# Uniform regret bounds over $R^d$ for the sequential linear regression problem with the square loss

Pierre Gaillard, Sébastien Gerchinovitz, Malo Huard, Gilles Stoltz

Not yet published — This is a working paper that just happens to be written in a standard JMLR layout

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

**Pierre Gaillard** PIERRE.GAILLARD@INRIA.FR
*Inria Paris*

**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, Orsay*

## Abstract

We consider the setting of online linear regression for arbitrary deterministic sequences, with the square loss. We are interested in regret bounds that hold uniformly over all vectors $\mathbf{u} \in \mathbb{R}^d$. Vovk (2001) showed a $d \ln T$ lower bound on this uniform regret. We exhibit forecasters with closed-form regret bounds that match this $d \ln T$ quantity. To the best of our knowledge, earlier works only provided closed-form regret bounds of $2d \ln T + \mathcal{O}(1)$.

**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 outputs $(y_t)_{t\geqslant 1}$ in $\mathbb{R}$ and a sequence of input vectors $(\mathbf{x}_t)_{t\geqslant 1}$ in $\mathbb{R}^d$. The output sequence $(y_t)_{t\geqslant 1}$ is initially hidden to the learner, while the input sequence may be given in advance or be initially hidden as well, depending on the setting considered: "beforehand-known features" 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 output $y_t \in \mathbb{R}$ is then revealed and instance $t+1$ starts. 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 ,

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

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

$$\sup_{\mathbf{u}\in\mathbb{R}^d} \mathcal{R}_T(\mathbf{u}) \,,$$

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 $y_t$.

The two settings described above are summarized in 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 $y_t \in \mathbb{R}$ | • observes $y_t \in \mathbb{R}$ |
| • 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

**Literature review.** 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 $\ell_1$–norm. We refer the interested reader to the monograph (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 (2), which achieves a regret of order $\mathcal{O}(d \ln T)$ uniformly over vectors $\mathbf{u}$ with bounded $\ell_2$–norm. Vovk (2001) also provided a matching minimax lower bound $dB^2 \ln T - \mathcal{O}(1)$ on the worst-case uniform regret over $\mathbb{R}^d$ of any forecaster, where $B$ is a bound on the outputs $|y_t|$. 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 assumptions on the sequences $(\mathbf{x}_t)_{t \geqslant 1}$ and $(y_t)_{t \geqslant 1}$ of features and observations. 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 (2017, Chapter 3) shows 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), which will be a building block for our new analyses. We also state and re-prove (Section 2.2) the regret lower bound by Vovk (2001), 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 constant equals 1. Our proof reuses several ideas from the original proof (features $\mathbf{x}_t$ with only one non-zero input equal to 1, Bernoulli observations $y_t$, the use of randomization with a Beta prior distribution, etc.). The main difference lies in the exposition: we resort 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 analysis 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 $t$, 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 (9) to get an optimal $dB^2 \ln(1 + T/d) + dB^2$ bound on the uniform regret over $\mathbb{R}^d$. To the best of our knowledge, this is perhaps the first optimal closed-form regret bound for the uniform regret: previous closed-form bounds typically had an extra 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 7 of Section 3.

Question then is (Section 4) whether a $dB^2 \ln T + \mathcal{O}(1)$ regret bound can be achieved on the uniform regret in the most interesting setting of sequentially revealed features. Among the several approaches taken, one worked and surprised us; it provides a new understanding of this well-known algorithm. Indeed, it turns out that 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 $dB^2 \ln T + \mathcal{O}_T(1)$. The leading term is thus the optimal one. Also, no parameter needs to be tuned, which is a true relief. The only drawback of this bound, compared to other bounds mentioned above, is that the $\mathcal{O}_T(1)$ remainder term depends on the sequence of features. For each sequence of features, it is a constant, but that constant may be large. But all in all, we think that it was worth proving that the regularization term $\lambda \|\mathbf{u}\|^2$ in the defining equation (2) of the non-linear ridge regression algorithm is not so useful, while the seemingly harmless regularization term $(\mathbf{u} \cdot \mathbf{x}_t)^2$ is crucial.

## 2. Sequentially revealed features / Known results

In this section, we recall and reestablish some known results regarding the regret with the quadratic 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). 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 indicates that the upper bounds $dB^2 \ln T + \mathcal{O}(1)$ obtained later in this article are first-order optimal; in particular, they get the optimal $dB^2$ constant.

### 2.1. Upper bound on the regret / Reminder of a known result

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. Note that this definition 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 $\Gamma$), the the vector $\widehat{\mathbf{u}}_t$ used should also be just divided by $\gamma$ (or multiplied by $\Gamma^{-1}$). We may similarly define what a scale invariant bound is.

**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 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 $r_t = \mathrm{rank}(\mathbf{G}_t)$ 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 as 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 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 $\mathbf{u} \in \mathbb{R}^d$,*

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

Question is whether this regret bound, which is seemingly non uniform over $\mathbb{R}^d$, can lead to a uniform regret bound. The answer is yes, but the best bound which we could obtain from it is of the form $2dB^2 \ln(T) + \mathcal{O}_T(1)$, and holds under an additional boundedness assumption: there exists $X > 0$ such that $\|\mathbf{x}_t\| \leqslant X$ for all $t \geqslant 1$. However, as we show in the next sections the constant 2 in the leading term $2dB^2 \ln T$ of the regret is suboptimal. Despite all our efforts, we were unable to get the known-to-be optimal constant 1 for the non-linear ridge regression algorithm. But working on the derivation below and trying to improve on it, we designed the forecaster of Section 3, which achieves the optimal constant 1 in its uniform regret bound.

