use of a proper metric $\|\cdot\|_{\mathbf{G}_T}$ described in (8) instead of the Euclidean norm. This leads to a worst-case bound of $dB^2 \ln(1+T/d) + dB^2$ on the uniform regret over $\mathbb{R}^d$, which is optimal (with an optimal constant of 1) in view of the perfectly matching minimax lower bound of Section 2.2. To the best of our knowledge, earlier closed-form worst-case upper bounds were suboptimal by a factor of 2. See the corresponding discussions for the non-linear ridge regression algorithm, in Section 2.1, and for the minimax forecaster by Bartlett et al. (2015), in Remark 8 of Section 3.

The question then is (Section 4) whether a $dB^2 \ln T + \mathcal{O}_T(1)$ regret bound can be achieved on the uniform regret in the most interesting setting of sequentially revealed features. Surprisingly enough, even if the traditional bound for the non-linear ridge regression forecaster blows up when the regularization parameter vanishes, $\lambda = 0$ (see Section 2.1), an ad hoc analysis can be made in this case; it yields a uniform regret bound of $dB^2 \ln T + \mathcal{O}_T(1)$. This bound holds for any fixed sequence of features and bounded observations; we do not impose stringent conditions as in Malek and Bartlett (2018). Also, no parameter needs to be tuned, which is a true relief. The only drawback of this bound, compared to the bounds obtained in the case of beforehand-known features,

is that the $\mathcal{O}_T(1)$ remainder term depends on the sequence of features. We thus could not derive a worst-case uniform regret bound.

Therefore, a final open question is stated in Section 4.1 and consists in determining if such a twice uniform regret bound of order $dB^2 \ln T$ (over comparison vectors $\mathbf{u} \in \mathbb{R}^d$ and over feature vectors $\mathbf{x}_t \in \mathbb{R}^d$ and bounded observations y_t) may hold in the case of sequentially revealed features, or whether the lower bound should be improved. The proofs of the lower bound (the ones by Vovk, 2001, Takimoto and Warmuth, 2000, and our one) generate observations and feature vectors ex ante, independently of the strategy considered, and reveal them to the latter before the prediction game starts. However, it might be the case that truly sequential choices to annoy the strategy considered or generating feature sequences with difficult-to-predict sequences of Gram matrices lead to a larger regret being suffered.

2. Sequentially revealed features / Partially known results

In this section, we recall and reestablish some known results regarding the regret with the square loss function. We recall the definition and the regret bound (Section 2.1) of the non-linear ridge regression algorithm of Vovk (2001), Azoury and Warmuth (2001). The (proof of this) regret bound is used later in this article to design and study our new strategies. We reestablish as well a

$$dB^2\big(\ln T - (3 + \ln d) - \ln\ln T\big)$$

lower bound on the regret of any forecaster (Section 2.2), which implies that the worst-case uniform regret bound $dB^2 \ln(1+T/d) + dB^2$ obtained in Section 3 is first-order optimal: it gets the optimal $dB^2 \ln T$ main term.

2.1. Upper bound on the regret / Reminder of a known result + a new consequence of it

The *non-linear ridge regression algorithm* of Vovk, Azoury and Warmuth uses at each time-step t a vector $\widehat{\mathbf{u}}_t$ such that $\widehat{\mathbf{u}}_1 = (0, \ldots, 0)^{\mathrm{T}}$ and for $t \geqslant 2$,

$$\widehat{\mathbf{u}}_t \in \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^d} \left\{ \sum_{s=1}^{t-1} (y_s - \mathbf{u}\cdot\mathbf{x}_s)^2 + (\mathbf{u}\cdot\mathbf{x}_t)^2 + \lambda \|\mathbf{u}\|^2 \right\}, \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean norm, and predicts $\widehat{y}_t = \widehat{\mathbf{u}}_t \cdot \mathbf{x}_t$. No clipping can take place to form the prediction as the learner has no knowledge of the range $[-B, B]$ of the observations.

Note that the definition (2) is not scale invariant. By scale invariance, we mean that if the $\mathbf{x}_t$ are all multiplied by some $\gamma > 0$ (or even by an invertible matrix Γ), the vector $\widehat{\mathbf{u}}_t$ used should also be just divided by γ (or multiplied by Γ^{-1}). We may also define what a scale-invariant bound on the uniform regret is: a bound that is unaffected by a rescaling of the feature vectors $\mathbf{x}_t$ (as the vectors $\mathbf{u} \in \mathbb{R}^d$ compensate for the rescaling).

Notation 1 *Given features $\mathbf{x}_1, \mathbf{x}_2, \ldots \in \mathbb{R}^d$, we denote by $\mathbf{G}_t = \sum_{s=1}^t \mathbf{x}_s\mathbf{x}_s^{\mathrm{T}}$ the associated $d \times d$ Gram matrix at step $t \geqslant 1$. This matrix is symmetric and positive semidefinite; it admits d eigenvalues, which we sort in non-increasing order and refer to as $\lambda_1(\mathbf{G}_t), \ldots, \lambda_d(\mathbf{G}_t)$. Furthermore, we denote by $r_t = \operatorname{rank}(\mathbf{G}_t)$ the rank of $\mathbf{G}_t$. In particular, $\lambda_{r_t}(\mathbf{G}_t)$ is the smallest positive eigenvalue of G_t.*

For $\lambda > 0$, we have a unique, closed-form solution of (2): denoting $\mathbf{A}_t = \lambda \mathbf{I}_d + \mathbf{G}_t$, which is a symmetric definite positive thus invertible matrix, and $\mathbf{b}_{t-1} = \sum_{s=1}^{t-1} y_t \mathbf{x}_t$,

$$\widehat{\mathbf{u}}_t = \mathbf{A}_t^{-1} \mathbf{b}_{t-1} \,. \tag{3}$$

We recall the proof of the following theorem in Appendix B, mostly for the sake of completeness and because we will use some standard inequalities extracted from it.

Theorem 2 (see Theorem 11.8 of Cesa-Bianchi and Lugosi, 2006) *Let the non-linear ridge regression (2) be run with parameter $\lambda > 0$. For all $T \geqslant 1$, for all sequences $\mathbf{x}_1, \ldots, \mathbf{x}_T \in \mathbb{R}^d$ and all $y_1, \ldots, y_T \in [-B, B]$, for all $\mathbf{u} \in \mathbb{R}^d$,*

$$\mathcal{R}_T(\mathbf{u}) \leqslant \lambda \left\|\mathbf{u}\right\|^2 + B^2 \sum_{k=1}^{d} \ln\left(1 + \frac{\lambda_k(\mathbf{G}_T)}{\lambda}\right).$$

The regret bound above involves a $\lambda \left\|\mathbf{u}\right\|^2$ term, which blows up when the supremum over $\mathbf{u} \in \mathbb{R}^d$ is taken. However, under an additional boundedness assumption on the features $\mathbf{x}_t$, we could prove the following uniform regret bound. To the best of our knowledge, this is the first uniform regret bound proved for this well-known forecaster. Other uniform regret bounds (see Bartlett et al., 2015) were proved for ad-hoc and more involved forecasters, not for a standard, good old forecaster like the non-linear ridge regression (2).

However, despite our best efforts, the uniform regret bound we could prove is only of the form $2dB^2 \ln(T) + \mathcal{O}_T(1)$. It has two drawbacks: first, as we show in the next sections, the constant 2 in the leading term is suboptimal; second, the $\mathcal{O}_T(1)$ strongly depends on the sequence of feature vectors. The proof is provided in Section B and essentially consists in noting that it is unnecessary to worry about vectors $\mathbf{u} \in \mathbb{R}^d$ with too large a norm, as they never achieve the infimum in (1).

Corollary 3 *Let the non-linear ridge regression (2) be run with parameter $\lambda > 0$. For all $T \geqslant 1$, for all sequences $\mathbf{x}_1, \ldots, \mathbf{x}_T \in \mathbb{R}^d$ with $\left\|\mathbf{x}_t\right\| \leqslant X$ and all $y_1, \ldots, y_T \in [-B, B]$,*

$$\sup_{\mathbf{u} \in \mathbb{R}^d} \mathcal{R}_T(\mathbf{u}) \leqslant r_T B^2 \ln\left(1 + \frac{TX^2}{r_T \lambda}\right) + \frac{\lambda}{\lambda_{r_T}(\mathbf{G}_T)} T B^2 \,.$$

Proper choices for λ to minimize the upper bound above are roughly of the order of $1/T$, to get rid of the linear part of the bound given by TB^2; because of the T/λ term in the logarithm, the resulting bound has unfortunately a main term of order $dB^2 \ln T^2 = 2dB^2 \ln T$. For instance, the choice $\lambda = 1/T$, that does not require any beforehand knowledge of the features $\mathbf{x}_t$, together with the bound $r_T \leqslant d$ and the fact that $u \mapsto (1/u)\ln(1+u)$ is decreasing over $(0, +\infty)$, leads to a regret bound less than

$$2dB^2 \ln T + \frac{B^2}{\lambda_{r_T}(\mathbf{G}_T)} + dB^2 \ln\bigl(1 + X^2/d\bigr) \,.$$

The $B^2/\lambda_{r_T}(\mathbf{G}_T)$ quantity in the regret bound is not uniformly bounded over sequences of features $\mathbf{x}_t$. In this respect, Corollary 3 only constitues a minor improvement on Theorem 2. We note a scaling issue: for a fixed sequence of observations $y_1, y_2, \ldots$, while the uniform regret is not affected by a scaling of the feature vectors, the upper bound exhibited above is so. The deep reason for this issue is the lack of invariance of the non-linear ridge regression (2) itself.

2.2. Lower bound on the uniform regret / Improvement on known results

In this section, we study the uniform regret $\sup\{\mathcal{R}_T(\mathbf{u}) : \mathbf{u} \in \mathbb{R}^d\}$, in the minimax case (of beforehand-known features, which is the most difficult setting for a lower bound). That is, we are interested in

$$\mathcal{R}^\star_{T,[-B,B]} \stackrel{\text{def}}{=} \inf_{\text{forecasters}} \sup_{\mathbf{x}_1,\dots,\mathbf{x}_T \in \times[0,1]^d} \sup_{y_t \in [-B,B]} \left\{ \sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \inf_{\mathbf{u}\in\mathbb{R}^d} \sum_{t=1}^T (y_t - \mathbf{u}\cdot\mathbf{x}_t)^2 \right\}, \tag{4}$$

where the first infimum is over all forecasters (all forecasting strategies) that can possibly access beforehand all the features $\mathbf{x}_1, \dots, \mathbf{x}_T$ that are considered next, and the second supremum is over all individual sequences $y_1, \dots, y_T \in [-B, B]$, that are sequentially revealed. Our result is the following; we carefully explain in Remark 6 why this result slightly improves on the existing literature.

Theorem 4 *For all $T \geqslant 8$ and $B > 0$, we have $\mathcal{R}^\star_{T,[-B,B]} \geqslant dB^2\big(\ln T - (3 + \ln d) - \ln\ln T\big)$.*

Remark 5 Note that the features $\mathbf{x}_t$ could be any element of $\mathbb{R}^d$ (by a scaling property on the $\mathbf{u}$), they do not necessarily need to be restricted to $[0,1]^d$; it is merely that our proof relies on such $[0,1]^d$–valued features. Compare to Theorem 7, where no boundedness assumption is required on the features.

Remark 6 Our proof reuses several ideas from the original proof of Vovk (2001, Theorem 2), namely, taking features $\mathbf{x}_t$ with only one non-zero input equal to 1 and Bernoulli observations y_t, resorting to a randomization with a Beta prior distribution, etc.; see also the proof of Takimoto and Warmuth (2000, Theorem 4). However, we believe that we achieve a more satisfactory result than the $(d-\varepsilon)B^2 \ln T - C_\varepsilon$ lower bound of Vovk (2001, Theorem 2), where $\varepsilon > 0$ is a parameter and C_ε is a finite value; also, the proof technique somewhat relied on the boundedness of the features to derive the general case $d \geqslant 2$ from the special case $d = 1$. See also similar results for the case $d = 1$ in Takimoto and Warmuth (2000, Theorem 4), with the same issue for the generalization to $d \geqslant 2$. Our proof, on the contrary, directly tackles the d–dimensional case, which turns out to be more efficient (and more elegant). However, our alternative proof for the lower bound is admittedly a minor variation of existing results, it merely sheds a slightly different light on the bound, see the interpretation below in terms of parametric estimation rate.

The high-level idea of our proof of this known bound is to see the desired $d \ln T$ bound as a sum of parametric estimation errors in $\mathbb{R}^d$, each of order at least d/t. It is a classic result in parametric statistics that the estimation of a d–dimensional parameter based on a sample of size t can be performed at best at rate d/t in quadratic error, and this is exactly what is used in our proof. Vovk (2001, Theorem 2) was heavily relying on the fact that in a Bayesian stochastic context, the optimal strategy can be determined: his proof states that "since Nature's strategy is known, it is easy to find the best, on the average, strategy for Statistician (the Bayesian strategy)." In contrast, our argument does not require to explicitly compute the optimal strategy. It relies on the van Trees inequality (see Gill and Levit, 1995), that lower bounds the estimator error of any, possibly biased, forecaster—unlike the Cramér-Rao bound, which only holds for unbiased estimators. In this respect, the van Trees inequality could reveal itself a new tool of general interest for the machine learning community to derive lower bounds in other settings.

Proof (sketch) The complete proof can be bound in Appendix A and we merely indicate here its most salient arguments. We start with a case where $y_t \in [0,1]$ and explain later how to draw the result for the desired case where $y_t \in [-B,B]$.

We fix any forecaster. A sequence $J_1, \ldots, J_T$ is drawn independently and uniformly at random over $\{1,\ldots,d\}$ and we associate with it the sequence of feature vectors $\mathbf{e}_{J_1}, \ldots, \mathbf{e}_{J_T}$, where $\mathbf{e}_j$ denotes the unit vector $(0,\ldots,0,1,0,\ldots,0)^{\mathrm{T}}$ along the j-th coordinate (the 1 is in position j). The forecaster is informed of this sequence of feature vectors and the sequential prediction problem starts. We actually consider several prediction problems, each indexed by $\theta^\star \in [0,1]^d$: conditionally on the feature vectors $\mathbf{e}_{J_1}, \ldots, \mathbf{e}_{J_T}$, at each round t the observation Y_t is drawn independently according to a Bernoulli distribution with parameter $\theta^\star \cdot \mathbf{e}_{J_t} = \theta^\star_{J_t}$. Expectations with respect to the randomization thus defined will be denoted by $\mathbb{E}_{\theta^\star}$.

Now, given the features considered above, that are unit vectors, each forecasting strategy can be termed as picking only linear combinations $\widehat{y}_t = \widehat{\mathbf{u}}_t \cdot \mathbf{e}_{J_t}$ as predictions. Indeed, we denote by $\widehat{y}_t(j)$ the prediction output by the strategy when $J_t = j$ given the past observations $Y_1, \ldots, Y_s$ and the features $J_1, \ldots, J_T$. We then consider the vector $\widehat{\mathbf{u}}_t \in \mathbb{R}^d$ whose j–th component equals $\widehat{u}_{j,t} = \widehat{y}_t(j)$. This way, in our specific stochastic setting, outputting direct predictions $\widehat{y}_t$ of the observations or outputting vectors $\widehat{\mathbf{u}}_t \in \mathbb{R}^d$ to form linear combinations are the same thing.

