Proceedings of Machine Learning Research vol 195:1–19, 2023 36th Annual Conference on Learning Theory

A Lower Bound for Linear and Kernel Regression with Adaptive Covariates

Tor Lattimore LATTIMORE@DEEPMIND.COM
Google DeepMind, London

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We prove that the continuous time version of the concentration bounds by Abbasi-Yadkori et al. (2011) for adaptive linear regression cannot be improved in general, showing that there can be a significant price for sequential design. This resolves the continuous time version of the COLT open problem by Vakili et al. (2021b) on confidence intervals for kernel regression under sequential design. Experimental evidence suggests that improved confidence bounds are also not possible in discrete time.

1. Introduction

A statistician observes a stochastic process of covariates $(x_t)_{t \in [0,n]}$ in $\mathbb{R}^{d+1}$ and a process of cumulative responses, which is a stochastic process $(y_t)_{t \in [0,n]}$ satisfying the stochastic differential equation

$$\mathrm{d}y_t = \langle x_t, \theta_\star \rangle \, \mathrm{d}t + \mathrm{d}B_t \,, \tag{1}$$

where $(B_t)_t$ is a standard Brownian and $\theta_\star \in \mathbb{R}^{d+1}$ is an unobserved random element. We assume that (x_t) is adapted to the filtration $(\mathscr{G}_t)$ with $\mathscr{G}_t = \sigma((x_s, y_s)_{s \in [0,t)})$ and that $\theta_\star$ is independent of (B_t) and has law ξ for some probability measure ξ on $\mathbb{R}^{d+1}$. The statistician's job is to observe the (x_t, y_t) process on the interval $[0, n]$ and output an estimate $\mathcal{E}$ of $\langle v, \theta_\star \rangle$ where $v \in \mathbb{R}^{d+1}$ is a prespecified nonzero vector. Mathematically, $\mathcal{E}$ is any $\mathscr{G}_n$-measurable random variable. Note that the prior measure ξ is known.

Performance measure To manage expectations, let us remind ourselves about what happens when the covariates are chosen deterministically. Given a positive definite matrix A, let $\|x\|_A = \sqrt{x^\top A x}$ and $\|\cdot\|$ be the standard euclidean norm. Suppose that the covariate process (x_t) is deterministic and let $D_n = \lambda \mathbb{1} + \int_0^n x_t x_t^\top \, \mathrm{d}t$ be the regularised design matrix. When $\lambda = 0$ and D_n is invertible, then the least squares estimator of $\theta_\star$ is $\hat{\theta}_n = D_n^{-1} \int_0^n x_t \, \mathrm{d}y_t$. Since D_n is deterministic, the law of $\theta_\star$ is Gaussian with mean θ and covariance D_n^{-1}. Hence, by standard concentration bounds for Gaussian random variables (Boucheron et al., 2013, p22), for any $v \in \mathbb{R}^{d+1}$ with probability at least $1 - \delta$,

$$|\langle \hat{\theta}_n - \theta_\star, v \rangle| \leq \|v\|_{D_n^{-1}} \sqrt{2 \log(2/\delta)} \,.$$

On the other hand, when the covariate process is chosen adaptively, the bound by Abbasi-Yadkori et al. (2011) shows that for any $\lambda > 0$, with probability at least $1 - \delta$,

$$|\langle \hat{\theta}_n - \theta_\star, v \rangle| \leq \|v\|_{D_n^{-1}} \left[\sqrt{\lambda} \|\theta_\star\| + \sqrt{2 \log(1/\delta) + \log \det \left(D_0^{-1} D_n \right)} \right] \,, \tag{2}$$

where $\hat{\theta}_n = D_n^{-1} \int_0^n x_t \, \mathrm{d}y_t$ is the ridge regression estimator. These bounds are typically stated in discrete time, but the martingale arguments generalise without change to continuous processes driven by Brownian motion as we consider here. Our contribution is to show that the bound above is not improvable for a reasonably large class of problems with adaptively chosen covariates. In fact, we prove that in certain circumstances no estimator $\mathcal{E}$ of $\langle \theta_\star, v \rangle$ improves on the bound above. To measure the price of adaptivity we compare our statisticians' error relative to the above benchmark by defining the risk of an estimator $\mathcal{E}$ by

$$\mathfrak{R}_n(\mathcal{E}) = \mathbb{E}\left[\frac{(\mathcal{E} - \langle \theta_\star, v \rangle)^2}{\|v\|^2_{D_n^{-1}}}\right].$$

The notation of the left-hand side hides the dependence on λ and the direction v, as well as the prior and method of generating the covariates. Integrating the high probability bounds shows that when the covariate process is deterministic, $\mathfrak{R}_n(\mathcal{E}) = O(1)$, while in the adaptive setting, integrating Equation (2) only yields

$$\mathfrak{R}_n(\mathcal{E}) = O(\mathbb{E}[\lambda \|\theta_\star\|^2 + \log \det(D_0^{-1} D_n)]) . \tag{3}$$

Our main construction shows that this worse bound cannot be improved in general, including in the kernel regime where the log-determinant is large, which resolves a continuous-time analogue of the open problem posed by Vakili et al. (2021b).

A note on the set-up We have used continuous time because it leads to a more elegant analysis without (we believe) much loss in insight. As solace, the experiments use discrete time.