**Corollary 3** *Let the non-linear ridge regression 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 $\|\mathbf{x}_t\| \leqslant X$ 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 $\lambda$ 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 $T/\lambda$ 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

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

**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 for the proof of Theorem 6.

Theorem 2 indicates that

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

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

$$\begin{aligned}\|\mathbf{u}^\star\| = \left\|\mathbf{G}_T^{-1/2}\,\mathbf{G}_T^{-1/2}\mathbf{b}_T\right\| \leqslant \lambda_1\big(\mathbf{G}_T^{-1/2}\big)\,\left\|\mathbf{G}_T^{-1/2}\mathbf{b}_T\right\| &= \frac{1}{\sqrt{\lambda_d(\mathbf{G}_T)}}\,\left\|\mathbf{G}_T^{-1/2}\mathbf{b}_T\right\| \\ &\leqslant \frac{1}{\sqrt{\lambda_d(\mathbf{G}_T)}}\,B\sqrt{T}\,, \end{aligned} \tag{4}$$

where we used, for the final inequality, an elementary argument of orthogonal projection that is at the heart of the proof of Theorem 6: see (15) and the sentence after it. In addition, Jensen's inequality (or the alternative treatement 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

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

and the claimed bound follows by substituting the bound (4). ■

### 2.2. Lower bound on the uniform regret / Known result but new proof

In this section, we study the uniform regret $\sup\{\mathcal{R}_T(\mathbf{u}) : \mathbf{u}\in\mathbb{R}^d\}$, in the worst case, that is,

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

where the first infimum is over all forecasters (all forecasting strategies) and the supremum is over all individual sequences $(\mathbf{x}_t, y_t)$ in $[0,1]^d \times [-B,B]$, where $t \geqslant 1$. Vovk (2001, Theorem 2) indicated a result of the following form.

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

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 $\mathbb{R}^d$, of order each 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.

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. The main difference lies in the exposition: we resort 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, 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 some sense, our argument is more generic and works for any strategy, not only for the optimal one. 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. It indeed holds for any estimator (unlike the Cramér-Rao bound, which only holds for unbiased estimators).

**Remark 5** Note that the sequences considered in the definition of $\mathcal{R}^\star_{T,[-B,B]}$ are fixed beforehand, they do not need to be constructed in an adaptive way to fool the considered forecaster. Note also that the features $x_t$ 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 6, where no boundedness assumption is required on the features.

**Proof** We start with a case where $y_t \in [0,1]$ and explain at the end of the proof (in Appendix A) how to draw the result for the desired case where $y_t \in [-B,B]$.

We fix a forecaster and consider the following randomization over the possible individual sequences. Given $\theta^\star \in [0,1]^d$, we define a joint distribution $\mathcal{P}_{\theta^\star}$ on $[0,1]^d \times \{0,1\}$ as the distribution of the pair $(\mathbf{e}_J, Y)$ where $J$ is uniformly distributed over $\{1,\dots,d\}$, where $\mathbf{e}_j$ denotes the unit vector $(0,\dots,0,1,0,\dots,0)^{\mathrm{T}}$ along the $j$-th coordinate (the 1 is in position $j$), and where $Y$ has a conditional distribution with respect to $J$ given by a Bernoulli distribution with parameter $\theta^\star \cdot \mathbf{e}_J = \theta^\star_J$. We consider a sequence of i.i.d. pairs $(J_t, Y_t)$, for $t = 1, 2, \dots$.

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; and 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.

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

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

Since $\widehat{y}_t$ is measurable w.r.t. $\mathcal{F}_{t-1}$, the $\sigma$–algebra generated by $J_t$ and the $(J_s, Y_s)$ where $s \leqslant t-1$, a conditional bias–variance decomposition yields

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

where we used first that by construction, $\widehat{y}_t = \widehat{\mathbf{u}}_t \cdot \mathbf{e}_{J_t} = \widehat{u}_{J_t,t}$, and second, that the conditional distribution of $Y_t$ is a Bernoulli distribution with parameter $\theta^\star_{J_t}$. Similarly, for all $\mathbf{u} \in \mathbb{R}^d$,

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

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

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

Now, by resorting to the tower rule again, integrating over $J_t$ conditionally to the $(J_s, Y_s)$ where $s \leqslant t-1$, we get

$$\mathbb{E}\left[\sup_{\mathbf{u}\in\mathbb{R}^d} \mathcal{R}_T(\mathbf{u})\right] \geqslant \sum_{t=1}^T \mathbb{E}\Big[(\widehat{u}_{J_t,t} - \theta^\star_{J_t})^2\Big] = \sum_{t=1}^T \frac{1}{d}\,\mathbb{E}\Big[\,\big\|\,\widehat{\mathbf{u}}_t - \theta^\star\,\big\|_2^2\Big]\,, \tag{6}$$

and we now show that each term in the sum is larger than something of the order of $d/t$, from which the desired $d\ln T$ bound will follow.

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.

See Appendix A for full details and conclusion of this proof. ■

## 3. Beforehand-known features / New result

In this section we assume that the features are known beforehand and exhibit a simple forecaster with a 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. A uniform bound of the form $2dB^2 \ln T + \mathcal{O}_T(1)$ was already proved by Bartlett et al. (2015), in a different way (studying minimax values), see a more detailed discussion below.