The sketchy part of the proof starts here (again, details can be found in Appendix A). By exchanging an expectation and an infimum and by repeated uses of the tower rule, we have, for each $\theta^\star \in [0,1]^d$:

$$\mathbb{E}_{\theta^\star}\left[\sup_{\mathbf{u}\in\mathbb{R}^d} \mathcal{R}_T(\mathbf{u})\right] \geqslant \sum_{t=1}^T \mathbb{E}_{\theta^\star}\left[(Y_t - \widehat{y}_t)^2\right] - \inf_{\mathbf{u}\in\mathbb{R}^d} \sum_{t=1}^T \mathbb{E}_{\theta^\star}\left[(Y_t - \mathbf{u}\cdot\mathbf{e}_{J_t})^2\right] \geqslant \sum_{t=1}^T \frac{1}{d}\,\mathbb{E}_{\theta^\star}\left[\left\|\,\widehat{\mathbf{u}}_t - \theta^\star\,\right\|_2^2\right]. \tag{5}$$

The inequality above being valid for all forecasters accessing in advance to the entire sequence of feature vectors, we thus proved, for any prior π over $[0,1]^d$,

$$\mathcal{R}^\star_{T,\,[0,1]} \geqslant \inf_{\text{forecasters}} \sum_{t=1}^T \int_{[0,1]^d} \frac{1}{d}\,\mathbb{E}_{\theta^\star}\left[\left\|\,\widehat{\mathbf{u}}_t - \theta^\star\,\right\|_2^2\right] \mathrm{d}\pi(\theta^\star)\,.$$

An immediate application of the (multi-dimentional) van Trees inequality with a Beta(α,α) prior π shows that for all forecasters, all $t \geqslant 1$ and $\alpha \geqslant 3$,

$$\int_{[0,1]^d} \frac{1}{d}\,\mathbb{E}_{\theta^\star}\left[\left\|\,\widehat{\mathbf{u}}_t - \theta^\star\,\right\|_2^2\right] \mathrm{d}\pi(\theta^\star) \geqslant \frac{d}{4t + 2t/(\alpha-1) + 16d\alpha}\,,$$

which is roughly of order $d/(4t)$, as we will take large values of α (of order $\ln T$). Straightforward calculations conclude the proof and lead to a lower bound on $\mathcal{R}^\star_{T,\,[0,1]}$ of order $(d/4)\ln T$, which entails a lower bound of order $dB^2 \ln T$ on $\mathcal{R}^\star_{T,\,[-B,B]}$. ■

3. Beforehand-known features / New result

In this section we assume that the features are known beforehand and exhibit a simple forecaster with a closed-form regret bound of $dB^2 \ln T + \mathcal{O}_T(1)$ uniformly over $\mathbb{R}^d$ and all sequences of

features and of bounded observations. Combined with the minimax lower bound of Theorem 4, this upper bound implies that the minimax regret for beforehand-known features has a leading term exactly equal to $dB^2 \ln T$. It thus closes a gap between $dB^2 \ln T$ and $2dB^2 \ln T$ left open by earlier closed-form results such as those of Bartlett et al. (2015, Theorem 8). See a more detailed discussion below, in Remark 8.

The *non-linear ridge regression algorithm with adapted regularization* will pick weight vectors as follows: $\widehat{\mathbf{u}}_1 = (0, \dots, 0)^{\mathrm{T}}$ and for $t \geqslant 2$,

$$\widehat{\mathbf{u}}_t \in \operatorname*{argmin}_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{s=1}^{t-1} (y_s - \mathbf{u} \cdot \mathbf{x}_s)^2 + (\mathbf{u} \cdot \mathbf{x}_t)^2 + \lambda \sum_{s=1}^{T} (\mathbf{u} \cdot \mathbf{x}_s)^2 \right\} \tag{6}$$

with the constraint that $\widehat{\mathbf{u}}_t$ should be of minimal norm within all vectors of the stated argmin. It then predicts $\widehat{y}_t = \widehat{\mathbf{u}}_t \cdot \mathbf{x}_t$. As shown in Appendix C.2, the closed-form expression for $\widehat{\mathbf{u}}_t$ reads

$$\widehat{\mathbf{u}}_t = \big(\lambda \, \mathbf{G}_T + \mathbf{G}_t\big)^{\dagger} \mathbf{b}_{t-1} \,, \tag{7}$$

where $\dagger$ denotes the Moore-Penrose inverse of a matrix (see Appendix E.3) and where $\mathbf{b}_{t-1}$ was defined in (3).

The difference to (2) lies in the regularization term, which can be denoted by

$$\lambda \, \|\mathbf{u}\|^2_{\mathbf{G}_T} \stackrel{\text{def}}{=} \lambda \, \mathbf{u}^{\mathrm{T}} \mathbf{G}_T \mathbf{u} = \lambda \sum_{s=1}^{T} (\mathbf{u} \cdot \mathbf{x}_s)^2 \,; \tag{8}$$

that is, this regularization term can be seen as a metric adapted to the known-in-advance features $\mathbf{x}_1, \dots, \mathbf{x}_T$. This algorithm has the desirable property of being scale invariant. Actually, as will be clear from the equality (12) in the proof of Theorem 7, the strategy (6) considered here consists of "whitening" the feature vectors $\mathbf{x}_t$ into $\widetilde{\mathbf{x}}_t = \mathbf{G}_T^{-1/2} \mathbf{x}_t$ and applying the "classic" non-linear ridge regression (2) to these whitened feature vectors $\widetilde{\mathbf{x}}_t$. Their associated Gram matrix is the identity, which helps obtaining a sharper bound from Theorem 2 than the suboptimal but general bound obtained in Corollary 3.

Theorem 7 *Let the non-linear ridge regression algorithm with adapted regularization (6) be run with parameter $\lambda > 0$. For all $T \geqslant 1$, for all feature sequences $\mathbf{x}_1, \dots, \mathbf{x}_T \in \mathbb{R}^d$ and all $y_1, \dots, y_T \in [-B, B]$,*

$$\sup_{\mathbf{u} \in \mathbb{R}^d} \mathcal{R}_T(\mathbf{u}) \leqslant \lambda T B^2 + r_T B^2 \ln\Big(1 + \frac{1}{\lambda}\Big) \,,$$

where $r_T = \operatorname{rank}(\mathbf{G}_T)$.

By taking $\lambda = r_T/T$, we get the bound $r_T B^2 \big(1 + \ln(1 + T/r_T)\big)$. Of course, $r_T \leqslant d$ and since $u \mapsto (1/u)\ln(1+u)$ is decreasing over $(0, +\infty)$, the final optimized regret bound reads

$$\sup_{\mathbf{u} \in \mathbb{R}^d} \mathcal{R}_T(\mathbf{u}) \leqslant B^2 \left(r_T \ln\left(1 + \frac{T}{r_T}\right) + r_T \right) \leqslant d \, B^2 \ln\left(1 + \frac{T}{d}\right) + d \, B^2 \,.$$

Note that the leading constant is 1, which is known to be optimal because of Theorem 4.

Remark 8 Bartlett et al. (2015) study some minimax uniform regret, namely

$$\mathcal{R}_T^\star = \sup_{\substack{\mathbf{x}_1,\dots,\mathbf{x}_T\in\mathbb{R}^d \\ \text{satisfying (9)}}} \inf_{\widehat{y}_1} \sup_{y_1\in[-B,B]} \cdots \inf_{\widehat{y}_T} \sup_{y_T\in[-B,B]} \sup_{\mathbf{u}\in\mathbb{R}^d} \mathcal{R}_T(\mathbf{u})\,,$$

and design a forecaster called MM based on backward induction. It uses vectors $\widehat{\mathbf{u}}_t = \mathbf{P}_t\mathbf{b}_{t-1}$ for $t \geqslant 2$, where the sequence $\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_T$ is defined in a backward manner as

$$\mathbf{P}_T = \mathbf{G}_T^\dagger \qquad \text{and} \qquad \mathbf{P}_{t-1} = \mathbf{P}_t + \mathbf{P}_t\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}\mathbf{P}_t\,.$$

Because MM is minimax optimal if the (stringent) conditions

$$\forall t \in \{1,\dots,T\}, \qquad \sum_{s=1}^{t-1} \left|\mathbf{x}_s{}^{\mathrm{T}}\mathbf{P}_t^\dagger\mathbf{x}_t\right| \leqslant 1\,, \tag{9}$$

on the feature sequence $\mathbf{x}_1, \dots, \mathbf{x}_T$ are met, a consequence of Theorem 7 is that MM also satisfies the regret bound $dB^2\big(1+\ln(1+T/d)\big)$ for those feature sequences. Bartlett et al. (2015, Theorem 8) showed a regret bound with a leading term of $2dB^2\ln T$. This closed-form bound actually also held for any sequence of features, not only the ones satisfying (9), but it is suboptimal by a multiplicative factor of 2. See Appendix C.1 for further technical details. We do not know whether this suboptimal bound is unavoidable (i.e., is due to the algorithm itself) or whether a different analysis could lead to a better bound for the MM forecaster on sequences not satisfying (9).

Remark 9 It is worth to notice that our result holds in a less restrictive setting than beforehand-known features. Indeed, in the definition of the weight vector $\widehat{\mathbf{u}}_t$, see Equations (6) and (8), the only forward information used lies in the regularization term $\lambda\,\mathbf{u}^{\mathrm{T}}\mathbf{G}_T\mathbf{u}$. Therefore, our algorithm does not need to know the whole sequence of features $\mathbf{x}_1, \dots, \mathbf{x}_T$ in advance: it is enough to know the Gram matrix $\mathbf{G}_T$, in which case our results still hold true. A particular case is when the sequence of features is only known beforehand up to an unknown (and possibly random) permutation, as considered, e.g., by Kotłowski et al. (2017).

Proof In order to keep things simple, we will assume here that $\mathbf{G}_T$ is full rank; the proof in the general case can be found in Appendix C.2. Then, all matrices $\lambda\,\mathbf{G}_T + \mathbf{G}_t$ are full rank as well.

The proof of this theorem relies on the bound of the non-linear ridge regression algorithm of Section 2.1, applied on a modified sequence of features

$$\widetilde{\mathbf{x}}_t = \mathbf{G}_T^{-1/2}\mathbf{x}_t\,,$$

where $\mathbf{G}_T^{-1/2}$ is the inverse square root of the of the symmetric matrix $\mathbf{G}_T$. We successively prove the following two inequalities (where we replaced r_T by its value d, as $\mathbf{G}_T$ is full rank),

$$\sum_{t=1}^T (y_t - \widehat{\mathbf{u}}_t\cdot\mathbf{x}_t)^2 \leqslant \inf_{\mathbf{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^T (y_t - \mathbf{u}\cdot\widetilde{\mathbf{x}}_t)^2 + \lambda\,\|\mathbf{u}\|^2\right\} + dB^2\ln\left(1+\frac{1}{\lambda}\right) \tag{10}$$

$$\leqslant \inf_{\mathbf{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^T (y_t - \mathbf{u}\cdot\mathbf{x}_t)^2\right\} + \lambda T B^2 + dB^2\ln\left(1+\frac{1}{\lambda}\right). \tag{11}$$

Proof of (10). We first show that the strategy (2) on the $\widetilde{\mathbf{x}}_t$ leads to the same forecasts as the strategy (6) on the original $\mathbf{x}_t$; that is, we show that

$$\widetilde{\mathbf{u}}_t \cdot \widetilde{\mathbf{x}}_t = \widehat{\mathbf{u}}_t \cdot \mathbf{x}_t \,, \qquad \text{where} \qquad \widetilde{\mathbf{u}}_t \in \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^d} \left\{ \sum_{s=1}^{t-1} (y_s - \mathbf{u}\cdot\widetilde{\mathbf{x}}_s)^2 + (\mathbf{u}\cdot\widetilde{\mathbf{x}}_t)^2 + \lambda \left\|\mathbf{u}\right\|^2 \right\}. \tag{12}$$

The equality above follows from the definition $\widetilde{\mathbf{x}}_t = \mathbf{G}_T^{-1/2}\mathbf{x}_t$ and the fact that $\widetilde{\mathbf{u}}_t = \mathbf{G}_T^{1/2}\widehat{\mathbf{u}}_t$. Indeed, the closed-form expression (3) indicates that

$$\widetilde{\mathbf{u}}_t = \left(\lambda\,\mathbf{I}_d + \sum_{s=1}^{t} \widetilde{\mathbf{x}}_s\widetilde{\mathbf{x}}_s^{\mathrm{T}}\right)^{-1} \sum_{s=1}^{t-1} y_s\widetilde{\mathbf{x}}_s = \left(\lambda\mathbf{I}_d + \mathbf{G}_T^{-1/2}\mathbf{G}_t\mathbf{G}_T^{-1/2}\right)^{-1}\mathbf{G}_T^{-1/2}\mathbf{b}_{t-1}\,.$$

Now,

$$\left(\lambda\mathbf{I}_d + \mathbf{G}_T^{-1/2}\mathbf{G}_t\mathbf{G}_T^{-1/2}\right)^{-1} = \left(\mathbf{G}_T^{-1/2}\big(\lambda\mathbf{G}_T + \mathbf{G}_t\big)\mathbf{G}_T^{-1/2}\right)^{-1} = \mathbf{G}_T^{1/2}\big(\lambda\mathbf{G}_T + \mathbf{G}_t\big)^{-1}\mathbf{G}_T^{1/2}\,,$$

so that

$$\widetilde{\mathbf{u}}_t = \mathbf{G}_T^{1/2}\big(\lambda\mathbf{G}_T + \mathbf{G}_t\big)^{-1}\mathbf{G}_T^{1/2}\mathbf{G}_T^{-1/2}\mathbf{b}_{t-1} = \mathbf{G}_T^{1/2}\big(\lambda\mathbf{G}_T + \mathbf{G}_t\big)^{-1}\mathbf{b}_{t-1} = \mathbf{G}_T^{1/2}\widehat{\mathbf{u}}_t\,.$$

We apply the bound of Theorem 2 on sequences $\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_T \in \mathbb{R}^d$ and $y_1, \ldots, y_T \in [-B, B]$, to get, for all $\mathbf{u} \in \mathbb{R}^d$,

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 = \sum_{t=1}^{T}(y_t - \widetilde{\mathbf{u}}_t\cdot\widetilde{\mathbf{x}}_t)^2 \leqslant \sum_{t=1}^{T}(y_t - \mathbf{u}\cdot\widetilde{\mathbf{x}}_t)^2 + \lambda\left\|\mathbf{u}\right\|^2 + B^2\sum_{k=1}^{d}\ln\left(1 + \frac{\lambda_k\left(\sum_{t=1}^{T}\widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t^{\mathrm{T}}\right)}{\lambda}\right). \tag{13}$$

The Gram matrix of the $\widetilde{\mathbf{x}}_t$ equals

$$\sum_{t=1}^{T}\widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t^{\mathrm{T}} = \mathbf{G}_T^{-1/2}\left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t{}^{\mathrm{T}}\right)\mathbf{G}_T^{-1/2} = \mathbf{G}_T^{-1/2}\,\mathbf{G}_T\,\mathbf{G}_T^{-1/2} = \mathbf{I}_d\,, \tag{14}$$

so that

$$\sum_{k=1}^{d}\ln\left(1 + \frac{\lambda_k\left(\sum_{t=1}^{T}\widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t^{\mathrm{T}}\right)}{\lambda}\right) = d\ln\left(1 + \frac{1}{\lambda}\right).$$

Taking the infinimum over $\mathbf{u}$ in $\mathbb{R}^d$ in (13) concludes the proof of (10).