Motivation and related work Linear regression with adaptively chosen covariates arises naturally when the experimenter wishes to adapt the design online using previously collected data. A particular application where this is essential is in linear bandits, where the covariates need to be chosen adaptively to minimise the regret. The simplest algorithms for linear bandits are based on a combination of adaptive confidence intervals for the ridge regression estimator and the optimism principle (Abbasi-Yadkori et al., 2011). These results are known to be suboptimal when the number of actions is much less than exponential in the dimension. In this case more sophisticated algorithms are needed, which introduce various gadgets to obtain the kind of independence needed for tighter confidence bounds using classical methods (Auer, 2002; Valko et al., 2013; Li et al., 2019). Unfortunately these modifications of the basic principle generally lead to worse performance empirically. The example in the present work shows that any analysis of linear contextual bandits aimed at proving a similar result cannot completely decouple the concentration analysis and the algorithm. The same is true for kernelised bandits where the dimension-dependence arising from loose confidence bounds is especially pernicious and can be the difference between sublinear and linear regret (Vakili et al., 2021a,b).

The new lower bound also shows that the self-normalised concentration bound by Abbasi-Yadkori et al. (2011) has very little room for improvement. As a side effect, it shows that a law-of-the-iterated logarithm result is not possible except in dimension one. Regrettably this contradicts Theorem 8 by Lattimore and Szepesvári (2017), which has a mistake in the covering argument. The main theorem in that work is still correct with some adjustments, or by using a different analysis and algorithm (Degenne et al., 2020, for example). The same problem appears in the concentration result of Tirinzoni et al. (2020).

Of course there is a large literature on linear regression with sequential covariates, with early work by Lai and Wei (1982). A more recent work, which also gives a nice survey of the literature, is by Khamaru et al. (2021). At a very high level, these works prove positive results (consistency/asymptotic normality) when the process of covariates gives sufficient coverage, generally expressed in terms of a condition on the spectrum of the design matrix. In some cases the assumptions are such that the least squares estimator gets the job done, while in others some modification is needed (Zhang et al., 2021; Khamaru et al., 2021, for example). The continuous time model was also considered by Liptser and Spokoiny (2000), who prove finite-time positive results under certain conditioning assumptions on the covariates. Interestingly, the strongest positive results are just on the boundary of what is necessary to obtain logarithmic regret for bandits with static action sets, while for contextual bandits we do not see how they can be applied without making additional assumptions. There also exist some negative results. Most relevant is the construction in Exercise 20.2 of the book by Lattimore and Szepesvári (2020), which shows that when $\mathcal{E}$ is the inner product between the least squares estimate and v, then there exist instances for which $\mathfrak{R}_n(\mathcal{E}) = \Omega(d)$ with $n = \Theta(d)$.

Notation The indicator function of a set A is $\mathbf{1}_A$. We use $\mathbf{0}$ to denote the vector that is zero in all coordinates and the identity matrix is $\mathbb{1}$. In all cases we hope the reader can deduce the dimension of the relevant quantity from the context. The Gaussian distribution with mean μ and covariance Σ is $\mathcal{N}(\mu, \Sigma)$. Given positive definite matrices A and B, let $\operatorname{Breg}(A, B) = \log\det(A^{-1}B) + \operatorname{tr}(B^{-1}(A - B))$, which is the Bregman divergence with respect to the negative log determinant. We will use this notation and definition even when A and B are infinite positive definite matrices (i.e., positive self-adjoint operators). The determinant in this case is defined as the product of the eigenvalues and will only ever be applied to operators of the form $A = \mathbb{1} + T$ where T is positive, self adjoint and trace class, which ensures that the determinant as the product of the eigenvalues is well defined. When $D_n - \lambda\mathbb{1}$ has eigenvalues $(\lambda_m)_{m=1}^d$, then

$$\operatorname{Breg}(D_0, D_n) = \sum_{m=1}^{d} \left(\log\left(\frac{\lambda + \lambda_m}{\lambda}\right) + \frac{\lambda}{\lambda + \lambda_m} - 1 \right) \leq \log\det(D_0^{-1} D_n)\,.$$

The inequality is close to an equality when the log determinant is dominated by the contribution of eigenvalues λ_m with $\lambda_m \gg \lambda$.

2. Construction

Given any deterministic covariate process in $\mathbb{R}^d$, we show that a simple adaptive lifting to $\mathbb{R}^{d+1}$ dramatically increases the difficulty of estimation relative to the unlifted deterministic set-up. Note that we allow $d = \infty$. Let $\xi = \frac{1}{2}\delta_{\mathbf{0}} + \frac{1}{2}\mathcal{N}(\mu, \Sigma)$, where $\delta_{\mathbf{0}}$ is a Dirac on $\mathbf{0} \in \mathbb{R}^{d+1}$ and

$$\mu = (1, 0, \dots, 0) \in \mathbb{R}^{d+1} \qquad\qquad \Sigma = \begin{bmatrix} 0 & \mathbf{0}^\top \\ \mathbf{0} & \frac{1}{\lambda}\mathbb{1} \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}\,.$$

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space carrying a random vector $\theta_\star : \Omega \to \mathbb{R}^d$ with law ξ and let (B_t) be an independent Brownian motion adapted to a filtration $(\mathscr{F}_t)_{t \geq 0}$ satisfying the usual conditions. Let $v = (1, 0, \dots, 0)$, which means that $\langle v, \theta_\star \rangle \in \{0, 1\}$. Let $(b_t)_{t\in[0,n]}$ be a deterministic function in $\mathbb{R}^d$ and $H_t = \lambda\mathbb{1} + \int_0^t b_s b_s^\top \, \mathrm{d}s \in \mathbb{R}^{d\times d}$, which we assume exists and satisfies $\operatorname{tr}(H_n - \lambda\mathbb{1}) < \infty$.