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

$$\widehat{\mathbf{u}}_t \in \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^d} \left\{ \sum_{s=1}^{t-1} (y_s - \mathbf{u}\cdot\mathbf{x}_s)^2 + (\mathbf{u}\cdot\mathbf{x}_t)^2 + \lambda \sum_{s=1}^{T} (\mathbf{u}\cdot\mathbf{x}_s)^2 \right\} \tag{7}$$

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

$$\widehat{\mathbf{u}}_t = \big(\lambda\,\mathbf{G}_T + \mathbf{G}_t\big)^\dagger \mathbf{b}_{t-1}\,, \tag{8}$$

where † denotes the Moore-Penrose inverse of a matrix.

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

$$\lambda\,\|\mathbf{u}\|_{\mathbf{G}_T}^2 \stackrel{\text{def}}{=} \lambda\,\mathbf{u}^{\mathrm{T}}\mathbf{G}_T\mathbf{u} = \lambda\sum_{s=1}^{T}(\mathbf{u}\cdot\mathbf{x}_s)^2\,; \tag{9}$$

that is, this regularization term can be seen as a metric adapted to the known-in-advance features $\mathbf{x}_1,\dots,\mathbf{x}_T$. Note that this algorithm has the desirable property of being scale invariant.

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

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

*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 7** Bartlett et al. (2015) study some minimax uniform regret, namely

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

The minimax regret is also the regret of the associated minimax forecaster (which can be computed by backward induction); note that the latter forecaster strongly depends on $T$. Because of the minimax optimality, $\mathcal{R}_T^\star$ is smaller than the bound of Theorem 6.

However, this was not reflected, to the best of our understanding, in the original analysis by Bartlett et al. (2015), which rather gets a bound of order $2dB^2\ln T$, that is, which suffers from an extra multiplicative factor of 2. Indeed, Theorem 5 therein indicates, in the case where $d=1$ and $B=1$, that

$$\forall\, T \geqslant 1, \qquad \mathcal{R}_T^\star \leqslant f(T) \tag{10}$$

for any function $f : \{1,2,\dots\} \to \mathbb{R}_+$ 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 above. 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

$$e^{-f(T)/2} = \Omega\big(T^{-a/2}\big) \qquad \text{and} \qquad f(T+1) - f(T) = a \ln\left(1 + \frac{1}{T}\right) = \Omega\big(T^{-1}\big)\,,$$

hence, are not suitable choices for the bound (10).

**Remark 8** 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 (7) and (9), 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 $\mathbf{x}_1, \ldots, \mathbf{x}_T$ 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 Kotlowski 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. 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),

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

$$\leqslant \inf_{\mathbf{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}(y_t - \mathbf{u}\cdot\mathbf{x}_t)^2\right\} + \lambda T B^2 + dB^2\ln\left(1+\frac{1}{\lambda}\right). \tag{12}$$

*Proof of* (11). We first show that the strategy (2) on the $\widetilde{\mathbf{x}}_t$ leads to the same forecasts as the strategy (7) on the original $\mathbf{x}_t$; that is, we show that

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

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

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

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 $\widetilde{\mathbf{x}}_1, \dots, \widetilde{\mathbf{x}}_T \in \mathbb{R}^d$ and $y_1, \dots, y_T \in [-B, B]$, to get, for all $\mathbf{u} \in \mathbb{R}^d$,

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

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

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

Taking the infinimum over $\mathbf{u}$ in $\mathbb{R}^d$ in (13) concludes the proof of (11).

*Proof of* (12). We bound

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

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). To that end, 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

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

We conclude the proof of (12) by a change of dummy variable $\mathbf{v}=\mathbf{G}_T^{1/2}\mathbf{u}$,

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

## 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 $\lambda\left\|\mathbf{u}\right\|_{\mathbf{G}_T}^2$ in (7), which cannot be computed in advance. The *non-linear regression algorithm with almost no regularization* picks weight vectors as defined in Equations (2) or (7) with regularization parameter $\lambda=0$; that is, $\widehat{\mathbf{u}}_1=(0,\dots,0)^{\mathrm{T}}$ and for $t\geqslant 2$,

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

where the closed-form expression corresponds to (8). Note that no parameter requires to be tuned in this case, which can be a relief.

Why and how we were led to consider this forecaster is explained in the discussion following Theorem 10; we were surprised that 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 could be performed here. It provides a new understanding of this well-known non-linear regression algorithm: the regularization term $\lambda\left\|\mathbf{u}\right\|^2$ 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.

**Remark 9** The ridge regression with regularization factor $\lambda=0$, that is, the ordinary linear least-squares (OLS) regression, cannot achieve such a logarithmic bound on the regret. Even worse (but not surprisingly), its regret grows in a linear fashion. Indeed, consider $d=1$ and for a given $T$, consider the bounded sequences of scalar numbers

$$y_1=y_2=\cdots=y_T=1\,,\qquad\text{and}\qquad x_1=x_2=\cdots=x_{T-1}=\frac{1}{\sqrt{T}}\quad\text{while}\quad x_T=1\,.$$

Then OLS picks $\widehat{u}_1=0$ and $\widehat{u}_2=\cdots=\widehat{u}_T=\sqrt{T}$ and its cumulative loss satisfies

$$\sum_{t=1}^T(y_t-\widehat{u}_tx_t)^2\geqslant(y_T-\widehat{u}_Tx_T)^2=\left(1-\sqrt{T}\right)^2\sim T\,,$$

while, for the choice $v = (1+\sqrt{T})/2$,

$$\inf_{u\in\mathbb{R}} \sum_{t=1}^{T} (y_t - ux_t)^2 \leqslant \sum_{t=1}^{T} (y_t - vx_t)^2 = (T-1)\left(1 - \frac{1}{2\sqrt{T}} - \frac{1}{2}\right)^2 + \left(1 - \frac{1}{2} - \frac{\sqrt{T}}{2}\right)^2 \sim \frac{T}{2}.$$

This proves that the regret of OLS grows as $T/2$.