Proof of (11). We bound

$$\inf_{\mathbf{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}(y_t - \mathbf{u}\cdot\widetilde{\mathbf{x}}_t)^2 + \lambda\left\|\mathbf{u}\right\|^2\right\},$$

by evaluating it at $\mathbf{u}^\star \in \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}(y_t - \mathbf{u}\cdot\widetilde{\mathbf{x}}_t)^2\right\}$, which is a singleton with closed-form expression

$$\mathbf{u}^\star = \left(\sum_{t=1}^T \widetilde{\mathbf{x}}_t \widetilde{\mathbf{x}}_t^{\mathrm{T}}\right)^{-1} \left(\sum_{t=1}^T y_t \widetilde{\mathbf{x}}_t\right) = \mathbf{G}_T^{-1/2} \mathbf{b}_T \,,$$

where we used (14) and where $\mathbf{b}_T$ was defined in (3). We first bound $\|\mathbf{u}^\star\|^2$. By denoting

$$\mathbf{X}_T = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_T \end{bmatrix} \qquad \text{and} \qquad \mathbf{y}_T = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix},$$

which are respectively, a $d \times T$ and a $T \times 1$ matrix, we have

$$\mathbf{u}^\star = \mathbf{G}_T^{-1/2} \mathbf{X}_T \mathbf{y}_T \,, \qquad \text{thus} \qquad \|\mathbf{u}^\star\|^2 = \mathbf{y}_T^{\mathrm{T}} \mathbf{X}_T^{\mathrm{T}} \mathbf{G}_T^{-1} \mathbf{X}_T \mathbf{y}_T \,. \tag{15}$$

Noting that $\mathbf{X}_T^{\mathrm{T}} \mathbf{G}_T^{-1} \mathbf{X}_T$ is an orthogonal projection (on the image of $\mathbf{X}_T^{\mathrm{T}}$) entails the inequalities $\|\mathbf{u}^\star\|^2 \leqslant \|\mathbf{y}_T\|^2 \leqslant TB^2$. Putting all elements together, we proved so far

$$\inf_{\mathbf{u}\in\mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u}\cdot\widetilde{\mathbf{x}}_t)^2 + \lambda \|\mathbf{u}\|^2 \right\} \leqslant \inf_{\mathbf{u}\in\mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u}\cdot\widetilde{\mathbf{x}}_t)^2 \right\} + \lambda T B^2 \,.$$

We conclude the proof of (11) by a change of dummy variable $\mathbf{v} = \mathbf{G}_T^{1/2}\mathbf{u}$ and the fact that since $\mathbf{G}_T$ is full rank, its image is $\mathbb{R}^d$:

$$\inf_{\mathbf{u}\in\mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u}\cdot\widetilde{\mathbf{x}}_t)^2 \right\} = \inf_{\mathbf{u}\in\mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{G}_T^{1/2}\mathbf{u}\cdot\mathbf{x}_t)^2 \right\} = \inf_{\mathbf{v}\in\mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{v}\cdot\mathbf{x}_t)^2 \right\} . \quad \blacksquare$$

4. Sequentially revealed features / New result

In this section we do not assume that the features are known beforehand (i.e., unlike in the previous section) and yet exhibit a simple forecaster with a regret bound of $dB^2 \ln T + \mathcal{O}_T(1)$ holding uniformly over $\mathbb{R}^d$. Perhaps unexpectedly, the solution that we propose is just to remove the regularization term $\lambda \|\mathbf{u}\|^2_{\mathbf{G}_T}$ in (6), which cannot be computed in advance. This amounts to considering the standard non-linear ridge regression algorithm (2) with a regularization factor $\lambda = 0$. The reason why this is a natural choice is explained in Remark 13 below.

Thus, weights vectors defined as in Equations (2) or (6) with regularization parameter $\lambda = 0$ are picked: $\widehat{\mathbf{u}}_1 = (0, \ldots, 0)^{\mathrm{T}}$ and for $t \geqslant 2$,

$$\widehat{\mathbf{u}}_t \in \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^d} \left\{ \sum_{s=1}^{t-1} (y_s - \mathbf{u}\cdot\mathbf{x}_s)^2 + (\mathbf{u}\cdot\mathbf{x}_t)^2 \right\}, \qquad \text{hence} \qquad \widehat{\mathbf{u}}_t = \mathbf{G}_t^\dagger \mathbf{b}_{t-1} \,, \tag{16}$$

where the closed-form expression corresponds to (7). It then predicts $\widehat{y}_t = \widehat{\mathbf{u}}_t \cdot \mathbf{x}_t$ (and as already indicated after (2), no clipping can take place as B is unknown to the learner).

Note that no parameter requires to be tuned in this case, which can be a true relief.

Remark 10 The traditional bound for the non-linear ridge regression forecaster blows up when the regularization parameter is set as $\lambda = 0$ (see Section 2.1) but, perhaps surprisingly, an ad hoc analysis could be performed here—see Theorem 11. It provides a new understanding of this well-known non-linear regression algorithm: the regularization term $\lambda \|\mathbf{u}\|^2$ in its defining equation (2) is not so useful, while the seemingly harmless regularization term $(\mathbf{u} \cdot \mathbf{x}_t)^2$ therein is crucial.

The theorem below follows from a combination of arguments all already present in the literature, namely, (Forster and Warmuth, 2003, Theorem 3.2), Cesa-Bianchi et al. (2005, Lemma D.1), Luo et al. (2016, Theorem 4 of Appendix D), with a slightly more careful analysis at only one small point in the proof of the latter; see details in Appendix D. The proof is actually based on the proof of Theorem 2 but requires adaptations to account for the fact that $\widehat{\mathbf{u}}_t$ is defined in (16) in terms of a possibly non-invertible matrix $\mathbf{G}_t$. There are strong links between the results of Theorem 11 and Theorem 2; see Remark 12 below.

The result of Theorem 11 is not that straightforward, and in particular, some tricks that were suggested to us when presenting this work, e.g., neglecting finitely many rounds till the Gram matrix is full rank (if this ever happens), would probably work but would lead to an even larger constant term. Generally speaking, neglecting finitely many rounds may have important side-effects, see an illustration in Remark 12.

Theorem 11 *For all $T \geqslant 1$, for all sequences $\mathbf{x}_1, \dots, \mathbf{x}_T \in \mathbb{R}^d$ and all $y_1, \dots, y_T \in [-B, B]$, the non-linear ridge regression algorithm with $\lambda = 0$ as in (16) achieves the uniform regret bound*

$$\sup_{\mathbf{u} \in \mathbb{R}^d} \mathcal{R}_T(\mathbf{u}) \leqslant B^2 \sum_{t=1}^{T} \mathbf{x}_t^{\mathsf{T}} \mathbf{G}_t^{\dagger} \mathbf{x}_t \leqslant B^2 \sum_{k=1}^{r_T} \ln\bigl(\lambda_k(\mathbf{G}_T)\bigr) + B^2 \sum_{t \in [\![1,T]\!] \cap \mathcal{T}} \ln\left(\frac{1}{\lambda_{r_t}(\mathbf{G}_t)}\right) + r_T B^2$$

where r_t and λ_k are defined in Notation 1, and where the set $\mathcal{T}$ contains r_T rounds, given by the smallest $s \geqslant 1$ such that $\mathbf{x}_s$ is not null, and all the $s \geqslant 2$ for which $\operatorname{rank}(\mathbf{G}_{s-1}) \neq \operatorname{rank}(\mathbf{G}_s)$.

We recall in Appendix E.1 that $\operatorname{rank}(\mathbf{G}_t)$ is a non-decreasing sequence, with increments of 1, hence the claimed cardinality r_T of $\mathcal{T}$, and the fact that $\lambda_{r_t}(\mathbf{G}_t) > 0$ for all $t \in \mathcal{T}$.

Note that the regret bound obtained is scale invariant, which is natural and was expected, as the forecaster also is; to see why this is the case, note that it only involves quantities $\lambda_k(\mathbf{G}_T)/\lambda_{r_t}(\mathbf{G}_t)$.

The same (standard) arguments as the ones at the end of the proof of Corollary 3 show the following consequence of this bound (which is scale invariant as far as multiplications of the features by scalar factors only are concerned): for all $X > 0$, for all sequences $\mathbf{x}_1, \mathbf{x}_2, \dots$ of features with $\|\mathbf{x}_t\| \leqslant X$,

$$\sup_{\mathbf{u} \in \mathbb{R}^d} \mathcal{R}_T(\mathbf{u}) \leqslant dB^2 \ln T + \underbrace{dB^2 + B^2 \sum_{t \in [\![1,T]\!] \cap \mathcal{T}} \ln\left(\frac{X^2}{\lambda_{r_t}(\mathbf{G}_t)}\right)}_{\text{this is our } \mathcal{O}_T(1) \text{ here}} .$$

Note that the $\mathcal{O}_T(1)$ term stops increasing once the matrix $\mathbf{G}_T$ is full rank: then, for rounds $T' \geqslant T$, only the leading term increases to $dB^2 \ln T'$. But this $\mathcal{O}_T(1)$ can admittedly be large and blows up as all sequences of feature vectors are considered, just as in the bound of Corollary 3. The dependency on the eigenvalues is however slightly improved to a logarithmic one, here.

The bound of Theorem 11 thus still remains somewhat weak, hence our main open question.

4.1. Open question—Double uniformity over $\mathbb{R}^d$: for the $\mathbf{u}$ and for the $\mathbf{x}_t$

For the time being, no $dB^2\ln(T)+\mathcal{O}_T(1)$ regret bound simultaneously uniform over all comparison vectors $\mathbf{u}\in\mathbb{R}^d$ and over all features $\mathbf{x}_t$ with $\|\mathbf{x}_t\|\leqslant X$ and bounded observations $y_t\in[-B,B]$ is provided in the case of sequentially revealed features (what we called worst-case uniform regret bounds). Indeed, the bound of Theorem 2 is not uniform over the comparison vectors $\mathbf{u}\in\mathbb{R}^d$. The bound of Corollary 3 is of order $2dB^2\ln(T)$ (and is not uniform over even bounded feature vectors). The bound of Theorem 7 enjoys the double uniformity and is of proper order, but only holds for beforehand-known features. The bound of Theorem 11 is uniform over the comparison vectors $\mathbf{u}\in\mathbb{R}^d$, is of proper order $dB^2\ln(T)$ and holds in the sequential case, but is not uniform over bounded features $\mathbf{x}_t$ (its remainder term can be large).

The lower bound of Theorem 4 (and earlier lower bounds by Vovk, 2001 and Takimoto and Warmuth, 2000) are proved in the case of feature vectors that are initially revealed to the regression strategy. The open question is therefore whether we can improve the lower bound and make it larger for strategies that only discover the features on the fly, or if a doubly uniform regret upper bound of $dB^2\ln(T)+\mathcal{O}_T(1)$ over the $\mathbf{u}$ and the $\mathbf{x}_t$, y_t is also possible in the case of sequentially revealed features. For the lower bound, it seems that choosing random feature vectors that are independent over time might not be a good idea, since the final normalized Gram matrix $\mathbf{G}_T/T$ may be concentrated around its expectation $\mathbf{G}$, and the regression strategy might use the possibly known $\mathbf{G}$ to transform the features $\mathbf{x}_t$ as in Theorem 7. Instead, choosing random features $\mathbf{x}_t$ that are dependent over time might make the task of predicting the final Gram matrix $\mathbf{G}_T$ virtually impossible, and might help to improve the lower bound. Alternatively, we could construct features $\mathbf{x}_t$ in a truly sequential manner, as functions of the strategy's past predictions, so as to annoy the regression strategy.

4.2. Some further technical remarks

We provide details on two claims issued above.

Remark 12 (Links between Theorem 11 and Theorem 2) Assume that we use the non-linear ridge regression algorithm with $\lambda=0$ as in (16) but feed it first with d warm-up feature vectors $\mathbf{x}_{-t}=(0,\dots,0,\sqrt{\lambda},0,\dots,0)$, where the $\sqrt{\lambda}$ is in position $t\in\{1,\dots,d\}$, and that the observations are $y_{-t}=0$. Then for each $\mathbf{u}\in\mathbb{R}^d$, a cumulative loss of $\lambda\|\mathbf{u}\|^2$ is suffered, and to neglect these d additional rounds in the regret bound obtained by Theorem 11, we need to add a $\lambda\|\mathbf{u}\|^2$ term to it. As all terms corresponding to the new eigenvalues introduced $\lambda_{r_t}(\mathbf{G}_t)$ are equal to λ, given the choice of these warm-up features, we are thus essentially back to the bound of Theorem 2.

Remark 13 (How we came up with the forecaster (16)) A natural attempt to transform the forecaster (6) designed for the case of beforehand-known features into a fully sequential algorithm is to replace the matrix $\mathbf{G}_T$ that is unknown at the beginning of round t by its sequential estimate $\mathbf{G}_t$ and to regularize at time t with $(\mathbf{u}\cdot\mathbf{x}_t)^2+\lambda\|\mathbf{u}\|^2_{\mathbf{G}_t}$ instead of $(\mathbf{u}\cdot\mathbf{x}_t)^2+\lambda\|\mathbf{u}\|^2_{\mathbf{G}_T}$ as in (6). However, in this case, the closed-form expression for the vector $\widehat{\mathbf{u}}_t$ is $\widehat{\mathbf{u}}_t=\mathbf{G}_t^\dagger\mathbf{b}_{t-1}/(1+\lambda)$, that is, the λ only acts as a multiplicative bias to the vector otherwise considered in (16). The analysis we followed led to a regret bound increasing in λ, so that we finally picked $\lambda=0$ and ended up with our non-linear ridge regression algorithm with $\lambda=0$ as in (16).