In a moment we construct a stochastic process $(a_t)_{t\in[0,n]}$ and let $x_t = (a_t, b_t) \in \mathbb{R}^{d+1}$ be the lifted covariate process and y_t be the response process defined by Equation (1). Note that in the lifted process only the first coordinate is adaptive and is used to maximise the confusion of the statistician about the inner product $\langle v, \theta_\star\rangle$. The dynamics governing the process (a_t) are given by the stochastic differential equation

$$S_0 = \mathbf{0} \qquad \mathrm{d}S_t = b_t\langle x_t, \theta_\star\rangle\,\mathrm{d}t - a_t b_t\,\mathrm{d}t + b_t\,\mathrm{d}B_t \qquad a_t = -\langle \hat\varphi_t, b_t\rangle \qquad \hat\varphi_t = H_t^{-1} S_t\,.$$

Existence and uniqueness of a strong solution is guaranteed by the classical results of Itô (Karatzas and Shreve, 2012, Theorems 5.2.5 and 5.2.9). Note that $\mathrm{d}S_t = b_t\,\mathrm{d}y_t - a_t b_t\,\mathrm{d}t$, which means that (x_t) is adapted to the filtration $(\mathscr{G}_t)$ as we assumed it must be. To build a little intuition, let φ be the last d coordinates of $\theta_\star$ (all but the first when $d = \infty$). Then $\hat\varphi_t$ is the ridge regression estimator of φ on the event $\{\theta_\star \neq \mathbf{0}\}$. The process a_t is chosen adaptively so that $\langle x_t, (1, \hat\varphi_t)\rangle = 0 = \langle x_t, \mathbf{0}\rangle$, which means the statistician has trouble gaining information about the event $\{\theta_\star \neq \mathbf{0}\}$. Actually we will show that the statistician cannot gain *any* information about $\{\theta_\star \neq \mathbf{0}\}$. And since $\langle v, \theta_\star\rangle = \mathbf{1}_{\{\theta_\star \neq \mathbf{0}\}}$, there is no better estimate of $\langle v, \theta_\star\rangle$ than $\mathcal{E} = 1/2$. The proof of our main theorem below boils down to making the above intuition rigorous. What is nice about this construction and the use of continuous time is that the *exact* optimal risk can be computed analytically.

Theorem 1 *The optimal risk in the instance constructed above satisfies*

$$\mathfrak{R}_n^\star \triangleq \inf_{\mathcal{E}} \mathfrak{R}_n(\mathcal{E}) = \frac{\lambda + \mathrm{Breg}(H_0, H_n)}{4}\,,$$

where the infimum is over all random variables $\mathcal{E} : \Omega \to \mathbb{R}$ measurable with respect to the σ-algebra generated by the processes $(x_t)_{0\le t\le n}$ and $(y_t)_{0\le t\le n}$.

Before the proof, let us spend a moment to see under what conditions the lower bound in Theorem 1 matches the upper bound obtained in Equation (3). The following lemma relates the spectrum of the unlifted design matrix H_n and the lifted one D_n. Note that H_n is deterministic while D_n is a random variable.

Lemma 2 $\mathbb{E}[\log\det D_0^{-1} D_n] \ge \log\det H_0^{-1} H_n \ge \left(\frac{\lambda}{\lambda+1}\right) \mathbb{E}[\log\det D_0^{-1} D_n]$.

The proof can be found in Appendix B. Suppose that the following hold:
(a) $\lambda = \Theta(1)$; and
(b) $\mathrm{Breg}(H_0, H_n) = \Omega(\log\det H_0^{-1} H_n)$; and
(c) $\log\det H_0^{-1} H_n = \Omega(d)$.
By Lemma 2 and Theorem 1, the risk in the construction above is lower bounded by

$$\begin{aligned}
\mathfrak{R}_n^\star = \frac{\lambda + \mathrm{Breg}(H_0, H_n)}{4} &= \Omega\left(\frac{\lambda + \log\det(H_0^{-1} H_n)}{4}\right) && \text{by } (b) \\
&= \Omega\left(\frac{\lambda + \frac{\lambda}{\lambda+1}\mathbb{E}[\log\det D_0^{-1} D_n]}{4}\right) && \text{by Lemma 2} \\
&= \Omega(\mathbb{E}[\log\det D_0^{-1} D_n])\,. && \text{by } (a)
\end{aligned}$$

On the other hand, in the construction above we have $\mathbb{E}[\|\theta_\star\|^2] = O(d)$ and so the bound obtained by integrating the self-normalised inequality in Equation (3) gives $\mathfrak{R}_n^\star = O(d + \mathbb{E}[\log\det D_0^{-1} D_n])$. By Lemma 2 and *(c)*, $\mathbb{E}[\log\det D_0^{-1} D_n] \geq \log\det H_0^{-1} H_n = \Omega(d)$. Hence, under conditions *(a)*, *(b)* and *(c)* above, the lower and upper bounds match up to constant factors. The Bregman divergence is upper bounded by the log determinant, so you should be skeptical about when *(b)* might hold. Let $(\lambda_m)_{m=1}^d$ be the eigenvalues of $\int_0^n b_t b_t^\top \, \mathrm{d}t$. Then

$$\log\det(H_0^{-1} H_n) = \sum_{m=1}^d \log\left(1 + \frac{\lambda_m}{\lambda}\right) , \quad \mathrm{Breg}(H_0, H_n) = \sum_{m=1}^d \log\left(1 + \frac{\lambda_m}{\lambda}\right) - \frac{\lambda_m}{\lambda + \lambda_m} .$$

In the low-dimensional case where n is large relative to d and the covariate process $(b_t)_{t\in[0,n]}$ is well conditioned, then $\lambda_m = \Omega(n)$ and the Bregman divergence and log determinant have the same order. On the other hand, in the high-dimensional 'kernel' regime the situation is more nuanced. A Taylor expansion shows that

$$\log\left(1 + \frac{\lambda_m}{\lambda}\right) = \frac{\lambda_m}{\lambda} + o\left(\frac{\lambda_m}{\lambda}\right) , \qquad \log\left(1 + \frac{\lambda_m}{\lambda}\right) - \frac{\lambda_m}{\lambda + \lambda_m} = \frac{\lambda_m^2}{2\lambda^2} + o\left(\frac{\lambda_m^2}{\lambda^2}\right) .$$

This shows that if the log determinant is dominated by eigenvalues with $\lambda_m \ll \lambda$, then the Bregman divergence may be very small relative to the log determinant, while otherwise the two will have the same order of magnitude.