The proof of the regret bound below for the non-linear regression algorithm with almost no regularization can be found in Appendix D.

**Theorem 10** *For all $T \geqslant 1$, for all sequences $\mathbf{x}_1, \dots, \mathbf{x}_T \in \mathbb{R}^d$ and all $y_1, \dots, y_T \in [-B, B]$, for all $\mathbf{u} \in \mathbb{R}^d$, the non-linear regression algorithm with almost no regularization achieves the uniform regret bound*

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

*where the set $\mathcal{T}$ contains $r_T$ rounds, given by $s = 1$ and the $s \geqslant 2$ when $\operatorname{rank}(\mathbf{G}_{s-1}) \neq \operatorname{rank}(\mathbf{G}_s)$.*

Note that the regret bound obtained is scale invariant, which is natural and was expected, as the forecaster also is. The same (standard) arguments as the ones at the end of the proof of Corollary 3 show the following consequence of this bound: for all $X > 0$, for all sequences $\mathbf{x}_1, \mathbf{x}_2, \dots$ of features with $\|\mathbf{x}_t\| \leqslant X$,

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

**Discussion of this $\mathcal{O}_T(1)$ term.** The $\mathcal{O}_T(1)$ term above has two desirable properties. First, it only increases when the rank of the Gram matrix $\mathbf{G}_t$ increases, which occurs at most $d$ times. Once the matrix $\mathbf{G}_T$ is full rank, it thus stops increasing; for rounds $T' \geqslant T$, only the leading term increases to $dB^2 \ln T'$. Second, it is scale invariant in the features $\mathbf{x}_t$, which is not a surprise, at the algorithm itself is scale invariant—in constrast to the non-linear Ridge forecaster (2), which was not scale invariant mostly because of the regularization factor $\lambda$.

A drawback of this $\mathcal{O}_T(1)$ term is however that it strongly depends on the input sequence $(\mathbf{x}_t)_{t\geqslant 1}$. In particular, it is not invariant by a permutation of the $(\mathbf{x}_t)_{t\geqslant 1}$ and is not uniformly bounded over all sequences $(x_t)_{t\geqslant 1}$. The underlying issue is that the algorithm does not know in advance the correct metric to use on the feature space and may not scale back or regularize properly certain directions when they start being observed and get significant. Unfortunately, we could not get rid of this additional remainder term (we detail below some attempts made to do so). The deep reason might be that it is unavoidable due to the sequential nature of the problem: there might be a price to pay for not knowing $G_T$ in advance. The lower bound of Theorem 4 is perhaps improvable as far as its remainder terms are considered: could they explicitly depend on the sequence of features?

**Attempts made to improve on this $\mathcal{O}_T(1)$ term.** We first note that in dimension $d = 1$, the metrics endowed by $\|\cdot\|$ and $\|\cdot\|_{G_T}$ are equal up to a scaling factor. The adaptation to $\|\cdot\|_{G_T}$ thus reduces to a proper calibration of the regularization parameter $\lambda > 0$, which can be achieved without too much technicalities. But for dimensions $d \geqslant 2$, the metric endowed by $\|\cdot\|_{G_T}$ can be significantly different from the one associated with $\|\cdot\|$ because, in particular, of different scalings between directions.

*Our first attempt* to improve on this $\mathcal{O}_T(1)$ term when $d \geqslant 2$ was 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 $(\mathbf{u} \cdot \mathbf{x}_t)^2 + \lambda \|\mathbf{u}\|^2_{\mathbf{G}_T}$. 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 $\lambda$ only acts as a multiplicative bias to the vector otherwise considered. The analysis we followed led to a regret bound increasing in $\lambda$, so that we finally picked $\lambda = 0$ and ended up with our non-linear regression algorithm with almost no regularization.

*A second attempt* was to add $d$ artificial initial steps before the real game starts, indexed by $\tau = -d+1, \ldots, 0$, with features of the form $\widetilde{\mathbf{x}}_\tau = (0, \ldots, 0, \sqrt{\lambda}, 0, \ldots)$ and with observations $\widetilde{y}_\tau = 0$. This ensures that all increases of the ranks are controlled, with $\ln\bigl(1/\lambda_{r_t}(\mathbf{G}_t\bigr)$ terms all being equal to $\ln(1/\lambda)$. However, the additive price to pay in the regret bound equals $\lambda \|\mathbf{u}\|^2$ and we are thus essentially back to the bound of Theorem 2.

*A variation on this second attempt* was to ignore new directions brought by new inputs $\mathbf{x}_t$, by approximating $\mathbf{x}_t$ with its projection onto $\mathrm{Im}(\mathbf{G}_{t-1})$, and do so while these directions, which lie in $\mathrm{Ker}(G_{t-1})$, are not strong enough, i.e., have not been observed enough.

*A final attempt* aimed to discretize the space of possible Gram matrices $G_T$ together with the consideration of a meta-aggregation algorithm on the considered approximations of $G_T$. It did not correct the issues mentioned above and it worsened the constant factor of 1 in front of the leading $dB^2 \ln T$ term, not mentioning the prohibitive computational complexity associated with this approach (exponential in $d$).