Appendix A. Details on the proof of Theorem 4

Details on getting (5)

By exchanging an expectation and an infimum, the expectation of the uniform regret of any fixed forecaster considered can be bounded as

$$\mathbb{E}_{\theta^\star}\left[\sup_{\mathbf{u}\in\mathbb{R}^d} \mathcal{R}_T(\mathbf{u})\right] \geqslant \sum_{t=1}^T \mathbb{E}_{\theta^\star}\left[(Y_t - \widehat{y}_t)^2\right] - \inf_{\mathbf{u}\in\mathbb{R}^d} \sum_{t=1}^T \mathbb{E}_{\theta^\star}\left[(Y_t - \mathbf{u}\cdot\mathbf{e}_{J_t})^2\right]. \tag{17}$$

Since $\widehat{y}_t$ is measurable w.r.t. $\mathcal{F}_{t-1}$, the σ–algebra generated by the information available at the beginning of round t, namely, $J_1,\dots,J_T$ and $Y_1,\dots,Y_{t-1}$, and since Y_t is distributed, conditionally on $\mathcal{F}_{t-1}$ according to a Bernoulli distribution with parameter $\theta^\star_{J_t}$, a conditional bias–variance decomposition yields

$$\begin{aligned}\mathbb{E}_{\theta^\star}\left[(\widehat{y}_t - Y_t)^2 \,\middle|\, \mathcal{F}_{t-1}\right] &= (\widehat{y}_t - \theta^\star_{J_t})^2 + \mathbb{E}_{\theta^\star}\left[(Y_t - \theta^\star_{J_t})^2 \,\middle|\, \mathcal{F}_{t-1}\right] \\ &= \left(\widehat{u}_{J_t,t} - \theta^\star_{J_t}\right)^2 + \theta^\star_{J_t}\left(1-\theta^\star_{J_t}\right),\end{aligned}$$

where we also used that by construction, $\widehat{y}_t = \widehat{\mathbf{u}}_t\cdot\mathbf{e}_{J_t} = \widehat{u}_{J_t,t}$. Similarly, for all $\mathbf{u}\in\mathbb{R}^d$,

$$\mathbb{E}_{\theta^\star}\left[(Y_t - \mathbf{u}\cdot\mathbf{e}_{J_t})^2 \,\middle|\, \mathcal{F}_{t-1}\right] = \left(u_{J_t} - \theta^\star_{J_t}\right)^2 + \theta^\star_{J_t}\left(1-\theta^\star_{J_t}\right).$$

By the tower rule and since the variance terms $\theta^\star_{J_t}\left(1-\theta^\star_{J_t}\right)$ cancel out, we thus proved that

$$\begin{aligned}\mathbb{E}_{\theta^\star}\left[\sup_{\mathbf{u}\in\mathbb{R}^d} \mathcal{R}_T(\mathbf{u})\right] &\geqslant \sum_{t=1}^T \mathbb{E}_{\theta^\star}\left[(\widehat{y}_t - Y_t)^2\right] - \inf_{\mathbf{u}\in\mathbb{R}^d} \sum_{t=1}^T \mathbb{E}_{\theta^\star}\left[(Y_t - \mathbf{u}\cdot\mathbf{e}_{J_t})^2\right] \\ &= \sum_{t=1}^T \mathbb{E}_{\theta^\star}\left[\left(\widehat{u}_{J_t,t} - \theta^\star_{J_t}\right)^2\right] - \inf_{\mathbf{u}\in\mathbb{R}^d} \sum_{t=1}^T \mathbb{E}_{\theta^\star}\left[\left(u_{J_t} - \theta^\star_{J_t}\right)^2\right] \\ &= \sum_{t=1}^T \mathbb{E}_{\theta^\star}\left[(\widehat{u}_{J_t,t} - \theta^\star_{J_t})^2\right].\end{aligned}$$

Now, by resorting to the tower rule again, integrating over J_t conditionally on $Y_1,\dots,Y_{t-1}$ and $J_1,\dots,J_{t-1},J_{t+1},\dots,J_T$, we get

$$\mathbb{E}_{\theta^\star}\left[\sup_{\mathbf{u}\in\mathbb{R}^d} \mathcal{R}_T(\mathbf{u})\right] \geqslant \sum_{t=1}^T \mathbb{E}_{\theta^\star}\left[(\widehat{u}_{J_t,t} - \theta^\star_{J_t})^2\right] = \sum_{t=1}^T \frac{1}{d}\,\mathbb{E}_{\theta^\star}\left[\left\|\,\widehat{\mathbf{u}}_t - \theta^\star\,\right\|_2^2\right]. \tag{18}$$

We now show that each term in the sum is larger than something of the order of d/t. This order of magnitude d/t is the parametric rate of optimal estimation; indeed, due to the randomness of the J_s, over t periods, each component is used about t/d times, while the rate of convergence in quadratic error of any d–dimensional estimator based on $\tau = t/d$ unbiased i.i.d. observations is at best $d/\tau = d^2/t$. Taking into account the $1/d$ factor gets us the claimed d/t rate. The next steps (based on the van Trees inequality) transform this intuition into formal statements.

Conclusion of the proof, given the application of the van Trees inequality

We resume at (18) and consider a prior π over the $\theta^\star \in [0,1]^d$. Since an expectation is always smaller than a supremum, we have first, given the defining equation (4) of $\mathcal{R}^\star_{T,\,[0,1]}$,

$$\begin{aligned}\mathcal{R}^\star_{T,\,[0,1]} &\geqslant \inf_{\text{forecasters}} \int_{[0,1]^d} \mathbb{E}_{\theta^\star}\left[\sup_{\mathbf{u}\in\mathbb{R}^d} \mathcal{R}_T(\mathbf{u})\right] \mathrm{d}\pi(\theta^\star) \\ &\geqslant \inf_{\text{forecasters}} \sum_{t=1}^{T} \int_{[0,1]^d} \frac{1}{d}\,\mathbb{E}_{\theta^\star}\Big[\big\|\widehat{\mathbf{u}}_t - \theta^\star\big\|_2^2\Big]\,\mathrm{d}\pi(\theta^\star)\,,\end{aligned}$$

where the second inequality follows by mixing both sides of (18) according to π. Now, an immediate application of the (multi-dimentional) van Trees inequality with a Beta(α,α) prior π shows that for all forecasters, all $t \geqslant 1$ and $\alpha \geqslant 3$,

$$\int_{[0,1]^d} \frac{1}{d}\,\mathbb{E}_{\theta^\star}\Big[\big\|\widehat{\mathbf{u}}_t - \theta^\star\big\|_2^2\Big]\,\mathrm{d}\pi(\theta^\star) \geqslant \frac{d}{4(t-1) + 2(t-1)/(\alpha-1) + 16d\alpha}\,,$$

see Lemma 14 below. We thus proved

$$\begin{aligned}\mathcal{R}^\star_{T,\,[0,1]} &\geqslant \sum_{t=0}^{T-1} \frac{d}{\big(4+2/(\alpha-1)\big)t + 16d\alpha} \geqslant d\int_0^T \frac{1}{\big(4+2/(\alpha-1)\big)t + 16d\alpha}\,\mathrm{d}t \\ &= \frac{d}{4+2/(\alpha-1)} \ln \frac{\big(4+2/(\alpha-1)\big)T + 16d\alpha}{16d\alpha} \\ &\geqslant \frac{d}{4+2/(\alpha-1)} \ln \frac{4T}{16d\alpha} \\ &= \frac{d}{4+2/(\alpha-1)} \big(\ln T - \ln(4d\alpha)\big)\,,\end{aligned}$$

which we lower bound in a crude way by resorting to $1/(1+u) \geqslant 1-u$ and by taking α such that $\alpha - 1 = \ln T$; this is where our condition $T \geqslant 8 > \mathrm{e}^2$ is used, to ensure that $\alpha \geqslant 3$. We also use that since $T \geqslant \mathrm{e}^2$, we have $1 \leqslant (\ln T)/2$ thus $4d\alpha \leqslant 4d(1+\ln T) \leqslant 6d\ln T$. We get

$$\begin{aligned}\mathcal{R}^\star_{T,\,[0,1]} &\geqslant \frac{d}{4}\Bigg(1 - \underbrace{\frac{1}{2(\alpha-1)}}_{\geqslant 0}\Bigg)\big(\ln T - \ln(6d\ln T)\big) \\ &\geqslant \frac{d}{4}\left(1 - \frac{1}{2\ln T}\right)(\ln T - \ln(6d) - \ln\ln T) \geqslant \frac{d}{4}\big(\ln T - (3+\ln d) - \ln\ln T\big)\,. \qquad (19)\end{aligned}$$

The factor 3 above corresponds to $1/2 + \ln 6 \leqslant 3$. So, we covered the case of $\mathcal{R}^\star_{T,\,[0,1]}$ and now turn to $\mathcal{R}^\star_{T,\,[-B,B]}$ for a general $B > 0$.

Going from $\mathcal{R}^\star_{T,\,[0,1]}$ to $\mathcal{R}^\star_{T,\,[-B,B]}$

To get a lower bound of exact order $d\ln T$, that is, to get rid of the annoying multiplicative factor of $1/4$, we proceed as follows. With the notation above, $Z_t = 2B(Y_t - 1/2)$ lies in $[-B,B]$. Denoting

by $\widehat{z}_t$ the forecasts output by a given forecaster sequentially fed with the $(Z_s, \mathbf{e}_{J_s})$, we have

$$(\widehat{z}_t - Z_t)^2 = 4B^2(\widehat{y}_t - Y_t)^2 \qquad \text{where the} \qquad \widehat{y}_t = \frac{\widehat{z}_t + 1/2}{2B}$$

also correspond to predictions output by a legitimate forecaster, and

$$\inf_{\mathbf{v}\in\mathbb{R}^d} \sum_{t=1}^T \mathbb{E}\Big[(Z_t - \mathbf{v}\cdot\mathbf{e}_{J_t})^2\Big] = 4B^2 \inf_{\mathbf{v}\in\mathbb{R}^d} \sum_{t=1}^T \mathbb{E}\left[\left(Y_t - \frac{1}{2} - \frac{\mathbf{v}\cdot\mathbf{e}_{J_t}}{2B}\right)^2\right] = 4B^2 \inf_{\mathbf{u}\in\mathbb{R}^d} \sum_{t=1}^T \mathbb{E}\Big[(Y_t - \mathbf{u}\cdot\mathbf{e}_{J_t})^2\Big]$$

by considering the transformation $\mathbf{v} \leftrightarrow \mathbf{u}$ given by $u_j = v_j/(2B) - 1/2$. (We use here that the sum of the components of the $\mathbf{e}_{J_t}$ equal 1.) We thus showed that $\mathcal{R}^\star_{T,[-B,B]}$ is larger than $4B^2$ times the lower bound (19) exhibited on (17), which concludes the proof.

Details on the application of the van Trees inequality

The van Trees inequality is a Bayesian version of the Cramér-Rao bound, but holding for any estimator (not only the unbiased ones); see Gill and Levit (1995, Section 4) for a multivariate statement (and refer to Van Trees, 1968 for its first statement).

Recall that we denoted above by $\mathcal{P}_{\theta^\star}$ the distribution of the sequence of pairs (J_t, Y_t), with $1 \leqslant t \leqslant T$, considered in Section 2.2 for a given $\theta^\star \in [0,1]^d$. We also considered the family $\mathcal{P}$ of these distributions and thus, for clarity, indexed all expectations $\mathbb{E}$ by the underlying parameter $\theta^\star$ at hand. We introduce a product of independent Beta(α,α) distributions as a prior π on the $\theta^\star \in [0,1]^d$; its density with respect to the Lebesgue measure equals

$$\beta^{(d)}_{\alpha,\alpha}(t_1,\dots,t_d) \longmapsto \beta_{\alpha,\alpha}(t_1)\cdots\beta_{\alpha,\alpha}(t_d)\,, \qquad \text{where} \qquad \beta_{\alpha,\alpha} : t \mapsto \frac{\Gamma(2\alpha)}{\big(\Gamma(\alpha)\big)^2}\, t^{\alpha-1}(1-t)^{\alpha-1}\,.$$

The reason why Beta distributions are considered is because of the form of the Fisher information of the $\mathcal{P}$ family, see calculations (22) below.

The multivariate van Trees inequality ensures that for all estimators $\widehat{\mathbf{u}}_t$, that is, for all random variables which are measurable functions of $J_1,\dots,J_T$ and $Y_1,\dots,Y_{t-1}$, we have

$$\int_{[0,1]^d} \mathbb{E}_{\theta^\star}\Big[\big\|\,\widehat{\mathbf{u}}_t - \theta^\star\,\big\|_2^2\Big]\ \beta^{(d)}_{\alpha,\alpha}(\theta^\star)\ \mathrm{d}\theta^\star \geqslant \frac{(\operatorname{Tr}\mathbf{I}_d)^2}{\operatorname{Tr}\mathcal{I}\big(\beta^{(d)}_{\alpha,\alpha}\big) + \displaystyle\int_{[0,1]^d} \big(\operatorname{Tr}\mathcal{I}(\theta^\star)\big)\,\beta^{(d)}_{\alpha,\alpha}(\theta^\star)\ \mathrm{d}\theta^\star}\,, \quad (20)$$

where $\mathrm{d}\theta^\star$ denotes the integration w.r.t. Lebesgue measure, Tr is the trace operator, $\mathcal{I}(\theta^\star)$ stands for the Fisher information of the family $\mathcal{P}$ at $\theta^\star$, see (21), while each component (i,i) of the other matrix in the denominator is given by

$$\mathcal{I}\big(\beta^{(d)}_{\alpha,\alpha}\big)_{i,i} \stackrel{\text{def}}{=} \int_{[0,1]^d} \left(\frac{\partial\beta^{(d)}_{\alpha,\alpha}}{\partial\theta^\star_i}(\theta^\star)\right)^2 \frac{1}{\beta^{(d)}_{\alpha,\alpha}(\theta^\star)}\ \mathrm{d}\theta^\star\,,$$

which may equal $+\infty$ (in which case the lower bound is void). There are conditions for the inequality to be satisfied, we detail them in the proof of the lemma below.

Lemma 14 *When the family $\mathcal{P}$ is equipped with a prior given by a product of independent Beta(α,α) distributions, where $\alpha \geqslant 3$, it follows from the van Trees inequality and from simple calculations that*

$$\int_{[0,1]^d} \mathbb{E}_{\theta^\star}\Big[\big\| \widehat{\mathbf{u}}_t - \theta^\star \big\|_2^2 \Big] \, \beta_{\alpha,\alpha}^{(d)}(\theta^\star) \, \mathrm{d}\theta^\star \geqslant \frac{d^2}{16d\alpha + 4(t-1) + 2(t-1)/(\alpha-1)} \, .$$

Proof We denote by

$$f_{\theta^\star} : (j_1, \dots, j_T, \, y_1, \dots, y_{t-1}) \in \{1, \dots, d\}^T \times \{0,1\}^{t-1} \longmapsto \frac{1}{d^T} \prod_{s=1}^{t-1} {\theta^\star_{j_s}}^{y_s} (1 - \theta^\star_{j_s})^{1-y_s}$$

the density of $\mathcal{P}_{\theta^\star}$ w.r.t. to the counting measure μ on $\{1, \dots, d\}^T \times \{0,1\}^{t-1}$.

The sufficient conditions of Gill and Levit (1995, Section 4) for (20) are met, since on the one hand $\beta_{\alpha,\alpha}^{(d)}$ is C^1–smooth, vanishes on the border of $[0,1]^d$, and is positive on its interior, while on the other hand, $\theta^\star \mapsto f_{\theta^\star}(j_1, \dots, j_T, \, y_1, \dots, y_{t-1})$ is C^1–smooth for all $(j_1, \dots, j_T, \, y_1, \dots, y_{t-1})$, with, for all $i \in \{1, \dots, d\}$,

$$L_i(\theta^\star) = \frac{\partial}{\partial \theta_i^\star} \ln f_{\theta^\star}(J_1, \dots, J_T, \, Y_1, \dots, Y_{t-1}) = \sum_{s=1}^{t-1} \left(\frac{Y_s}{\theta^\star_{J_s}} - \frac{1 - Y_s}{1 - \theta^\star_{J_s}} \right) \mathbb{1}_{\{J_s = i\}}$$

being square integrable, so that the Fisher information matrix $\mathcal{I}(\theta^\star)$ of the $\mathcal{P}$ model at $\theta^\star$ exists and has a component (i,i) given by

$$\begin{aligned} \mathcal{I}(\theta^\star)_{i,i} \overset{\text{def}}{=} \mathbb{E}_{\theta^\star}\Big[L_i(\theta^\star)^2 \Big] &= (t-1) \, \mathbb{E}_{\theta^\star}\left[\left(\frac{Y_1}{\theta^\star_{J_1}} - \frac{1 - Y_1}{1 - \theta^\star_{J_1}} \right)^2 \mathbb{1}_{\{J_1 = i\}} \right] \\ &= \frac{t-1}{d} \left(\frac{1}{\theta_i^\star} + \frac{1}{1 - \theta_i^\star} \right) = \frac{t-1}{d \, \theta_i^\star (1 - \theta_i^\star)} \, , \end{aligned} \tag{21}$$

and therefore, is such that $\theta^\star \mapsto \sqrt{\mathcal{I}(\theta^\star)}$ is locally integrable w.r.t. the Lebesgue measure. The second inequality in (21) is because $L_i(\theta^\star)$ is a sum of $t-1$ centered, independent and identically distributed variables, while the third inequality is obtained by the tower rule, by first taking the conditional expectation with respect to J_1.