Low-dimensional application Consider the case where $\lambda = 1$ and $(b_t)_{t\in[0,n]}$ are chosen so that each of the d standard basis vectors appears in equal proportion. Then $H_n = (1 + \frac{n}{d})\mathbb{1}$ and our results show that

$$\mathfrak{R}_n^\star = \frac{1 + d\log(1 + n/d) - d + \frac{d^2}{d+n}}{4} = \Omega\left(d \log\left(\frac{n}{d}\right)\right) .$$

This contradicts the law-of-the-iterated logarithm style confidence bounds for least-squares estimators claimed by Lattimore and Szepesvári (2017) and Tirinzoni et al. (2020), both of which contain errors in their covering arguments (details in Appendix E). Note that the covariates in our construction always live in a space of at least dimension two, which is why there is no contradiction with the standard law-of-the-iterated logarithm.

Comments on infinite-dimensional case When $d = \infty$, then *(c)* above does not hold and in this case Equation (3) is vacuous because $\mathbb{E}[\|\theta_\star\|^2] = \infty$ in our construction. The analysis is useful nevertheless for analysing the behaviour of the ridge regression estimator in the kernel setting, as we shall see in Sections 4 and 5, where we resolve the open problem of Vakili et al. (2021b). You can also still use the analysis to bound the risk of any estimator in the kernel setting by applying Theorem 1 with a finite dimensional approximation of the kernel and control the approximation error in some other way, as done, for example, by Vakili et al. (2021a). This typically works when d is chosen to have the same order as the effective dimension.

3. Proof of Theorem 1

There are three steps. First we show that if the posterior distribution of $\langle v, \theta_\star \rangle$ is almost surely constant, then

$$\mathfrak{R}_n^\star = \frac{1}{4}\mathbb{E}\left[\frac{1}{\|v\|^2_{D_n^{-1}}}\right] .$$

In the second step we show that indeed the posterior distribution of $\langle v, \theta_\star \rangle$ is almost surely constant. In the third step we complete the proof by evaluating the expectation in the above display.

Step 1: Bayesian optimal estimator and risk Let $\mathscr{G}_t$ be the σ-algebra generated by the processes $(x_s)_{0 \leq s \leq t}$ and $(y_s)_{0 \leq s \leq t}$. The Bayesian risk of estimator $\mathscr{E}$ is

$$\mathfrak{R}_n(\mathscr{E}) = \mathbb{E}\left[\frac{(\mathscr{E} - \langle v, \theta_\star \rangle)^2}{\|v\|^2_{D_n^{-1}}}\right] = \mathbb{E}\left[\frac{\mathbb{E}\left[(\mathscr{E} - \langle v, \theta_\star \rangle)^2 | \mathscr{G}_n\right]}{\|v\|^2_{D_n^{-1}}}\right] .$$

which is minimised by $\mathscr{E} = \mathbb{E}[\langle v, \theta_\star \rangle | \mathscr{G}_n]$. By construction of the prior ξ, $\langle v, \theta_\star \rangle \in \{0, 1\}$ and hence $\mathscr{E} = \mathbb{P}(\langle v, \theta_\star \rangle = 1 | \mathscr{G}_n) = \mathbb{P}(\theta_\star \neq \mathbf{0} | \mathscr{G}_n)$. Therefore, if $\mathbb{P}(\theta_\star \neq \mathbf{0} | \mathscr{G}_n) = \mathbb{P}(\theta_\star = \mathbf{0} | \mathscr{G}_n) = \frac{1}{2}$ almost surely, then $\mathscr{E} = \frac{1}{2}$ almost surely and

$$\mathfrak{R}_n^\star = \mathbb{E}\left[\frac{\mathbb{E}[\langle v, \theta_\star \rangle^2 | \mathscr{G}_n] - \mathbb{E}[\langle v, \theta_\star \rangle | \mathscr{G}_n]^2}{\|v\|^2_{D_n^{-1}}}\right] = \frac{1}{4}\mathbb{E}\left[\frac{1}{\|v\|^2_{D_n^{-1}}}\right] . \tag{4}$$

Step 2: Evolution of posterior Let $\nu = \mathcal{N}(\mathbf{0}, \frac{1}{\lambda}\mathbb{1})$ be a Gaussian in $\mathbb{R}^d$. By Girsanov's theorem (Karatzas and Shreve, 2012, Theorem 3.5.1) and a standard integral (complete the square in φ),

$$\begin{aligned} \frac{\mathbb{P}(\theta^\star \neq \mathbf{0} | \mathscr{G}_t)}{\mathbb{P}(\theta^\star = \mathbf{0} | \mathscr{G}_t)} &= \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\int_0^t (a_s + \langle \varphi, b_s \rangle)^2 \,\mathrm{d}s + \int_0^t (a_s + \langle \varphi, b_s \rangle) \,\mathrm{d}y_s\right) \mathrm{d}\nu(\varphi) \\ &= \exp\left(\frac{1}{2}\|S_t\|^2_{H_t^{-1}} - \frac{1}{2}\int_0^t a_s^2 \,\mathrm{d}s + \int_0^t a_s \,\mathrm{d}y_s - \frac{1}{2}\log\det(H_0^{-1}H_t)\right) \\ &\triangleq \exp(M_t) , \end{aligned}$$