## 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.

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

Alan Malek. *Efficient Sequential Decision Making*. PhD thesis, EECS Department, University of California, Berkeley, 2017.

Roger Penrose. A generalized inverse for matrices. *Mathematical proceedings of the Cambridge philosophical society*, 51(3):406–413, 1955.

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.

## Appendix A. End of the proof of Theorem 4

**Proof** We resume at (6). We now (re)index all expectations relative to $\mathcal{P}_{\theta^\star}$ by the parameter $\theta^\star$ and consider a prior $\pi$ on these $\theta^\star \in [0,1]^d$. The randomizations considered in the main body of the proof can be mixed according to $\pi$, so that we get, by mixing both sides of (6)

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

(Recall that we are considering the interval $[0,1]$ here.) Now, an immediate application of the (multi-dimentional) van Trees inequality with a Beta$(\alpha,\alpha)$ prior $\pi$ shows that for all forecasters, all $t \geqslant 1$ and $\alpha \geqslant 3$,

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

see Lemma 11 below. We thus proved

$$\begin{aligned}\mathcal{R}^\star_{T,\,[0,1]} &\geqslant \sum_{t=1}^T \frac{d}{\big(4+2/(\alpha-1)\big)t + 16d\alpha} \geqslant d\int_1^{T+1} \frac{1}{\big(4+2/(\alpha-1)\big)t + 16d\alpha}\,\mathrm{d}t \\ &= \frac{d}{4+2/(\alpha-1)} \ln \frac{\big(4+2/(\alpha-1)\big)(T+1) + 16d\alpha}{\big(4+2/(\alpha-1)\big) + 16d\alpha} \\ &\geqslant \frac{d}{4+2/(\alpha-1)} \ln \frac{\big(4+2/(\alpha-1)\big)T}{\big(4+2/(\alpha-1)\big) + 16d\alpha} \\ &= \frac{d}{4+2/(\alpha-1)}\left(\ln T - \ln\left(1 + \frac{16d\alpha}{4+2/(\alpha-1)}\right)\right) \geqslant \frac{d}{4+2/(\alpha-1)}\big(\ln T - \ln(1+4d\alpha)\big)\,,\end{aligned}$$

which we lower bound in a crude way by resorting to $1/(1+u) \geqslant 1-u$ and by taking $\alpha$ 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 $1 + 4d\alpha \leqslant 1 + 4d(1+\ln T) \leqslant 7d\ln T$. We get

$$\begin{aligned}\mathcal{R}^\star_{T,\,[0,1]} &\geqslant \frac{d}{4}\underbrace{\left(1 - \frac{1}{2(\alpha-1)}\right)}_{\geqslant 0}\big(\ln T - \ln(7d\ln T)\big) \\ &\geqslant \frac{d}{4}\left(1 - \frac{1}{2\ln T}\right)(\ln T - \ln(7d) - \ln\ln T) \geqslant \frac{d}{4}\big(\ln T - (3 + \ln d) - \ln\ln T\big)\,. \qquad (16)\end{aligned}$$

The factor 3 above corresponds to $1/2 + \ln 7 \leqslant 3$. 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 $B > 0$.

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 (16) exhibited on (5), 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 i.i.d. pairs $(J, Y)$ 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$(\alpha,\alpha)$ distributions as a prior $\pi$ 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 (19) 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 the $(J_s, Y_s)$, where $1 \leqslant s \leqslant t$, we have

$$\int_{[0,1]^d} \mathbb{E}_{\theta^\star}\Big[\,\big\|\,\widehat{\mathbf{u}}_t - \theta^\star\,\big\|_2^2\Big]\ \beta^{(d)}_{\alpha,\alpha}(\theta^\star)\ \mathrm{d}\theta^\star \geqslant \frac{(\operatorname{Tr}\mathbf{I}_d)^2}{\operatorname{Tr}\mathcal{I}\big(\beta^{(d)}_{\alpha,\alpha}\big) + t\displaystyle\int_{[0,1]^d}\big(\operatorname{Tr}\mathcal{I}(\theta^\star)\big)\,\beta^{(d)}_{\alpha,\alpha}(\theta^\star)\ \mathrm{d}\theta^\star}\,, \quad (17)$$

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 (18), while each component $(i,i)$ of the other matrix in the denominator is given by

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

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

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

**Proof** We denote by

$$f_{\theta^\star} : (j,y) \in \{1,\dots,d\}\times\{0,1\} \longmapsto \frac{1}{d}\,\theta_j^{\star y}(1-\theta_j^\star)^{1-y}$$

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

The sufficient conditions of Gill and Levit (1995, Section 4) for (17) 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, $\theta^\star \mapsto f_{\theta^\star}(j,y)$ is $C^1$–smooth for all $(j,y)$, with, for all $i \in \{1,\dots,d\}$,

$$\frac{\partial}{\partial\theta_i^\star}\ln f_{\theta^\star}(J,Y) = \left(\frac{Y}{\theta_J^\star} - \frac{1-Y}{1-\theta_J^\star}\right)\mathbb{1}_{\{J=i\}}$$

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

$$\mathcal{I}(\theta^\star)_{i,i} = \mathbb{E}_{\theta^\star}\left[\left(\frac{Y}{\theta_J^\star} - \frac{1-Y}{1-\theta_J^\star}\right)^2\mathbb{1}_{\{J=i\}}\right] = \frac{1}{d}\left(\frac{1}{\theta_i^\star} + \frac{1}{1-\theta_i^\star}\right) = \frac{1}{d\,\theta_i^\star(1-\theta_i^\star)}\,, \tag{18}$$

and therefore, is such that $\theta^\star \mapsto \sqrt{\mathcal{I}(\theta^\star)}$ is locally integrable w.r.t. the Lebesgue measure.