We now compute all elements of the denominator of (20). First, by symmetry and then by substituting (21),

$$\begin{aligned} &\int_{[0,1]^d} \big(\mathrm{Tr} \, \mathcal{I}(\theta^\star) \big) \, \beta_{\alpha,\alpha}^{(d)}(\theta^\star) \, \mathrm{d}\theta^\star \\ &= d \int_{[0,1]^d} \mathcal{I}(\theta^\star)_{1,1} \, \beta_{\alpha,\alpha}^{(d)}(\theta^\star) \, \mathrm{d}\theta^\star \\ &= d \int_{[0,1]^d} \frac{t-1}{d \, \theta_1^\star (1 - \theta_1^\star)} \, \frac{\Gamma(2\alpha)}{\big(\Gamma(\alpha)\big)^2} \, (\theta_1^\star)^{\alpha-1} (1 - \theta_1^\star)^{\alpha-1} \, \beta_{\alpha,\alpha}(\theta_2^\star) \cdots \beta_{\alpha,\alpha}(\theta_d^\star) \, \mathrm{d}\theta^\star \\ &= (t-1) \, \frac{\Gamma(2\alpha)}{\big(\Gamma(\alpha)\big)^2} \int_{[0,1]} z^{\alpha-2} (1-z)^{\alpha-2} \, \mathrm{d}z = (t-1) \, \frac{\Gamma(2\alpha)}{\big(\Gamma(\alpha)\big)^2} \frac{\big(\Gamma(\alpha-1)\big)^2}{\Gamma\big(2(\alpha-1)\big)} \, , \end{aligned} \tag{22}$$

where we used the expression of the density of the Beta$(\alpha-1,\alpha-1)$ distribution for the last equality. Using that $x\,\Gamma(x)=\Gamma(x+1)$ for all real numbers $x>0$, we finally get

$$\int_{[0,1]^d}\big(\operatorname{Tr}\mathcal{I}(\theta^\star)\big)\,\beta_{\alpha,\alpha}^{(d)}(\theta^\star)\;\mathrm{d}\theta^\star=\frac{(2\alpha-1)(2\alpha-2)}{(\alpha-1)^2}\,(t-1)=\frac{4\alpha-2}{\alpha-1}\,(t-1)=4(t-1)+\frac{2(t-1)}{\alpha-1}\,.$$

Second, as far as the $\operatorname{Tr}\mathcal{I}\big(\beta_{\alpha,\alpha}^{(d)}\big)$ in (20) is concerned, because $\beta_{\alpha,\alpha}^{(d)}$ is a product of univariate distributions,

$$\mathcal{I}\big(\beta_{\alpha,\alpha}^{(d)}\big)_{i,i}=\int_{[0,1]^d}\left(\frac{\partial\beta_{\alpha,\alpha}^{(d)}}{\partial\theta_i^\star}(\theta^\star)\right)^2\frac{1}{\beta_{\alpha,\alpha}^{(d)}(\theta^\star)}\;\mathrm{d}\theta^\star=\int_{[0,1]}\left(\frac{\partial\beta_{\alpha,\alpha}}{\partial z}(z)\right)^2\frac{1}{\beta_{\alpha,\alpha}}(z)\;\mathrm{d}z\,,$$

so that $\operatorname{Tr}\mathcal{I}\big(\beta_{\alpha,\alpha}^{(d)}\big)$ equals d times this value, that is, d times

$$\begin{aligned}
&\int_{[0,1]}\frac{\Gamma(2\alpha)}{\big(\Gamma(\alpha)\big)^2}\,\frac{\big((\alpha-1)\,z^{\alpha-2}(1-z)^{\alpha-1}-(\alpha-1)\,z^{\alpha-1}(1-z)^{\alpha-2}\big)^2}{z^{\alpha-1}(1-z)^{\alpha-1}}\;\mathrm{d}z\\
&=\frac{(\alpha-1)^2\,\Gamma(2\alpha)}{\big(\Gamma(\alpha)\big)^2}\int_{[0,1]}(1-2z)^2\;z^{\alpha-3}(1-z)^{\alpha-3}\,\mathrm{d}z\\
&=\frac{(\alpha-1)^2\,\Gamma(2\alpha)}{\big(\Gamma(\alpha)\big)^2}\,\frac{\big(\Gamma(\alpha-2)\big)^2}{\Gamma\big(2(\alpha-2)\big)}\,\mathbb{E}\Big[(1-2Z_{\alpha-2})^2\Big]=\frac{(\alpha-1)^2\,\Gamma(2\alpha)}{\big(\Gamma(\alpha)\big)^2}\,\frac{\big(\Gamma(\alpha-2)\big)^2}{\Gamma\big(2(\alpha-2)\big)}\,4\operatorname{Var}(Z_{\alpha-2})
\end{aligned}$$

where $Z_{\alpha-2}$ is a random variable following the Beta$(\alpha-2,\alpha-2)$ distribution; its expectation equals indeed $\mathbb{E}[Z_{\alpha-2}]=1/2$ by symmetry of the distribution w.r.t. $1/2$, so that

$$\mathbb{E}\Big[(1-2Z_{\alpha-2})^2\Big]=4\;\mathbb{E}\Big[(1/2-Z_{\alpha-2})^2\Big]=4\operatorname{Var}(Z_{\alpha-2})\qquad\text{where}\qquad\operatorname{Var}(Z_{\alpha-2})=\frac{1}{4(2\alpha-3)}$$

by a classical formula. Collecting all elements together and using again that $x\,\Gamma(x)=\Gamma(x+1)$ for all real numbers $x>0$, we get

$$\operatorname{Tr}\mathcal{I}\big(\beta_{\alpha,\alpha}^{(d)}\big)=d\;\underbrace{\frac{(\alpha-1)^2\,\big(\Gamma(\alpha-2)\big)^2}{\big(\Gamma(\alpha)\big)^2}}_{1/(\alpha-2)^2}\;\underbrace{\frac{\Gamma(2\alpha)}{(2\alpha-3)\,\Gamma\big(2(\alpha-2)\big)}}_{=(2\alpha-1)(2\alpha-2)(2\alpha-4)}=d\;\frac{4(2\alpha-1)(\alpha-1)}{\alpha-2}$$

hence the upper bound $\operatorname{Tr}\mathcal{I}\big(\beta_{\alpha,\alpha}^{(d)}\big)\leqslant 16d\alpha$ for $\alpha\geqslant 3$, which concludes the proof. ■

Appendix B. Proof of Theorem 2 and of Corollary 3

We start with the proof of Corollary 3.

Proof We assume that the Gram matrix G_T is full rank; otherwise, we may adapt the proof below by resorting to Moore-Penrose pseudoinverses, just as we do in Appendix C.2 for the proof of Theorem 7.

Theorem 2 indicates that

$$\sum_{t=1}^{T}(y_t - \widehat{\mathbf{u}}_t \cdot \mathbf{x}_t)^2 \leqslant \inf_{\mathbf{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}(y_t - \mathbf{u}\cdot\mathbf{x}_t)^2 + \lambda\left\|\mathbf{u}\right\|^2\right\} + B^2\sum_{k=1}^{d}\ln\left(1+\frac{\lambda_k(\mathbf{G}_T)}{\lambda}\right).$$

Now, as in (3), we have a closed-form expression of the unique vector achieving the following, infimum:

$$\inf_{\mathbf{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}(y_t - \mathbf{u}\cdot\mathbf{x}_t)^2\right\} = \sum_{t=1}^{T}(y_t - \mathbf{u}^\star\cdot\mathbf{x}_t)^2$$

Namely, $\mathbf{u}^\star = \mathbf{G}_T^{-1}\mathbf{b}_T$, so that

$$\begin{aligned}\left\|\mathbf{u}^\star\right\| = \left\|\mathbf{G}_T^{-1/2}\,\mathbf{G}_T^{-1/2}\mathbf{b}_T\right\| \leqslant \lambda_1\big(\mathbf{G}_T^{-1/2}\big)\,\left\|\mathbf{G}_T^{-1/2}\mathbf{b}_T\right\| &= \frac{1}{\sqrt{\lambda_d(\mathbf{G}_T)}}\,\left\|\mathbf{G}_T^{-1/2}\mathbf{b}_T\right\| \\ &\leqslant \frac{1}{\sqrt{\lambda_d(\mathbf{G}_T)}}\,B\sqrt{T}\,, \end{aligned} \tag{23}$$

where we used, for the final inequality, an elementary argument of orthogonal projection that is at the heart of the proof of Theorem 7: see (15) and the sentence after it. In addition, Jensen's inequality (or the alternative treatment of Cesa-Bianchi and Lugosi, 2006, page 320) indicates that

$$\sum_{k=1}^{d}\ln\left(1+\frac{\lambda_k(\mathbf{G}_T)}{\lambda}\right) \leqslant d\ln\left(1+\frac{\sum_{k=1}^{d}\lambda_k(\mathbf{G}_T)}{d\lambda}\right) = d\ln\left(1+\frac{\mathrm{Tr}(\mathbf{G}_T)}{d\lambda}\right) \leqslant d\ln\left(1+\frac{TX^2}{d\lambda}\right)$$

where Tr is the trace operator. All in all, we get

$$\sum_{t=1}^{T}(y_t - \widehat{\mathbf{u}}_t \cdot \mathbf{x}_t)^2 \leqslant \inf_{\mathbf{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}(y_t - \mathbf{u}\cdot\mathbf{x}_t)^2 +\right\} + \lambda\left\|\mathbf{u}^\star\right\|^2 + dB^2\ln\left(1+\frac{TX^2}{d\lambda}\right)$$

and the claimed bound follows by substituting the bound (23). ■

Now, we move to the proof of Theorem 2, which we essentially extract from Cesa-Bianchi and Lugosi (2006, Chapter 11). We merely provide it because we will later need the first inequality of (24) in the proof of Theorem 11 and we wanted this article to be self-complete. But of course, this is extremely standard content and it should be skipped by any reader familiar with the basic results of sequential linear regression.

Proof We successively prove the following two inequalities,

$$\mathcal{R}_T(\mathbf{u}) \leqslant \lambda\left\|\mathbf{u}\right\|^2 + \sum_{t=1}^{T} y_t^2\,\mathbf{x}_t^{\mathrm{T}}\mathbf{A}_t^{-1}\mathbf{x}_t \leqslant \lambda\left\|\mathbf{u}\right\|^2 + B^2\sum_{k=1}^{d}\ln\left(1+\frac{\lambda_k(\mathbf{G}_T)}{\lambda}\right) \tag{24}$$

Proof of the first inequality in (24). We denote by L_{t-1}^{reg} the cumulative loss up to round $t-1$ included, to which we add the regularization term:

$$L_{t-1}^{\mathrm{reg}}(\mathbf{u}) = \sum_{s=1}^{t-1}(y_s - \mathbf{u}\cdot\mathbf{x}_s)^2 + \lambda\left\|\mathbf{u}\right\|^2$$

For all $t \geqslant 1$, we denote by

$$\check{\mathbf{u}}_t \in \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^d} \left\{ \sum_{s=1}^{t-1} (y_s - \mathbf{u}\cdot\mathbf{x}_s)^2 + \lambda \|\mathbf{u}\|^2 \right\} = \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^d} L_{t-1}^{\text{reg}}(\mathbf{u}) \,,$$

the vector output by the (ordinary) ridge regression; that is, when no $(\mathbf{u}\cdot\mathbf{x}_t)^2$ term is added to the regularization. In particular, $\check{\mathbf{u}}_1 = (0,\dots,0)^{\mathrm{T}}$. By the very definition of $\check{\mathbf{u}}_{T+1}$, for all $\mathbf{u}\in\mathbb{R}^d$,

$$L_T^{\text{reg}}(\check{\mathbf{u}}_{T+1}) \leqslant \sum_{t=1}^{T} (y_t - \mathbf{u}\cdot\mathbf{x}_t)^2 + \lambda \|\mathbf{u}\|^2 \;,$$

so that, for all $\mathbf{u}\in\mathbb{R}^d$,

$$\begin{aligned}\mathcal{R}_T(\mathbf{u}) &\leqslant \sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 + \lambda \|\mathbf{u}\|^2 - L_T^{\text{reg}}(\check{\mathbf{u}}_{T+1}) \\ &= \lambda \|\mathbf{u}\|^2 + \sum_{t=1}^{T} \big((y_t - \widehat{y}_t)^2 + L_{t-1}^{\text{reg}}(\check{\mathbf{u}}_t) - L_t^{\text{reg}}(\check{\mathbf{u}}_{t+1}) \big) \,,\end{aligned}$$

where the equality comes from a telescoping argument together with $L_0^{\text{reg}}(\check{\mathbf{u}}_0) = 0$. We will prove by means of direct calculations that

$$(y_t - \widehat{y}_t)^2 + L_{t-1}^{\text{reg}}(\check{\mathbf{u}}_t) - L_t^{\text{reg}}(\check{\mathbf{u}}_{t+1}) = (\check{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t)^{\mathrm{T}} \mathbf{A}_t (\check{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t) - (\widehat{\mathbf{u}}_t - \check{\mathbf{u}}_t)^{\mathrm{T}} \mathbf{A}_{t-1} (\widehat{\mathbf{u}}_t - \check{\mathbf{u}}_t) \,; \quad (25)$$

the first inequality in (24) will then be obtained, as the second term in (25) is negative and as the first term in (25) can be rewritten as $y_t^2\, \mathbf{x}_t^{\mathrm{T}} \mathbf{A}_t^{-1} \mathbf{x}_t$ thanks to the equality (27) below, which states $\mathbf{A}_t(\check{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t) = y_t \mathbf{x}_t$.