where the last equality serves as the definition of M_t. We will now show that $\mathrm{d}M_t = 0$, which implies that the posterior belief that $\theta_\star = \mathbf{0}$ does not change over time (almost surely). Note that we are not arguing that the statistician gains no information from the data. Only that they do not gain information about the event $\{\theta_\star = \mathbf{0}\}$. The calculation above holds also in discrete time, but in what follows we make full use of the continuous time model. In discrete time the lower-order terms in the discrete derivative would need to be handled, leading to a messier computation. Taking the derivative of M_t,

$$\mathrm{d}M_t = \mathrm{d}\frac{1}{2}\|S_t\|^2_{H_t^{-1}} - \frac{1}{2}a_t^2 \,\mathrm{d}t + a_t \,\mathrm{d}y_t - \mathrm{d}\frac{1}{2}\log\det H_0^{-1}H_t . \tag{5}$$

Recall that $\mathrm{d}S_t = b_t(\mathrm{d}y_t - a_t\,\mathrm{d}t)$ and $\mathrm{d}H_t = b_t b_t^\top\,\mathrm{d}t$ and so $\mathrm{d}H_t^{-1} = -H_t^{-1} b_t b_t^\top H_t^{-1}\,\mathrm{d}t$ and by Itô's formula (Karatzas and Shreve, 2012, Theorem 3.3.6),

$$\begin{aligned}
\mathrm{d}\frac{1}{2}\|S_t\|^2_{H_t^{-1}} &= \langle H_t^{-1}S_t, \mathrm{d}S_t\rangle - \frac{1}{2}\|S_t\|^2_{H_t^{-1}b_tb_t^\top H_t^{-1}}\,\mathrm{d}t + \frac{1}{2}\|b_t\|^2_{H_t^{-1}}\,\mathrm{d}t \\
&= \langle\hat\varphi_t, \mathrm{d}S_t\rangle - \frac{1}{2}\langle\hat\varphi_t, b_t\rangle^2\,\mathrm{d}t + \frac{1}{2}\|b_t\|^2_{H_t^{-1}}\,\mathrm{d}t \\
&= \langle\hat\varphi_t, b_t\rangle\,\mathrm{d}y_t - a_t\langle\hat\varphi_t, b_t\rangle\,\mathrm{d}t - \frac{1}{2}\langle\hat\varphi_t, b_t\rangle^2\,\mathrm{d}t + \frac{1}{2}\|b_t\|^2_{H_t^{-1}}\,\mathrm{d}t \\
&= -a_t\,\mathrm{d}y_t + \frac{1}{2}a_t^2\,\mathrm{d}t + \frac{1}{2}\|b_t\|^2_{H_t^{-1}}\,\mathrm{d}t\,,
\end{aligned} \tag{6}$$

where in the first equality we used Itô's formula (the last term arises from the second derivative). The second equality is because $\hat\varphi_t = H_t^{-1}S_t$ and the third by substituting the definition of $\mathrm{d}S_t$ and finally using that $\langle\hat\varphi_t, b\rangle = -a_t$. Furthermore,

$$\mathrm{d}\frac{1}{2}\log\det H_t = \frac{1}{2}\|b_t\|^2_{H_t^{-1}} \tag{7}$$

Combining Equations (5) to (7) shows that $\mathrm{d}M_t = 0$, which means the posterior distribution about whether or not $\theta_\star = \mathbf{0}$ does not change: $\mathbb{P}(\theta_\star = \mathbf{0}|\mathscr{G}_t) = 1/2$ *a.s.* for all $0 \leq t \leq n$. Let $E_\mathbf{0} = \{\theta_\star = \mathbf{0}\}$. By the above argument, $\mathbb{P}(E_\mathbf{0}|\mathscr{G}_t) = 1/2$ for all t almost surely. By Bayes' law it follows that the law of the process (x_t, y_t) is the same under all of $\mathbb{P}$ and $\mathbb{P}(\cdot|E_\mathbf{0})$ and $\mathbb{P}(\cdot|E_\mathbf{0}^c)$.

Step 3: Law of design matrix Next we need to consider the law of $\|v\|^2_{D_n^{-1}}$. By construction,

$$D_n = \begin{bmatrix} \lambda + \int_0^n \langle b_t, \hat\varphi_t\rangle^2\,\mathrm{d}t & -\int_0^n \langle b_t, \hat\varphi_t\rangle b_t^\top\,\mathrm{d}t \\ -\int_0^n \langle b_t, \hat\varphi_t\rangle b_t\,\mathrm{d}t & \lambda\mathbb{1} + \int_0^n b_tb_t^\top\,\mathrm{d}t \end{bmatrix}. \tag{8}$$

Letting $G_t = \int_0^t b_sb_s^\top\,\mathrm{d}s$ and $U_t = \int_0^t b_sb_s^\top H_s^{-1}G_s\,\mathrm{d}s$ and $L_t = \int_0^s b_sb_s^\top H_s^{-1}\,\mathrm{d}s$,

$$H_t - U_t = H_0 + \int_0^t b_sb_s^\top(\mathbb{1} - H_s^{-1}G_s)\,\mathrm{d}s = H_0 + \int_0^t b_sb_s^\top H_s^{-1}H_0\,\mathrm{d}t = H_0 + H_0L_t\,. \tag{9}$$