We now compute all elements of the denominator of (17). First, by symmetry and then by substituting (18),

$$\begin{aligned}
&\int_{[0,1]^d}\big(\mathrm{Tr}\,\mathcal{I}(\theta^\star)\big)\,\beta_{\alpha,\alpha}^{(d)}(\theta^\star)\ \mathrm{d}\theta^\star \\
&= d\int_{[0,1]^d}\mathcal{I}(\theta^\star)_{1,1}\,\beta_{\alpha,\alpha}^{(d)}(\theta^\star)\ \mathrm{d}\theta^\star \\
&= d\int_{[0,1]^d}\frac{1}{d\,\theta_1^\star(1-\theta_1^\star)}\,\frac{\Gamma(2\alpha)}{\big(\Gamma(\alpha)\big)^2}\,(\theta_1^\star)^{\alpha-1}(1-\theta_1^\star)^{\alpha-1}\ \beta_{\alpha,\alpha}(\theta_2^\star)\cdots\beta_{\alpha,\alpha}(\theta_d^\star)\ \mathrm{d}\theta^\star \qquad (19)\\
&= \frac{\Gamma(2\alpha)}{\big(\Gamma(\alpha)\big)^2}\int_{[0,1]} t^{\alpha-2}(1-t)^{\alpha-2}\,\mathrm{d}t = \frac{\Gamma(2\alpha)}{\big(\Gamma(\alpha)\big)^2}\,\frac{\big(\Gamma(\alpha-1)\big)^2}{\Gamma\big(2(\alpha-1)\big)}\,,
\end{aligned}$$

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

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

Second, as far as the $\mathrm{Tr}\,\mathcal{I}\big(\beta_{\alpha,\alpha}^{(d)}\big)$ in (17) is concerned, because $\beta_{\alpha,\alpha}^{(d)}$ is a product of univariate distributions,

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

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

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

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

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

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

## Appendix B. Proof of Theorem 2

We essentially extract the proof from Cesa-Bianchi and Lugosi (2006, Chapter 11). We merely provide it because we will later need the first inequality of (20) in the proof of Theorem 10 and we wanted our submission to be self-complete. But of course, all the content of this section is extremely standard and should be skipped by any reader familiar with the basic results of sequential linear regression.

**Proof** We successively prove the following two inequalities,

$$\mathcal{R}_T(\mathbf{u}) \leqslant \lambda \|\mathbf{u}\|^2 + \sum_{t=1}^{T} y_t^2 \, \mathbf{x}_t^{\mathrm{T}} \mathbf{A}_t^{-1} \mathbf{x}_t \leqslant \lambda \|\mathbf{u}\|^2 + B^2 \sum_{k=1}^{d} \ln\left(1 + \frac{\lambda_k(\mathbf{G}_T)}{\lambda}\right) \tag{20}$$

*Proof of the first inequality in* (20). We denote by $L_{t-1}^{\text{reg}}$ the cumulative loss up to round $t-1$ included, to which we add the regularization term:

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

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

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

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, $\breve{\mathbf{u}}_1 = (0, \ldots, 0)^{\mathrm{T}}$. By the very definition of $\breve{\mathbf{u}}_{T+1}$, for all $\mathbf{u} \in \mathbb{R}^d$,

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

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

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

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

$$(y_t - \widehat{y}_t)^2 + L_{t-1}^{\text{reg}}(\breve{\mathbf{u}}_t) - L_t^{\text{reg}}(\breve{\mathbf{u}}_{t+1}) = (\breve{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t)^{\mathrm{T}} \mathbf{A}_t (\breve{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t) - (\widehat{\mathbf{u}}_t - \breve{\mathbf{u}}_t)^{\mathrm{T}} \mathbf{A}_{t-1} (\widehat{\mathbf{u}}_t - \breve{\mathbf{u}}_t) \,; \tag{21}$$

the first inequality in (20) will then be obtained, as the second term in (21) is negative and as the first term in (21) can be rewritten as $y_t^2 \, \mathbf{x}_t^{\mathrm{T}} \mathbf{A}_t^{-1} \mathbf{x}_t$ thanks to the equality (23) below, which states $\mathbf{A}_t(\breve{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t) = y_t \mathbf{x}_t$.

To prove (21), 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 $\breve{\mathbf{u}}_{t+1} = \mathbf{A}_t^{-1}\mathbf{b}_t$. Now, $L_t^{\mathrm{reg}}$ rewrites, for all $\mathbf{u} \in \mathbb{R}^d$,

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

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

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

In particular,

$$L_{t-1}^{\mathrm{reg}}(\breve{\mathbf{u}}_t) - L_t^{\mathrm{reg}}(\breve{\mathbf{u}}_{t+1}) = -y_t^2 + \breve{\mathbf{u}}_{t+1}^{\mathrm{T}}\mathbf{A}_t\breve{\mathbf{u}}_{t+1} - \breve{\mathbf{u}}_t^{\mathrm{T}}\mathbf{A}_{t-1}\breve{\mathbf{u}}_t\,. \tag{22}$$