To prove (25), we recall the closed-form expression (3), that is, $\widehat{\mathbf{u}}_t = \mathbf{A}_t^{-1}\mathbf{b}_{t-1}$, and note that we similarly have $\check{\mathbf{u}}_{t+1} = \mathbf{A}_t^{-1}\mathbf{b}_t$. Now, L_t^{reg} rewrites, for all $\mathbf{u}\in\mathbb{R}^d$,

$$L_t^{\text{reg}}(\mathbf{u}) = \left(\sum_{s=1}^{t} y_s^2 \right) - 2\mathbf{b}_t^{\mathrm{T}}\mathbf{u} + \mathbf{u}^{\mathrm{T}}\mathbf{A}_t\mathbf{u} \,,$$

so that the minimum of this quadratic form, achieved at $\mathbf{u} = \check{\mathbf{u}}_{t+1} = \mathbf{A}_t^{-1}\mathbf{b}_t$, equals

$$L_t^{\text{reg}}(\check{\mathbf{u}}_{t+1}) = \left(\sum_{s=1}^{t} y_s^2 \right) - 2\underbrace{\mathbf{b}_t^{\mathrm{T}}\mathbf{A}_t^{-1}}_{=\check{\mathbf{u}}_{t+1}^{\mathrm{T}}} \mathbf{A}_t\check{\mathbf{u}}_{t+1} + \check{\mathbf{u}}_{t+1}^{\mathrm{T}}\mathbf{A}_t\check{\mathbf{u}}_{t+1} = \left(\sum_{s=1}^{t} y_s^2 \right) - \check{\mathbf{u}}_{t+1}^{\mathrm{T}}\mathbf{A}_t\check{\mathbf{u}}_{t+1} \,.$$

In particular,

$$L_{t-1}^{\text{reg}}(\check{\mathbf{u}}_t) - L_t^{\text{reg}}(\check{\mathbf{u}}_{t+1}) = -y_t^2 + \check{\mathbf{u}}_{t+1}^{\mathrm{T}}\mathbf{A}_t\check{\mathbf{u}}_{t+1} - \check{\mathbf{u}}_t^{\mathrm{T}}\mathbf{A}_{t-1}\check{\mathbf{u}}_t \,. \quad (26)$$

We now expand the first term in (25). To that end, we use that from the closed-form expressions of $\widehat{\mathbf{u}}_t$ and $\check{\mathbf{u}}_{t+1}$,

$$\mathbf{A}_t(\check{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t) = \mathbf{A}_t\big(\mathbf{A}_t^{-1}\mathbf{b}_t - \mathbf{A}_t^{-1}\mathbf{b}_{t-1}\big) = \mathbf{b}_t - \mathbf{b}_{t-1} = y_t\mathbf{x}_t \,. \quad (27)$$

Therefore, $y_t\widehat{y}_t = y_t\mathbf{x}_t^{\mathrm{T}}\widehat{\mathbf{u}}_t = (\breve{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t)^{\mathrm{T}}\mathbf{A}_t\widehat{\mathbf{u}}_t$ and

$$\begin{aligned}(y_t - \widehat{y}_t)^2 = y_t^2 - 2y_t\widehat{y}_t + \widehat{y}_t^2 &= y_t^2 - 2(\breve{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t)^{\mathrm{T}}\mathbf{A}_t\widehat{\mathbf{u}}_t + \widehat{\mathbf{u}}_t^{\mathrm{T}}\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}\widehat{\mathbf{u}}_t \\ &= y_t^2 - 2(\breve{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t)^{\mathrm{T}}\mathbf{A}_t\widehat{\mathbf{u}}_t + \widehat{\mathbf{u}}_t^{\mathrm{T}}(\mathbf{A}_t - \mathbf{A}_{t-1})\widehat{\mathbf{u}}_t\,, \qquad (28)\end{aligned}$$

where in the last equality we used that by definition $\mathbf{A}_t - \mathbf{A}_{t-1} = \mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}$.

Putting (26) and (28) together, we proved

$$\begin{aligned}&(y_t - \widehat{y}_t)^2 + L_{t-1}^{\text{reg}}(\breve{\mathbf{u}}_t) - L_t^{\text{reg}}(\breve{\mathbf{u}}_{t+1}) \\ &= -2(\breve{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t)^{\mathrm{T}}\mathbf{A}_t\widehat{\mathbf{u}}_t + \widehat{\mathbf{u}}_t^{\mathrm{T}}(\mathbf{A}_t - \mathbf{A}_{t-1})\widehat{\mathbf{u}}_t + \breve{\mathbf{u}}_{t+1}^{\mathrm{T}}\mathbf{A}_t\breve{\mathbf{u}}_{t+1} - \breve{\mathbf{u}}_t^{\mathrm{T}}\mathbf{A}_{t-1}\breve{\mathbf{u}}_t \\ &= \breve{\mathbf{u}}_{t+1}^{\mathrm{T}}\mathbf{A}_t\breve{\mathbf{u}}_{t+1} - 2\breve{\mathbf{u}}_{t+1}^{\mathrm{T}}\mathbf{A}_t\widehat{\mathbf{u}}_t + \widehat{\mathbf{u}}_t^{\mathrm{T}}\mathbf{A}_t\widehat{\mathbf{u}}_t - \big(\widehat{\mathbf{u}}_t^{\mathrm{T}}\mathbf{A}_{t-1}\widehat{\mathbf{u}}_t - 2\widehat{\mathbf{u}}_t^{\mathrm{T}}\underbrace{\mathbf{A}_t\widehat{\mathbf{u}}_t}_{=\mathbf{A}_{t-1}\breve{\mathbf{u}}_t} + \breve{\mathbf{u}}_t^{\mathrm{T}}\mathbf{A}_{t-1}\breve{\mathbf{u}}_t\big)\,.\end{aligned}$$

In the last equation, we are about to use the equality $\mathbf{A}_t\widehat{\mathbf{u}}_t = \mathbf{A}_{t-1}\breve{\mathbf{u}}_t = \mathbf{b}_{t-1}$, which we get from the closed-form expressions of $\widehat{\mathbf{u}}_t$ and $\breve{\mathbf{u}}_t$. We then recognize the desired difference between two quadratic forms:

$$\begin{aligned}&(y_t - \widehat{y}_t)^2 + L_{t-1}^{\text{reg}}(\breve{\mathbf{u}}_t) - L_t^{\text{reg}}(\breve{\mathbf{u}}_{t+1}) \\ &= \big(\breve{\mathbf{u}}_{t+1}^{\mathrm{T}}\mathbf{A}_t\breve{\mathbf{u}}_{t+1} - 2\breve{\mathbf{u}}_{t+1}^{\mathrm{T}}\mathbf{A}_t\widehat{\mathbf{u}}_t + \widehat{\mathbf{u}}_t^{\mathrm{T}}\mathbf{A}_t\widehat{\mathbf{u}}_t\big) - \big(\widehat{\mathbf{u}}_t^{\mathrm{T}}\mathbf{A}_{t-1}\widehat{\mathbf{u}}_t - 2\widehat{\mathbf{u}}_t^{\mathrm{T}}\mathbf{A}_{t-1}\breve{\mathbf{u}}_t + \breve{\mathbf{u}}_t^{\mathrm{T}}\mathbf{A}_{t-1}\breve{\mathbf{u}}_t\big) \\ &= (\breve{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t)^{\mathrm{T}}\mathbf{A}_t(\breve{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t) - (\widehat{\mathbf{u}}_t - \breve{\mathbf{u}}_t)^{\mathrm{T}}\mathbf{A}_{t-1}(\widehat{\mathbf{u}}_t - \breve{\mathbf{u}}_t)\,.\end{aligned}$$

Proof of the second inequality in (24). Because $y_t^2 \leqslant B^2$, we only need to prove

$$\sum_{t=1}^{T}\mathbf{x}_t^{\mathrm{T}}\mathbf{A}_t^{-1}\mathbf{x}_t \leqslant \sum_{k=1}^{d}\ln\left(1 + \frac{\lambda_k(\mathbf{G}_T)}{\lambda}\right)\,.$$

Now, Lemma 15 below shows that

$$\sum_{t=1}^{T}\mathbf{x}_t^{\mathrm{T}}\mathbf{A}_t^{-1}\mathbf{x}_t = \sum_{t=1}^{T}\left(1 - \frac{\det(\mathbf{A}_{t-1})}{\det(\mathbf{A}_t)}\right)\,.$$

We then use $1 - u \leqslant -\ln u$ for $u > 0$ and identify a telescoping sum,

$$\sum_{t=1}^{T}\left(1 - \frac{\det(\mathbf{A}_{t-1})}{\det(\mathbf{A}_t)}\right) \leqslant \sum_{t=1}^{T}\ln\frac{\det(\mathbf{A}_t)}{\det(\mathbf{A}_{t-1})} = \ln\frac{\det(\mathbf{A}_T)}{\det(\mathbf{A}_0)}\,.$$

All in all, we proved so far

$$\sum_{t=1}^{T}\mathbf{x}_t^{\mathrm{T}}\mathbf{A}_t^{-1}\mathbf{x}_t \leqslant \ln\frac{\det(\mathbf{A}_T)}{\det(\mathbf{A}_0)}\,,$$

and may conclude by noting that

$$\det(\mathbf{A}_T) = \det(\lambda\,\mathbf{I}_d + \mathbf{G}_T) = \prod_{k=1}^{d}(\lambda + \lambda_k(\mathbf{G}_T)) \qquad \text{and} \qquad \det(\mathbf{A}_0) = \det(\lambda\,\mathbf{I}_d) = \lambda^d\,.$$

■

Lemma 15 *Let V an arbitrary $d \times d$ full-rank matrix, let $\mathbf{u}$ and $\mathbf{v}$ two arbitrary vectors of $\mathbb{R}^d$, and let $\mathbf{U} = \mathbf{V} - \mathbf{u}\mathbf{v}^\mathrm{T}$. Then*

$$\mathbf{v}^\mathrm{T}\mathbf{V}^{-1}\mathbf{u} = 1 - \frac{\det(\mathbf{U})}{\det(\mathbf{V})}\,.$$

Proof If $V = \mathbf{I}_d$, we are left to show that $\det(\mathbf{I}_d - \mathbf{u}\mathbf{v}^\mathrm{T}) = 1 - \mathbf{v}^\mathrm{T}\mathbf{u}$. The result follows from taking the determinant of every term of the equality

$$\begin{bmatrix} \mathbf{I}_d & 0 \\ \mathbf{v}^\mathrm{T} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I}_d - \mathbf{u}\mathbf{v}^\mathrm{T} & -\mathbf{u} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I}_d & 0 \\ -\mathbf{v}^\mathrm{T} & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{I}_d & -\mathbf{u} \\ 0 & 1 - \mathbf{v}^\mathrm{T}\mathbf{u} \end{bmatrix}\,.$$

Now, we can reduce the case of a general $\mathbf{V}$ to this simpler case by noting that

$$\det(\mathbf{U}) = \det\bigl(\mathbf{V} - \mathbf{u}\mathbf{v}^\mathrm{T}\bigr) = \det(\mathbf{V})\det\Bigl(\mathbf{I}_d - \bigl(\mathbf{V}^{-1}\mathbf{u}\bigr)\mathbf{v}^\mathrm{T}\Bigr) = \det(\mathbf{V})\bigl(1 - \mathbf{v}^\mathrm{T}\mathbf{V}^{-1}\mathbf{u}\bigr)\,.$$

■

Appendix C. Technical complements to Section 3

In this section we provide some additional discussions to those of Remark 8 (Section C.1) and also extend the proof of Theorem 7 to work in the general case (Section C.2).

C.1. Complements to Remark 8

We detail here why the derivation of a closed-form bound as led by Bartlett et al. (2015) only entails a bound of the order of $2dB^2 \ln T$ and why it cannot easily be improved.

Indeed, Theorem 5 by Bartlett et al. (2015) indicates, in the case where $d = 1$ and $B = 1$, that

$$\forall\, T \geqslant 1, \qquad \mathcal{R}_T^\star \leqslant f(T) \tag{29}$$

for any function $f : \{1, 2, \ldots\} \to \mathbb{R}_+$ satisfying $e^{-f(T)/2} \leqslant f(T+1) - f(T)$ for all $T \geqslant 1$. As they showed, the function $f(T) = 2\ln(1 + T/2) + 1$ is a suitable choice, but it leads to the extra multiplicative factor of 2 that we pointed out.

However, this choice for f does not seem to be easily improvable; for instance, functions f of the form $T \mapsto a \ln T + b$ for some $a < 2$ and $b \in \mathbb{R}$ are such that

$$e^{-f(T)/2} = \Theta_T\bigl(T^{-a/2}\bigr) \qquad \text{and} \qquad f(T+1) - f(T) = a\ln\left(1 + \frac{1}{T}\right) = \mathcal{O}_T\bigl(T^{-1}\bigr)\,,$$

hence, are not suitable choices for the bound (29).

C.2. Proof of Theorem 7 in the general case

In this section we extend the proof of Theorem 7, provided only in the case of a full-rank Gram matrix G_T in Section 3, to the general case of a possibly non-invertible Gram matrix G_T.

To that end, we first explain how the closed-form expression (7) is derived. We rewrite the definition equation (6) of $\widehat{\mathbf{u}}_t$ as

$$\widehat{\mathbf{u}}_t \in \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^d} \bigl\{\mathbf{u}^\mathrm{T}(\lambda\,\mathbf{G}_T + \mathbf{G}_t)\mathbf{u} - 2\mathbf{b}_{t-1}^\mathrm{T}\mathbf{u}\bigr\}\,.$$

Because the matrix $\lambda \mathbf{G}_T + \mathbf{G}_t$ is positive semidefinite, the considered argmin is also the set of values $\mathbf{u}'$ where the gradient vanishes: $(\lambda \mathbf{G}_T + \mathbf{G}_t)\mathbf{u}' = \mathbf{b}_{t-1}$. This system is possibly underdefined because $\mathbf{u}' \in \mathbb{R}^d$ and $\lambda \mathbf{G}_T + \mathbf{G}_t$ is a matrix of size $\mathrm{d} \times \mathrm{d}$, possibly not full rank. The system has at least one solution but the one with minimal Euclidean norm is given by the Moore-Penrose inverse, see Corollary 19 (e):

$$\widehat{\mathbf{u}}_t = \big(\lambda \mathbf{G}_T + \mathbf{G}_t\big)^\dagger \mathbf{b}_{t-1}\,.$$

We may now turn to the general proof of Theorem 7. For an integer $k \geqslant 1$, we denote therein by $\mathbf{I}_k$ the $k \times k$ identity matrix.

Proof As a consequence of the spectral theorem applied to the symmetric matrix $\mathbf{G}_T$, there exists a matrix $\mathbf{U}$ of size $d \times r_T$ and a full rank square matrix $\mathbf{\Sigma}$ of size $r_T \times r_T$ such that $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_{r_T}$ and $\mathbf{G}_T = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\mathrm{T}}$. We could even impose that the matrix $\mathbf{\Sigma}$ be diagonal but this property will not be used in this proof.