Using the definitions of $\hat\varphi_t$ and Fact 7 in Appendix A,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\|v\|^2_{D_n^{-1}}}\right] &= \mathbb{E}\left[\frac{1}{(D_n^{-1})_{1,1}}\middle| E_\mathbf{0}^c\right] \\
&= \lambda + \left[\underbrace{\mathbb{E}\left[\int_0^n \langle b_t, \hat\varphi_t\rangle^2\,\mathrm{d}t\middle| E_\mathbf{0}^c\right]}_{\text{(A)}} - \underbrace{\mathbb{E}\left[\left\|\int_0^n \langle b_t, \hat\varphi_t\rangle b_t\,\mathrm{d}t\right\|^2_{H_n^{-1}}\middle| E_\mathbf{0}^c\right]}_{\text{(B)}}\right]
\end{aligned} \tag{10}$$

By the standard posterior for Bayesian linear regression, $\mathbb{P}(\varphi = \cdot | E_{\mathbf{0}}^c, \mathscr{G}_t) = \mathcal{N}(\hat{\varphi}_t, H_t)$. Therefore,

$$
\begin{aligned}
\text{(A)} &= \mathbb{E}\left[\int_0^n \langle b_t, \hat{\varphi}_t\rangle^2 \,\mathrm{d}t \middle| E_{\mathbf{0}}^c\right] \\
&= \mathbb{E}\left[\int_0^n \langle b_t, \hat{\varphi}_t - \varphi\rangle^2 \,\mathrm{d}t - \int_0^n \langle b_t, \varphi\rangle^2 \,\mathrm{d}t + 2\int_0^n \langle b_t, \varphi\rangle\langle b_t, \hat{\varphi}_t\rangle \middle| E_{\mathbf{0}}^c\right] \\
&= \int_0^n \|b_t\|^2_{H_t^{-1}} \,\mathrm{d}t - \frac{\operatorname{tr}(G_n)}{\lambda} + \frac{2}{\lambda}\int_0^n \operatorname{tr}\left(b_t b_t^\top H_t^{-1} G_t\right) \mathrm{d}t \\
&= \log\left(\det(H_0^{-1} H_n)\right) - \frac{\operatorname{tr}(G_n)}{\lambda} + \frac{2\operatorname{tr}(U_n)}{\lambda}\,,
\end{aligned}
$$

where the third equality follows from the covariance of φ given $\mathscr{G}_t$ and the definition of $\hat{\varphi}_t$. The last inequality follow from the definition of U_n. For the second term in Equation (10), expanding the square, the Itô isometry and integrating by parts yields

$$
\begin{aligned}
\text{(B)} &= \mathbb{E}\left[\left\|\int_0^n \langle b_t, \hat{\varphi}_t\rangle b_t \,\mathrm{d}t\right\|^2_{H_n^{-1}} \middle| E_{\mathbf{0}}^c\right] \\
&= \mathbb{E}\left[\left\|\int_0^n b_t\langle b_t, H_t^{-1} G_t \varphi\rangle \,\mathrm{d}t + \int_0^n \int_0^t b_t \langle b_t, H_t^{-1} b_s\rangle \,\mathrm{d}B_s \,\mathrm{d}t\right\|^2_{H_n^{-1}} \middle| E_{\mathbf{0}}^c\right] \\
&= \frac{1}{\lambda}\operatorname{tr}\left(U_n H_n^{-1} U_n\right) + \int_0^n \|(L_n - L_s) b_s\|^2_{H_n^{-1}} \,\mathrm{d}s \\
&= \frac{1}{\lambda}\operatorname{tr}\left(U_n H_n^{-1} U_n\right) - \operatorname{tr}(L_n H_n^{-1} L_n H_0) - 2\operatorname{tr}\left(L_n H_n^{-1} H_0\right) + 2\operatorname{tr}\left(H_n^{-1}(H_n - H_0)\right)\,.
\end{aligned}
$$

Combining and simplifying by substituting $H_0 = \lambda \mathbb{1}$ and Equation (9) shows that

$$
\text{(A)} - \text{(B)} = \log\det(H_0^{-1} H_n) + \operatorname{tr}(H_n^{-1}(H_0 - H_n)) = \operatorname{Breg}(H_0, H_n)\,.
$$

The result follows by substituting this in Equation (10) and then Equation (4).

4. Ridge regression estimator

Theorem 1 bounds the risk for *any* estimator. For the classical ridge regression estimator we can say a little more. Let

$$
\hat{\theta}_n = D_n^{-1} \int_0^n x_t \,\mathrm{d}y_t\,. \tag{11}
$$

Let $\mathbb{P}_{\mathbf{0}}$ be the law of $(x_t, y_t)_{t \in [0,n]}$ under $\mathbb{P}(\cdot | E_{\mathbf{0}})$ and let $\mathbb{E}_{\mathbf{0}}$ be the corresponding expectation operator. By the calculation in Step 2 in the proof of Theorem 1, the law of $(x_t, y_t)_{t\in[0,n]}$ under the unconditioned measure $\mathbb{P}$ is in fact equal to $\mathbb{P}_{\mathbf{0}}$.

Theorem 3 *With the same set-up as in Theorem 1,*

$$
\mathbb{E}_{\mathbf{0}}\left[\frac{\langle \hat{\theta}_n - \theta_\star, v\rangle^2}{\|v\|^2_{D_n^{-1}}}\right] \geq \frac{\operatorname{Breg}(H_n, H_0)^2}{\lambda + \operatorname{Breg}(H_n, H_0)}\,.
$$

The bound is slightly weaker than the result of Theorem 1 because it only applies to the ridge regression estimator and the risk lower bound is slightly smaller. The advantage is that it holds for $\theta_\star = \mathbf{0}$. Note that some kind of bad dependence on λ is essential since the risk of the ridge regression estimator for $\theta_\star = \mathbf{0}$ obviously vanishes as $\lambda \to \infty$.