We now expand the first term in (21). To that end, we use that from the closed-form expressions of $\widehat{\mathbf{u}}_t$ and $\breve{\mathbf{u}}_{t+1}$,

$$\mathbf{A}_t(\breve{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t) = \mathbf{A}_t\big(\mathbf{A}_t^{-1}\mathbf{b}_t - \mathbf{A}_t^{-1}\mathbf{b}_{t-1}\big) = \mathbf{b}_t - \mathbf{b}_{t-1} = y_t\mathbf{x}_t\,. \tag{23}$$

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

$$\begin{aligned}(y_t - \widehat{y}_t)^2 = y_t^2 - 2y_t\widehat{y}_t + \widehat{y}_t^2 &= y_t^2 - 2(\breve{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t)^{\mathrm{T}}\mathbf{A}_t\widehat{\mathbf{u}}_t + \widehat{\mathbf{u}}_t^{\mathrm{T}}\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}\widehat{\mathbf{u}}_t \\ &= y_t^2 - 2(\breve{\mathbf{u}}_{t+1} - \widehat{\mathbf{u}}_t)^{\mathrm{T}}\mathbf{A}_t\widehat{\mathbf{u}}_t + \widehat{\mathbf{u}}_t^{\mathrm{T}}(\mathbf{A}_t - \mathbf{A}_{t-1})\widehat{\mathbf{u}}_t\,, \end{aligned} \tag{24}$$

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 (22) and (24) together, we proved

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

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:

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

*Proof of the second inequality in* (20). 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 12 below shows that

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

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

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

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

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

**Lemma 12** *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

$$\det(\mathbf{U}) = \det\big(\mathbf{V} - \mathbf{u}\mathbf{v}^{\mathrm{T}}\big) = \det(\mathbf{V}) \det\Big(\mathbf{I}_d - \big(\mathbf{V}^{-1}\mathbf{u}\big)\mathbf{v}^{\mathrm{T}}\Big) = \det(\mathbf{V})\big(1 - \mathbf{v}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{u}\big) \,. \quad \blacksquare$$

## Appendix C. Proof of Theorem 6 in the general case

In this section we extend the proof of Theorem 6, 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 (8) is derived. We rewrite the definition equation (7) of $\widehat{\mathbf{u}}_t$ as

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

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 16 (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 6. 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 6 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 (7) 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 (8), which gives that

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

To simplify this expression, we use twice the property of Moore-Penrose inverses stated in Corollary 16 (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 16 (b), as well as the matrix equalities in Corollary 16 (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

$$\mathbf{U}\mathbf{U}^{\mathrm{T}}(\lambda\mathbf{G}_T + \mathbf{G}_t)\mathbf{U}\mathbf{U}^{\mathrm{T}} = \lambda\mathbf{G}_T + \mathbf{G}_t\,. \tag{25}$$

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 (34) 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 (25) 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 6 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 $\mathbf{v} \in \mathbb{R}^{r_T}$,

$$\sum_{t=1}^{T}(y_t - \widehat{\mathbf{u}}_t \cdot \mathbf{x}_t)^2 = \sum_{t=1}^{T}(y_t - \widetilde{\mathbf{u}}_t \cdot \widetilde{\mathbf{x}}_t)^2 \leqslant \sum_{t=1}^{T}(y_t - \mathbf{v} \cdot \widetilde{\mathbf{x}}_t)^2 + \lambda T B^2 + r_T B^2 \ln\!\left(1 + \frac{1}{\lambda}\right). \tag{26}$$

To conclude the proof, its only remains to show that

$$\inf_{\mathbf{v}\in\mathbb{R}^{r_T}} \sum_{t=1}^{T}(y_t - \mathbf{v} \cdot \widetilde{\mathbf{x}}_t)^2 = \inf_{\mathbf{u}\in\mathbb{R}^d} \sum_{t=1}^{T}(y_t - \mathbf{u} \cdot \mathbf{x}_t)^2\,. \tag{27}$$

Now, a basic argument of linear algebra, recalled in (33) of Section 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 (27) 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 10

**Proof** We successively prove the following two inequalities,

$$\mathcal{R}_T(\mathbf{u}) \leqslant \sum_{t=1}^{T} y_t^2 \mathbf{x}_t^\intercal \mathbf{G}_t^\dagger \mathbf{x}_t \leqslant B^2 \sum_{k=1}^{r_T} \ln\bigl(\lambda_k(\mathbf{G}_T)\bigr) + B^2 \sum_{t \in [\![1,T]\!] \cap \mathcal{T}} \ln\left(\frac{1}{\lambda_{r_t}(\mathbf{G}_t)}\right) + r_T B^2 \tag{28}$$

*Proof of the first inequality in* (28). We start by exactly rewriting the first inequality of (20):

$$\sum_{t=1}^{T} \bigl(y_t - \mathbf{x}_t^\intercal (\lambda \mathbf{I}_d + \mathbf{G}_t)^{-1} \mathbf{b}_{t-1}\bigr)^2 - \sum_{t=1}^{T} (y_t - \mathbf{u} \cdot \mathbf{x}_t)^2 \leqslant \sum_{t=1}^{T} y_t^2 \, \mathbf{x}_t^\intercal (\lambda \mathbf{I}_d + \mathbf{G}_t)^{-1} \mathbf{x}_t + \lambda \left\|\mathbf{u}\right\|^2 . \tag{29}$$

Since

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

we note that $\mathbf{x}_t^\intercal (\lambda \mathbf{I}_d + \mathbf{G}_t)^{-1}$ is the last line of the matrix $\mathbf{X}_t^\intercal \bigl(\lambda \mathbf{I}_d + \mathbf{X}_t \mathbf{X}_t^\intercal\bigr)^{-1}$, which tends to $\mathbf{X}^\dagger$ when $\lambda \to 0$ as indicated by Corollary 16 (d). Now, $\mathbf{X}^\dagger = \mathbf{X}_t^\intercal \bigl(\mathbf{X}_t \mathbf{X}_t^\intercal\bigr)^\dagger = \mathbf{X}_t^\intercal \mathbf{G}_t^\dagger$ by Corollary 16 (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,

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

is obtained by taking the limit $\lambda \to 0$ in (29).