We will apply the (already proven) bound of Theorem 7 in the full rank case. To that end, we consider the modified sequence of features

$$\widetilde{\mathbf{x}}_t = \mathbf{U}^{\mathrm{T}}\mathbf{x}_t$$

and first prove that the strategy (6) on the $\widetilde{\mathbf{x}}_t$ leads to the same forecasts as the same strategy on the original features $\mathbf{x}_t$; that is,

$$\widetilde{\mathbf{u}}_t \cdot \widetilde{\mathbf{x}}_t = \widehat{\mathbf{u}}_t \cdot \mathbf{x}_t\,, \qquad \text{where} \qquad \widetilde{\mathbf{u}}_t \in \operatorname*{argmin}_{\mathbf{v} \in \mathbb{R}^{r_T}} \left\{ \sum_{s=1}^{t-1} (y_s - \mathbf{v} \cdot \widetilde{\mathbf{x}}_s)^2 + (\mathbf{v} \cdot \widetilde{\mathbf{x}}_t)^2 + \lambda \sum_{s=1}^{T} (\mathbf{v} \cdot \widetilde{\mathbf{x}}_s)^2 \right\}.$$

It suffices to prove $\mathbf{U}\widetilde{\mathbf{u}}_t = \widehat{\mathbf{u}}_t$, which we do below. Then, from this equality and the definition $\widetilde{\mathbf{x}}_t = \mathbf{U}^{\mathrm{T}}\mathbf{x}_t$, we have, as desired,

$$\widetilde{\mathbf{u}}_t \cdot \widetilde{\mathbf{x}}_t = \widetilde{\mathbf{u}}_t \cdot \big(\mathbf{U}^{\mathrm{T}}\mathbf{x}_t\big) = \big(\mathbf{U}\widetilde{\mathbf{u}}_t\big) \cdot \mathbf{x}_t = \widehat{\mathbf{u}}_t \cdot \mathbf{x}_t\,.$$

Now, to prove $\mathbf{U}\widetilde{\mathbf{u}}_t = \widehat{\mathbf{u}}_t$, we resort to the closed-form expression (7), which gives that

$$\mathbf{U}\widetilde{\mathbf{u}}_t = \mathbf{U}\left(\lambda \sum_{s=1}^{T} \widetilde{\mathbf{x}}_s\widetilde{\mathbf{x}}_s^{\mathrm{T}} + \sum_{s=1}^{t} \widetilde{\mathbf{x}}_s\widetilde{\mathbf{x}}_s^{\mathrm{T}}\right)^\dagger \sum_{s=1}^{t-1} y_s\widetilde{\mathbf{x}}_s = \mathbf{U}\Big(\mathbf{U}^{\mathrm{T}}(\lambda\mathbf{G}_T + \mathbf{G}_t)\mathbf{U}\Big)^\dagger \mathbf{U}^{\mathrm{T}}\mathbf{b}_{t-1}\,.$$

To simplify this expression, we use twice the property of Moore-Penrose pseudoinverses stated in Corollary 19 (b), once with $\mathbf{M} = \mathbf{U}$ and the second time with $\mathbf{N} = \mathbf{U}^{\mathrm{T}}$, which both satisfy the required condition for Corollary 19 (b), as well as the matrix equalities in Corollary 19 (c), and we get

$$\mathbf{U}\Big(\mathbf{U}^{\mathrm{T}}(\lambda\mathbf{G}_T + \mathbf{G}_t)\mathbf{U}\Big)^\dagger \mathbf{U}^{\mathrm{T}} = \Big(\mathbf{U}\mathbf{U}^{\mathrm{T}}(\lambda\mathbf{G}_T + \mathbf{G}_t)\mathbf{U}\mathbf{U}^{\mathrm{T}}\Big)^\dagger = (\lambda\mathbf{G}_T + \mathbf{G}_t)^\dagger\,,$$

where the last equality comes from

$$\mathbf{U}\mathbf{U}^{\mathrm{T}}(\lambda\mathbf{G}_T + \mathbf{G}_t)\mathbf{U}\mathbf{U}^{\mathrm{T}} = \lambda\mathbf{G}_T + \mathbf{G}_t\,. \tag{30}$$

Indeed, from $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_{r_T}$ we get $\mathbf{U}\mathbf{U}^{\mathrm{T}} = P_{\mathrm{Im}(\mathbf{G}_T)}$, the orthogonal projector on the image of $\mathbf{G}_T$; we recall in (39) why $\mathrm{Im}(\mathbf{G}_t) \subseteq \mathrm{Im}(\mathbf{G}_T)$, which implies $\mathbf{U}\mathbf{U}^{\mathrm{T}}(\lambda\mathbf{G}_T + \mathbf{G}_t) = \lambda\mathbf{G}_T + \mathbf{G}_t$.

Transposing this leads to $(\lambda \mathbf{G}_T + \mathbf{G}_t)\mathbf{U}\mathbf{U}^{\mathrm{T}} = \lambda \mathbf{G}_T + \mathbf{G}_t$, from which the desired equality (30) follows by a left multiplication again by $\mathbf{U}\mathbf{U}^{\mathrm{T}} = P_{\mathrm{Im}(\mathbf{G}_T)}$.

We may now apply the bound of the Theorem 7 in the full rank case on feature sequences $\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_T \in \mathbb{R}^{r_T}$ and observations $y_1, \ldots, y_T \in [-B, B]$; this is because the associated Gram matrix $\mathbf{U}^{\mathrm{T}}\mathbf{G}_T\mathbf{U} = \mathbf{\Sigma}$ is now full rank. We get, for all $\mathbf{v} \in \mathbb{R}^{r_T}$,

$$\sum_{t=1}^{T}(y_t - \widehat{\mathbf{u}}_t \cdot \mathbf{x}_t)^2 = \sum_{t=1}^{T}(y_t - \widetilde{\mathbf{u}}_t \cdot \widetilde{\mathbf{x}}_t)^2 \leqslant \sum_{t=1}^{T}(y_t - \mathbf{v} \cdot \widetilde{\mathbf{x}}_t)^2 + \lambda T B^2 + r_T B^2 \ln\left(1 + \frac{1}{\lambda}\right). \quad (31)$$

To conclude the proof, its only remains to show that

$$\inf_{\mathbf{v} \in \mathbb{R}^{r_T}} \sum_{t=1}^{T}(y_t - \mathbf{v} \cdot \widetilde{\mathbf{x}}_t)^2 = \inf_{\mathbf{u} \in \mathbb{R}^d} \sum_{t=1}^{T}(y_t - \mathbf{u} \cdot \mathbf{x}_t)^2 . \quad (32)$$

Now, a basic argument of linear algebra, recalled in (38) of Appendix E, indicates $\mathrm{Im}(\mathbf{G}_t) = \mathrm{Im}(\mathbf{X}_t)$. Together with the inclusion $\mathrm{Im}(\mathbf{G}_t) \subseteq \mathrm{Im}(\mathbf{G}_T)$ and the fact that $\mathbf{U}\mathbf{U}^{\mathrm{T}} = P_{\mathrm{Im}(\mathbf{G}_T)}$, both already used above, we get $\mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{x}_t = \mathbf{x}_t$. A direct consequence is that for any $\mathbf{u}$ in $\mathbb{R}^d$,

$$\mathbf{u} \cdot \mathbf{x}_t = \mathbf{u} \cdot \big(\mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{x}_t\big) = \big(\mathbf{U}^{\mathrm{T}}\mathbf{u}\big) \cdot \big(\mathbf{U}^{\mathrm{T}}\mathbf{x}_t\big) = \big(\mathbf{U}^{\mathrm{T}}\mathbf{u}\big) \cdot \widetilde{\mathbf{x}}_t \,,$$

from which (32) follows, by considering $\mathbf{v} = \mathbf{U}^{\mathrm{T}}\mathbf{u}$ and by the surjectivity of $\mathbf{U}^{\mathrm{T}}$ onto $\mathbb{R}^{r_T}$ (recall that $\mathbf{U}$ and $\mathbf{U}^{\mathrm{T}}$ are of rank r_T). ■

Appendix D. Proof of Theorem 11

We recall in Appendix E many basic properties of Gram matrices and Moore-Penrose pseudoinverses to be used in the proof below.

Proof We successively prove the following two inequalities,

$$\mathcal{R}_T(\mathbf{u}) \leqslant \sum_{t=1}^{T} y_t^2 \mathbf{x}_t^{\mathrm{T}} \mathbf{G}_t^{\dagger} \mathbf{x}_t \leqslant B^2 \sum_{k=1}^{r_T} \ln\big(\lambda_k(\mathbf{G}_T)\big) + B^2 \sum_{t \in [\![1,T]\!] \cap \mathcal{T}} \ln\left(\frac{1}{\lambda_{r_t}(\mathbf{G}_t)}\right) + r_T B^2 \,, \quad (33)$$

where actually, the first inequality is a classic inequality already proved by Forster and Warmuth (2003, Theorem 3.2). We provide its derivation for the sake of completeness only.

Proof of the first inequality in (33). We obtain it as a limit case. To do so, we start by exactly rewriting the first inequality of (24), where a $\lambda > 0$ regularization factor was considered:

$$\sum_{t=1}^{T}\big(y_t - \mathbf{x}_t^{\mathrm{T}}(\lambda \mathbf{I}_d + \mathbf{G}_t)^{-1}\mathbf{b}_{t-1}\big)^2 - \sum_{t=1}^{T}(y_t - \mathbf{u}\cdot\mathbf{x}_t)^2 \leqslant \sum_{t=1}^{T} y_t^2\, \mathbf{x}_t^{\mathrm{T}}(\lambda \mathbf{I}_d + \mathbf{G}_t)^{-1}\mathbf{x}_t + \lambda \left\|\mathbf{u}\right\|^2 . \quad (34)$$

Since

$$\mathbf{G}_t = \mathbf{X}_t\mathbf{X}_t^{\mathrm{T}} \qquad \text{where} \qquad \mathbf{X}_t = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_t \end{bmatrix}$$

we note that $\mathbf{x}_t^\intercal(\lambda \mathbf{I}_d + \mathbf{G}_t)^{-1}$ is the last line of the matrix $\mathbf{X}_t^\intercal\big(\lambda \mathbf{I}_d + \mathbf{X}_t\mathbf{X}_t^\intercal\big)^{-1}$, which tends to $\mathbf{X}^\dagger$ when $\lambda \to 0$ as indicated by Corollary 19 (d). Now, $\mathbf{X}^\dagger = \mathbf{X}_t^\intercal\big(\mathbf{X}_t\mathbf{X}_t^\intercal\big)^\dagger = \mathbf{X}_t^\intercal\mathbf{G}_t^\dagger$ by Corollary 19 (a), thus

$$\lim_{\lambda\to 0} \mathbf{x}_t^\intercal(\lambda \mathbf{I}_d + \mathbf{G}_t)^{-1} = \mathbf{x}_t^\intercal\mathbf{G}_t^\dagger\,.$$

Therefore, the desired inequality for the considered forecaster,

$$\mathcal{R}_T(\mathbf{u}) = \sum_{t=1}^T \big(y_t - \mathbf{x}_t^\intercal\mathbf{G}_t^\dagger\mathbf{b}_{t-1}\big)^2 - \sum_{t=1}^T (y_t - \mathbf{u}\cdot\mathbf{x}_t)^2 \leqslant \sum_{t=1}^T y_t^2\mathbf{x}_t^\intercal\mathbf{G}_t^\dagger\mathbf{x}_t\,,$$

is obtained by taking the limit $\lambda \to 0$ in (34).

Proof of the second inequality in (33). The first part of our derivation is similar to what is performed in Luo et al. (2016, Theorem 4 of Appendix D), while the second part slightly improves on their result thanks to a more careful analysis using however the same ingredients.

Because $y_t^2 \leqslant B^2$, we only need to prove

$$\sum_{t=1}^T \mathbf{x}_t^\intercal\mathbf{G}_t^\dagger\mathbf{x}_t \leqslant \sum_{k=1}^{r_T} \ln\big(\lambda_k(\mathbf{G}_T)\big) + \sum_{t\in\mathcal{T}\cap[\![1,T]\!]} \ln\left(\frac{1}{\lambda_{r_t}(\mathbf{G}_t)}\right) + r_T\,.$$

Now, Lemma 16 below shows that

$$\sum_{t=1}^T \mathbf{x}_t^\intercal\mathbf{G}_t^\dagger\mathbf{x}_t = \sum_{t=1}^T \left(1 - \prod_{k=1}^{r_t} \frac{\lambda_k(\mathbf{G}_{t-1})}{\lambda_k(\mathbf{G}_t)}\right);$$

we assumed with no loss of generality that $\mathbf{x}_1$ is not the null vector, hence all $\mathbf{G}_t$ are at least of rank 1. Indeed, when $\mathbf{x}_t$ is the null vector, all linear combinations result in the same prediction equal to 0 and incur the same instantaneous quadratic loss.

Now, given the definition of the set $\mathcal{T}$, whose cardinality is r_T, we have $\lambda_{r_t}(\mathbf{G}_{t-1}) = 0$ when $t \in \mathcal{T}$ (and this includes $t = 1$, with the convention that $\mathbf{G}_0$ is the null matrix), while $r_{t-1} = r_t$ if $t \notin \mathcal{T}$. Therefore,

$$\begin{aligned}\sum_{t=1}^T \mathbf{x}_t^\intercal\mathbf{G}_t^\dagger\mathbf{x}_t &\leqslant \sum_{t\in\mathcal{T}\cap[\![1,T]\!]} \left(1 - \prod_{k=1}^{r_t} \frac{\lambda_k(\mathbf{G}_{t-1})}{\lambda_k(\mathbf{G}_t)}\right) + \sum_{t\in[\![1,T]\!]\setminus\mathcal{T}} \left(1 - \prod_{k=1}^{r_t} \frac{\lambda_k(\mathbf{G}_{t-1})}{\lambda_k(\mathbf{G}_t)}\right)\\ &= r_T + \sum_{t\in[\![1,T]\!]\setminus\mathcal{T}} \left(1 - \frac{D_{t-1}}{D_t}\right),\end{aligned}$$

where $D_t = \prod_{k=1}^{r_t} \lambda_k(\mathbf{G}_t)$ is the product of the positive eigenvalues of $\mathbf{G}_t$.

Now (this is where our analysis is more careful), using $1 - u \leqslant -\ln u$ for $u > 0$, we get an almost telescoping sum,

$$\sum_{t\in[\![1,T]\!]\setminus\mathcal{T}} \left(1 - \frac{D_{t-1}}{D_t}\right) \leqslant \sum_{t\in[\![1,T]\!]\setminus\mathcal{T}} \ln\frac{D_t}{D_{t-1}} = \ln\frac{D_T}{D_1} + \sum_{t\in\mathcal{T}\cap[\![2,T]\!]} \ln\frac{D_{t-1}}{D_t}$$

(note that we dealt separately with $t = 1$, which belongs to $\mathcal{T}$). Because eigenvalues cannot decrease with t, see (40), we have in particular $\lambda_k(\mathbf{G}_{t-1}) \leqslant \lambda_k(\mathbf{G}_t)$ for all $1 \leqslant k \leqslant r_t - 1$. Thus, for $t \in \mathcal{T}$ with $t \neq 1$, we have

$$\ln \frac{D_{t-1}}{D_t} \leqslant \ln\left(\frac{1}{\lambda_{r_t}(\mathbf{G}_t)}\right),$$

Substituting the definition of D_T and the equality $D_1 = \lambda_{r_1}(\mathbf{G}_1)$, and collecting all bounds together leads to the second inequality in (33). ■

The lemma below was essentially stated and proved by Cesa-Bianchi et al. (2005, Lemma D.1).