Proof By the definition of v,

$$\mathbb{E}_{\mathbf{0}}\left[\frac{\langle \hat\theta_n - \theta_\star, v\rangle^2}{\|v\|^2_{D_n^{-1}}}\right] = \mathbb{E}_{\mathbf{0}}\left[\frac{(\hat\theta_n)_1^2}{(D_n^{-1})_{1,1}}\right] .$$

Using the definition of D_n and the inversion formula for block matrices (Fact 7),

$$(\hat\theta_n)_1 = (D_n^{-1})_{1,1}\Bigg[\underbrace{\int_0^n a_t \,\mathrm{d}B_t - \left\langle \int_0^n a_t b_t \,\mathrm{d}t, H_n^{-1}\int_0^n b_t \,\mathrm{d}B_t\right\rangle}_{\text{(A)}}\Bigg] .$$

Therefore, by Jensen's inequality and the convexity of $(x,y) \mapsto x^2/y$ on $\mathbb{R} \times (0,\infty)$,

$$\mathbb{E}_{\mathbf{0}}\left[\frac{(\hat\theta_n)_1^2}{(D_n^{-1})_{1,1}}\right] = \mathbb{E}_{\mathbf{0}}\left[\frac{\text{(A)}^2}{1/(D_n^{-1})_{1,1}}\right] \geq \frac{\mathbb{E}_{\mathbf{0}}[\text{(A)}]^2}{\mathbb{E}_{\mathbf{0}}[1/(D_n^{-1})_{1,1}]} .$$

By the last remark in the second step of the proof of Theorem 1 and by the third step of the same,

$$\mathbb{E}_{\mathbf{0}}[1/(D_n^{-1})_{1,1}] = \mathbb{E}[1/(D_n^{-1})_{1,1}] = \lambda + \operatorname{Breg}(H_0, H_n) .$$

Evaluating $\mathbb{E}_{\mathbf{0}}[\text{(A)}]$ is another exercise in Itô calculus:

$$\begin{aligned}\mathbb{E}_{\mathbf{0}}[\text{(A)}] &= \mathbb{E}_{\mathbf{0}}\left[\int_0^n a_t \,\mathrm{d}B_t - \int_0^n a_t b_t \,\mathrm{d}t H_n^{-1}\int_0^n b_t \,\mathrm{d}B_t\right] \\ &= \mathbb{E}_{\mathbf{0}}\left[\int_0^n \langle \hat\varphi_t, b_t\rangle b_t \,\mathrm{d}t H_n^{-1}\int_0^n b_t \,\mathrm{d}B_t\right] .\end{aligned}$$

Let $S_t = \int_0^t b_t(y_t - a_t)\,\mathrm{d}t$ and recall that $a_t = -\langle b_t, \hat\varphi_t\rangle = -\langle b_t, H_t^{-1}S_t\rangle$. Therefore, since $y_t \,\mathrm{d}t = \mathrm{d}B_t$, $\mathrm{d}S_t = b_t\,\mathrm{d}B_t + b_t b_t^\top H_t^{-1} S_t \,\mathrm{d}t$, which has a unique strong solution of $S_t = H_t \int_0^t H_s^{-1} b_s \,\mathrm{d}B_s$ and hence

$$\begin{aligned}\mathbb{E}_{\mathbf{0}}[(A)] &= \mathbb{E}_{\mathbf{0}}\left[\int_0^n S_t^\top H_t^{-1} b_t b_t^\top \,\mathrm{d}t H_n^{-1}\int_0^n b_t\,\mathrm{d}B_t\right] \\ &= \mathbb{E}_{\mathbf{0}}\left[\int_0^n \int_0^t b_s^\top H_s^{-1}\,\mathrm{d}B_s b_t b_t^\top \,\mathrm{d}t H_n^{-1}\int_0^n b_t\,\mathrm{d}B_t\right] \\ &= \mathbb{E}_{\mathbf{0}}\left[\int_0^n b_s^\top H_s^{-1}\,\mathrm{d}B_s \int_s^n b_t b_t^\top \,\mathrm{d}t \,\mathrm{d}B_s H_n^{-1}\int_0^n b_t\,\mathrm{d}B_t\right] \\ &= \mathbb{E}_{\mathbf{0}}\left[\int_0^n b_s^\top H_s^{-1}(H_n - H_s)\,\mathrm{d}B_s H_n^{-1}\int_0^n b_t\,\mathrm{d}B_t\right] \\ &= \mathbb{E}_{\mathbf{0}}\left[\int_0^n b_s^\top H_s^{-1}\,\mathrm{d}B_s\int_0^n b_t\,\mathrm{d}B_t - \int_0^n b_s^\top \,\mathrm{d}B_s H_n^{-1}\int_0^n b_t\,\mathrm{d}B_t\right] \\ &= \log\det(H_0^{-1}H_n) - \operatorname{tr}((H_n - H_0)H_n^{-1}) = \operatorname{Breg}(H_0, H_n) .\end{aligned}$$

Combining the parts completes the proof. ■

The only missing ingredient in Theorem 3 is that the process (a_t) is not uniformly bounded almost surely, which we need in order to formally resolve the continuous-time version of the open problem by Vakili et al. (2021b). To prove such a bound, define a stopping time

$$\tau = \min\left\{t \in [0, n] : a_t^2 \geq \frac{2\|b_t\|^2}{\lambda}\left(8 + \frac{2}{\lambda} + 5 \log\det(H_0^{-1} H_n)\right)\right\}, \tag{12}$$

where the minimum of the empty set is defined to be n.