*Proof of the second inequality in* (28). Because $y_t^2 \leqslant B^2$, we only need to prove

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

Now, Lemma 13 below shows that

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

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 $\mathbf{G}_t$.

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 (35), 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 (28). ■

**Lemma 13 (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*

$$\mathbf{x}^\mathsf{T}\mathbf{A}^\dagger\mathbf{x} = 1 - \prod_{k=1}^{r} \frac{\lambda_k(\mathbf{B})}{\lambda_k(\mathbf{A})}. \tag{30}$$

**Proof** This lemma is a consequence of the less general Lemma 12. 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 12 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 15 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

$$\sup_{0 \neq \mathbf{u} \in \mathbb{R}^r} \frac{\mathbf{u}^\mathsf{T}\mathbf{\Gamma}\mathbf{u}}{\mathbf{u}^\mathsf{T}\mathbf{u}} = \sup_{0 \neq \mathbf{u} \in \mathbb{R}^r} \frac{(\mathbf{U}\mathbf{u})^\mathsf{T}\mathbf{B}(\mathbf{U}\mathbf{u})}{(\mathbf{U}\mathbf{u})^\mathsf{T}\mathbf{U}\mathbf{u}}. \tag{31}$$

Next we show that

$$\sup_{0\neq\mathbf{u}\in\mathbb{R}^r} \frac{(\mathbf{Uu})^{\mathrm{T}}\mathbf{B}(\mathbf{Uu})}{(\mathbf{Uu})^{\mathrm{T}}\mathbf{Uu}} = \sup_{0\neq\mathbf{v}\in\mathbb{R}^d} \frac{\mathbf{v}^{\mathrm{T}}\mathbf{B}(\mathbf{v})}{\mathbf{v}^{\mathrm{T}}\mathbf{v}}\,, \tag{32}$$

which indicates, together with (31) and the characterization (36) 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 (32), 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 (32) 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 $\mathbf{A} = \mathbf{U\Sigma U}^{\mathrm{T}}$ 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 $\mathbf{y} \in \mathbb{R}^d$,

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

where we used $\mathbf{A} = \mathbf{B} + \mathbf{xx}^{\mathrm{T}}$ 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 $t \geqslant 1$,

$$\mathrm{Im}(\mathbf{G}_t) = \mathrm{Im}(\mathbf{X}_t)\,. \tag{33}$$

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

$$\mathrm{Im}(\mathbf{G}_{t-1}) \subseteq \mathrm{Im}(\mathbf{G}_t)\,. \tag{34}$$

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 (33) 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 $\mathbf{u}^{\mathrm{T}}\mathbf{G}_t\mathbf{u} = \|\mathbf{X}_t^{\mathrm{T}}\mathbf{u}\|^2$, which yields the opposite inclusion $\ker\big(\mathbf{G}_t\big) \subseteq \ker\big(\mathbf{X}_t^{\mathrm{T}}\big)$.

The inclusion (34) follows from (33) 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 $t \geqslant 1$,

$$\lambda_k\big(\mathbf{G}_{t-1}\big) \leqslant \lambda_k\big(\mathbf{G}_t\big)\,, \tag{35}$$

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 $\mathbf{M}$,

$$\lambda_k(\mathbf{M}) = \max\left\{ \min_{\mathbf{u}} \left\{ \frac{\mathbf{u}^{\mathrm{T}}\mathbf{M}\mathbf{u}}{\mathbf{u}^{\mathrm{T}}\mathbf{u}} \mid \mathbf{u} \in U \text{ and } \mathbf{u} \neq 0 \right\} \,\middle|\, U \text{ vector space with } \dim(U) = k \right\} \tag{36}$$

### E.3. Moore-Penrose inverses: 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 14 (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*

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

*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 15** *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)* | $\left(\mathbf{M}\mathbf{M}^\dagger\right)^{\mathrm{T}} = \mathbf{M}\mathbf{M}^\dagger$ |
| *(P2)* | $\mathbf{M}^\dagger\mathbf{M}\mathbf{M}^\dagger = \mathbf{M}^\dagger$ | *(P4)* | $\left(\mathbf{M}^\dagger\mathbf{M}\right)^{\mathrm{T}} = \mathbf{M}^\dagger\mathbf{M}$ |

The proof can be found in Penrose (1955). In particular, in our analysis we use the following consequences of Proposition 15. (We leave the standard proofs to the reader.)

**Corollary 16** *Let $\mathbf{M}$ be a $m \times n$ matrix and $\mathbf{N}$ a $n \times p$ matrix. Then,*

*(a)* $\mathbf{M}^\dagger = \mathbf{M}^{\mathrm{T}}\left(\mathbf{M}\mathbf{M}^{\mathrm{T}}\right)^\dagger$;

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

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

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

*(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.*