Lemma 16 (Rewriting of $\mathbf{x}^\mathsf{T}\mathbf{A}^\dagger\mathbf{x}$) *Let $\mathbf{B}$ be a $d \times d$ symmetric positive semidefinite matrix (possibly the null matrix), let $\mathbf{x} \in \mathbb{R}^d$, and and let $\mathbf{A} = \mathbf{B} + \mathbf{x}\mathbf{x}^\mathsf{T}$. Denote by r the rank of $\mathbf{A}$ and assume that $r \geqslant 1$. Then*

$$\mathbf{x}^\mathsf{T}\mathbf{A}^\dagger\mathbf{x} = 1 - \prod_{k=1}^{r} \frac{\lambda_k(\mathbf{B})}{\lambda_k(\mathbf{A})}. \tag{35}$$

Proof This lemma is a consequence of the less general Lemma 15. As a consequence of the spectral theorem applied to the symmetric matrix $\mathbf{A}$, there exists a matrix $\mathbf{U}$ of size $d \times r$ and a full rank square matrix $\mathbf{\Sigma}$ of size $r \times r$ such that $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}_r$ and $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}$. We can and will even impose that the matrix $\mathbf{\Sigma}$ is diagonal, with diagonal values equal to $\lambda_1(\mathbf{A}), \ldots, \lambda_r(\mathbf{A})$, the positive eigenvalues of $\mathbf{A}$. Let $\mathbf{\Gamma} = \mathbf{\Sigma} - \mathbf{U}^\mathsf{T}\mathbf{x}(\mathbf{U}^\mathsf{T}\mathbf{x})^\mathsf{T}$. Lemma 15 with $\mathbf{\Gamma}$, $\mathbf{\Sigma}$ and $\mathbf{U}^\mathsf{T}\mathbf{x}$ indicates that

$$\mathbf{x}^\mathsf{T}\big(\mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\big)\mathbf{x} = \big(\mathbf{U}^\mathsf{T}\mathbf{x}\big)^\mathsf{T}\mathbf{\Sigma}^{-1}\big(\mathbf{U}^\mathsf{T}\mathbf{x}\big) = 1 - \frac{\det(\mathbf{\Gamma})}{\det(\mathbf{\Sigma})} \qquad \text{where} \qquad \det(\mathbf{\Sigma}) = \prod_{k=1}^{r} \lambda_k(\mathbf{A}).$$

Now, it can be easily checked (by noting that all four properties in Proposition 18 are satisfied) that $\mathbf{A}^\dagger = \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}$, so that from the above equality, it suffices to show that

$$\det(\mathbf{\Gamma}) = \prod_{k=1}^{r} \lambda_k(\mathbf{B})$$

to conclude the proof. To do so, we first remark that $\mathbf{B} = \mathbf{A} - \mathbf{x}\mathbf{x}^\mathsf{T} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\mathsf{T} - \mathbf{x}\mathbf{x}^\mathsf{T}$, which yields

$$\mathbf{U}^\mathsf{T}\mathbf{B}\mathbf{U} = \mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\mathbf{U} - \mathbf{U}^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{U} = \mathbf{\Sigma} - \mathbf{U}^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{U} = \mathbf{\Gamma}.$$

Using again that $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}_r$, we note that $\mathbf{u}^\mathsf{T}\mathbf{u} = (\mathbf{U}\mathbf{u})^\mathsf{T}\mathbf{U}\mathbf{u}$ for all $\mathbf{u} \in \mathbb{R}^r$. From this and $\mathbf{U}^\mathsf{T}\mathbf{B}\mathbf{U} = \Gamma$, we get in particular

$$\sup_{0 \neq \mathbf{u} \in \mathbb{R}^r} \frac{\mathbf{u}^\mathsf{T}\mathbf{\Gamma}\mathbf{u}}{\mathbf{u}^\mathsf{T}\mathbf{u}} = \sup_{0 \neq \mathbf{u} \in \mathbb{R}^r} \frac{(\mathbf{U}\mathbf{u})^\mathsf{T}\mathbf{B}(\mathbf{U}\mathbf{u})}{(\mathbf{U}\mathbf{u})^\mathsf{T}\mathbf{U}\mathbf{u}}. \tag{36}$$

Next we show that

$$\sup_{0 \neq \mathbf{u} \in \mathbb{R}^r} \frac{(\mathbf{U}\mathbf{u})^\mathsf{T}\mathbf{B}(\mathbf{U}\mathbf{u})}{(\mathbf{U}\mathbf{u})^\mathsf{T}\mathbf{U}\mathbf{u}} = \sup_{0 \neq \mathbf{v} \in \mathbb{R}^d} \frac{\mathbf{v}^\mathsf{T}\mathbf{B}(\mathbf{v})}{\mathbf{v}^\mathsf{T}\mathbf{v}}, \tag{37}$$

which indicates, together with (36) and the characterization (41) of the eigenvalues of symmetric positive semidefinite matrices, that $\mathbf{B}$ and $\mathbf{\Gamma}$ have the same top r eigenvalues, as claimed. Now, to show (37), we recall that for a symmetric matrix $\mathbf{B}$, we have $\mathbb{R}^d = \ker(\mathbf{B}) \oplus \mathrm{Im}(\mathbf{B})$, so that,

$$\sup_{0 \neq \mathbf{v} \in \mathbb{R}^d} \frac{\mathbf{v}^{\mathrm{T}} \mathbf{B}(\mathbf{v})}{\mathbf{v}^{\mathrm{T}} \mathbf{v}} = \sup_{0 \neq \mathbf{v} \in \mathrm{Im}(\mathbf{B})} \frac{\mathbf{v}^{\mathrm{T}} \mathbf{B}(\mathbf{v})}{\mathbf{v}^{\mathrm{T}} \mathbf{v}} \,.$$

This leads to (37) via the inclusions

$$\mathrm{Im}(\mathbf{B}) \subseteq \mathrm{Im}(\mathbf{U}) \subseteq \mathbb{R}^d$$

which themselves follow from the inclusions

$$\mathrm{Im}(\mathbf{B}) \subseteq \mathrm{Im}(\mathbf{A}) \subseteq \mathrm{Im}(\mathbf{U}) \,.$$

Indeed, $\mathrm{Im}(\mathbf{A}) \subseteq \mathrm{Im}(\mathbf{U})$ because $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\mathrm{T}}$ and $\mathrm{Im}(\mathbf{B}) \subseteq \mathrm{Im}(\mathbf{A})$, or equivalently, given that we are considering symmetric matrices, $\ker \mathbf{A} \subseteq \ker \mathbf{B}$, as for all $\mathbf{y} \in \mathbb{R}^d$,

$$\mathbf{A}\mathbf{y} = 0 \implies \mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{y} = 0 \implies \Big[\mathbf{y}^{\mathrm{T}}\mathbf{B}\mathbf{y} = 0 \text{ and } \mathbf{y}^{\mathrm{T}}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{y} = 0\Big] \implies \sqrt{\mathbf{B}}\mathbf{y} = 0 \implies \mathbf{B}\mathbf{y} = 0 \,,$$

where we used $\mathbf{A} = \mathbf{B} + \mathbf{x}\mathbf{x}^{\mathrm{T}}$ to get the second implication, and where we multiplied $\sqrt{\mathbf{B}}\mathbf{y}$ by $\sqrt{\mathbf{B}}$ to get the final implication. ■

Appendix E. Some basic facts of linear algebra

We gather in this appendix some useful results of linear algebra, that are either reminder of well-known facts or are easy to prove (yet, we prefer prove them here rather for the proofs above to be more focused).

E.1. Gram matrices versus matrices of features

Recall that we denoted by

$$\mathbf{X}_t = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_t \end{bmatrix}$$

the $d \times t$ matrix consisting the first t features. By definition,

$$\mathrm{Im}(\mathbf{X}_t) = \mathrm{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_t\} \qquad \text{and} \qquad \mathbf{G}_t = \mathbf{X}_t \mathbf{X}_t^{\mathrm{T}} \,.$$

The aim of this section is to show that, for all $t \geqslant 1$,

$$\mathrm{Im}(\mathbf{G}_t) = \mathrm{Im}(\mathbf{X}_t) \,. \tag{38}$$

which in turn implies that for all $t \geqslant 2$,

$$\mathrm{Im}(\mathbf{G}_{t-1}) \subseteq \mathrm{Im}(\mathbf{G}_t) \,, \tag{39}$$

and that $\mathrm{rank}(\mathbf{G}_{t-1})$ and $\mathrm{rank}(\mathbf{G}_t)$ differ from at most 1.

First, as for any (not necessarily square) matrix $\mathbf{M}$ we have $\mathrm{Im}(\mathbf{M}) = \ker\big(\mathbf{M}^{\mathrm{T}}\big)^{\perp}$, we note that (38) is equivalent to $\ker\big(\mathbf{G}_t\big)^{\perp} = \ker\big(\mathbf{X}_t^{\mathrm{T}}\big)^{\perp}$, thus to $\ker\big(\mathbf{G}_t\big) = \ker\big(\mathbf{X}_t^{\mathrm{T}}\big)$. It is clear by definition of $\mathbf{G}_t$ that $\ker\big(\mathbf{X}_t^{\mathrm{T}}\big) \subseteq \ker\big(\mathbf{G}_t\big)$; furthermore, for any vector $\mathbf{u} \in \mathbb{R}^d$, we have the equality $\mathbf{u}^{\mathrm{T}}\mathbf{G}_t\mathbf{u} = \|\mathbf{X}_t^{\mathrm{T}}\mathbf{u}\|^2$, which yields the opposite inclusion $\ker\big(\mathbf{G}_t\big) \subseteq \ker\big(\mathbf{X}_t^{\mathrm{T}}\big)$.

The inclusion (39) follows from (38) as by definition, the image of $\mathbf{X}_t$ is generated by the image of $\mathbf{X}_{t-1}$ and $\mathbf{x}_t$.

E.2. Dynamic of the eigenvalues of Gram matrices

The above result gives us an idea of how eigenspaces and eigenvalues of the covariance matrix evolve. Another relationship is the following one: for $t \geqslant 1$,

$$\lambda_k\big(\mathbf{G}_{t-1}\big) \leqslant \lambda_k\big(\mathbf{G}_t\big)\,, \tag{40}$$

where we recall that $\lambda_k\big(\mathbf{G}_t\big)$ denotes the k$^{\text{th}}$ eigenvalue of $\mathbf{G}_t$ in decreasing order. To prove this we remark that for all $\mathbf{u} \in \mathbb{R}^d$, we have

$$\mathbf{u}^{\mathrm{T}}\mathbf{G}_{t-1}\mathbf{u} \leqslant \mathbf{u}^{\mathrm{T}}\mathbf{x}_t\mathbf{u} + \mathbf{u}^{\mathrm{T}}\mathbf{G}_{t-1}\mathbf{u} = \mathbf{u}^{\mathrm{T}}\mathbf{G}_t\mathbf{u}$$

and use the fact that for all symmetric positive semidefinite matrices $\mathbf{M}$,

$$\lambda_k(\mathbf{M}) = \max\left\{ \min_{\mathbf{u}} \left\{ \frac{\mathbf{u}^{\mathrm{T}}\mathbf{M}\mathbf{u}}{\mathbf{u}^{\mathrm{T}}\mathbf{u}} \mid \mathbf{u} \in U \text{ and } \mathbf{u} \neq 0 \right\} \,\middle|\, U \text{ vector space with } \dim(U) = k \right\} \tag{41}$$

E.3. Moore-Penrose pseudoinverses: definition and basic properties

In this appendix, we recall the definition and some basic properties of the Moore-Penrose pseudoinverse. It was introduced by E.H. Moore in 1920 and is a generalization of the inverse operator for non-invertible (and non-square) matrices.

Definition 17 (Moore-Penrose pseudoinverse) *The Moore-Penrose pseudoinverse of an $m \times n$ matrix $\mathbf{M}$ is a $n \times m$ matrix denoted by $\mathbf{M}^\dagger$ and defined as*

$$\mathbf{M}^\dagger \stackrel{\text{def}}{=} \lim_{\alpha \to 0} \big(\mathbf{M}^{\mathrm{T}}\mathbf{M} + \alpha I_n\big)^{-1}\mathbf{M}^{\mathrm{T}}\,,$$

where $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is the identity matrix and $\alpha \to 0$ while $\alpha > 0$.

We have the following characterization of $\mathbf{M}^\dagger$.

Proposition 18 *Let $\mathbf{M}$ be a $m \times n$ matrix. Its Moore-Penrose pseudoinverse $\mathbf{M}^\dagger$ is unique and is characterized as the only $n \times m$ matrix simultaneously satisfying the following four properties:*

(P1)	$\mathbf{M}\mathbf{M}^\dagger\mathbf{M} = \mathbf{M}$	*(P3)*	$\big(\mathbf{M}\mathbf{M}^\dagger\big)^{\mathrm{T}} = \mathbf{M}\mathbf{M}^\dagger$
(P2)	$\mathbf{M}^\dagger\mathbf{M}\mathbf{M}^\dagger = \mathbf{M}^\dagger$	*(P4)*	$\big(\mathbf{M}^\dagger\mathbf{M}\big)^{\mathrm{T}} = \mathbf{M}^\dagger\mathbf{M}$

The proof can be found in Penrose (1955). In particular, in our analysis we use the following consequences of Proposition 18. (We leave the standard proofs to the reader.)

Corollary 19 *Let $\mathbf{M}$ be a $m \times n$ matrix and $\mathbf{N}$ a $n \times p$ matrix. Then,*

(a) $\mathbf{M}^\dagger = \mathbf{M}^{\mathrm{T}}\big(\mathbf{M}\mathbf{M}^{\mathrm{T}}\big)^\dagger$;

(b) if $\mathbf{M}^{\mathrm{T}}\mathbf{M} = \mathbf{I}_n$ *or* $\mathbf{N}\mathbf{N}^{\mathrm{T}} = \mathbf{I}_n$ *then* $\big(\mathbf{M}\mathbf{N}\big)^\dagger = \mathbf{N}^\dagger\mathbf{M}^\dagger$;

(c) if $\mathbf{M}^{\mathrm{T}}\mathbf{M} = \mathbf{I}_n$, *then* $\mathbf{M}^{\mathrm{T}} = \mathbf{M}^\dagger$ *and* $\mathbf{M} = \big(\mathbf{M}^{\mathrm{T}}\big)^\dagger$;

(d) $\mathbf{M}^\dagger = \lim_{\alpha\to 0} \mathbf{M}^{\mathrm{T}}\big(\lambda\mathbf{I}_m + \mathbf{M}\mathbf{M}^{\mathrm{T}}\big)^{-1}$;

(e) if the equation $\mathbf{M}\mathbf{x} = \mathbf{z}$ with unknown $\mathbf{z} \in \mathbb{R}^m$ admits a solution $\mathbf{x} \in \mathbb{R}^n$, then $\mathbf{M}^\dagger\mathbf{z}$ is the solution in $\mathbb{R}^n$ with minimal Euclidean norm.

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.

Peter L. Bartlett, Wouter M. Koolen, Alan Malek, Eiji Takimoto, and Manfred K. Warmuth. Minimax fixed-design linear regression. *JMLR: Workshop and Conference Proceedings*, 40:1–14, 2015. Proceedings of COLT'2015.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. A second-order perceptron algorithm. *SIAM Journal on Computing*, 34(3):640–668, 2005.

Jürgen Forster and Manfred K. Warmuth. Relative loss bounds for temporal-difference learning. *Machine Learning*, 51:23–50, 2003.

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

Richard D. Gill and Boris Y. Levit. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, 1(1–2):59–79, 1995.

Wojciech Kotłowski, Wouter M. Koolen, and Alan Malek. Random permutation online isotonic regression. In *Advances in Neural Information Processing Systems*, pages 4183–4192, 2017.

Haipeng Luo, Alekh Agarwal, Nicolò Cesa-Bianchi, and John Langford. Efficient second order online learning by sketching. In *Advances in Neural Information Processing Systems*, pages 902–910, 2016.

Alan Malek and Peter L. Bartlett. Horizon-independent minimax linear regression. In *Advances in Neural Information Processing Systems*, pages 5264–5273, 2018.

Roger Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.

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

Harry L. Van Trees. *Detection, Estimation and Modulation Theory*. Wiley & Sons, 1968.

Vladimir Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.