Theorem 4 *Under the same conditions as Theorem 3,*

$$\mathbb{E}_{\mathbf{0}}\left[\frac{\langle \hat\theta_\tau - \theta_\star, v\rangle^2}{\|v\|^2_{D_\tau^{-1}}}\right] \geq \frac{\operatorname{Breg}(H_0, H_n)^2}{\lambda + \operatorname{Breg}(H_0, H_n)} - 1\,.$$

The proof is based on Theorem 3 in combination with a concentration of measure argument showing that $\tau = n$ with suitably large probability. Details are available in Appendix C.

Corollary 5 *Let $(b_t)_{t\in[0,n]}$ be a process in $\mathbb{R}^d$ and $a_t = -\langle \hat\varphi_t, b_t\rangle$ as in Section 2. With τ as in Equation* (12)*, define a process of covariates $(x_t)_{t\in[0,n]}$ in $\mathbb{R}^{d+1}$ by $x_t = (a_t, , b_t)$ for $t \leq \tau$ and $x_t = \mathbf{0}$ otherwise. Then with $\hat\theta_n$ the ridge regression estimator in Equation* (11) *and $\theta_\star = \mathbf{0}$,*

$$\mathbb{E}_{\mathbf{0}}\left[\frac{\langle \hat\theta_n - \theta_\star, v\rangle^2}{\|v\|^2_{D_n^{-1}}}\right] \geq \frac{\operatorname{Breg}(H_0, H_n)^2}{\lambda + \operatorname{Breg}(H_0, H_n)} - 1\,.$$

Proof Simply note that $\hat\theta_n = \hat\theta_\tau$ and apply Theorem 4. ■

5. Open problem of Vakili et al. (2021b)

Vakili et al. (2021b) ask whether or not the kernel version of the self-normalised inequality in Equation (2) is tight. Our lower bound answers this question positively in certain instances. Let $\mathcal{X}$ be a compact metric space and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel of Mercer type, which means there exists an orthonormal basis $(e_m)_{m=1}^\infty$ for $L_2(\mathcal{X})$ and eigenvalues $(\alpha_m)_{m=1}^\infty$ such that $k(x, y) = \langle \phi(x), \phi(y)\rangle$, where $\phi(x) = (\sqrt{\alpha_1} e_1(x), \sqrt{\alpha_2} e_2(x), \ldots)$ and the RKHS associated with the kernel is $\mathcal{H} = \{\sum_{m=1}^\infty w_m \sqrt{\alpha_m} e_m : w \in \ell_2\}$ with norms $\|f\| = \min\{\|w\| : w \in \ell_2, \sum_{m=1}^\infty w_m \sqrt{\alpha_m} e_m = f\}$. In the continuous time kernel regression problem there is a covariate process $(x_t)_{t\in[0,n]}$ in $\mathcal{X}$ and the learner observes the covariates and responses $(y_t)_{t\in[0,n]}$, which is a process satisfying $\mathrm{d}y_t = \langle \theta_\star, \phi(x_t)\rangle\,\mathrm{d}t + \mathrm{d}B_t$. The self-normalised inequality of Abbasi-Yadkori et al. (2011)[1] shows that if $D_n = \lambda\mathbb{1} + \int_0^n \phi(x_t)\phi(x_t)^\top\,\mathrm{d}t$ and

$$\hat\theta_n = D_n^{-1} \int_0^n \phi(x_t)\,\mathrm{d}y_t \tag{13}$$

[1] More precisely, the infinite-dimensional generalisation of the self-normalised inequality, which appears in the thesis of Abbasi-Yadkori (2013).

is the regularised least-squares estimator of $\theta_\star$, then with probability at least $1-\delta$,

$$|\langle \hat{\theta}_n - \theta_\star, v\rangle| \leq \|v\|_{D_n^{-1}} \left[\|\theta_\star\|\sqrt{\lambda} + \sqrt{2\log(1/\delta) + \log\det(D_0^{-1}D_n)}\right] .$$

The log determinant is data-dependent, but is often naively upper bounded by a quantity called the information gain, which is

$$\gamma_n(\phi;\lambda) = \sup_{(x_t)\in\mathcal{X}^{[0,n]}} \log\det\left(\mathbb{1} + \frac{1}{\lambda}\int_0^n \phi(x_t)\phi(x_t)^\top \,\mathrm{d}t\right) .$$

Substituting this into the self-normalised bound and integrating shows that

$$\mathbb{E}\left[\frac{\langle \hat{\theta}_n - \theta_\star, v\rangle^2}{\|v\|^2_{D_n^{-1}}}\right] = O\left(\lambda\|\theta_\star\|^2 + \gamma_n(\phi;\lambda)\right) . \tag{14}$$

Corollary 5 will show that for certain kernels and data generating processes this bound leaves little room for improvement.

Augmented Fourier kernels We now give an explicit calculation using a kernel that is the sum of the euclidean kernel and a Fourier kernel. Given $x \in \mathbb{R}$, let $[x]_{2\pi}$ be the $y \in [0, 2\pi]$ such that $y + 2\pi k = x$ for some integer k. Let $M > 0$ be a constant to be chosen later and $\mathcal{X} = [-M, M] \times [0, 2\pi]$. Define a sequence of reals $(\alpha_m)_{m=1}^\infty$ and a sequence of functions $(e_m)_{m=1}^\infty$ from $\mathbb{R}$ to $\mathbb{R}$ by $\alpha_m = 1/m^2$ and $e_m(x) = \cos(mx)$. Next, let ϕ be the feature map given by

$$\phi(x) = (x_1, \sqrt{\alpha_1}e_1(x_2), \sqrt{\alpha_2}e_2(x_2), \ldots) . \tag{15}$$

The associated kernel is $k(x, y) = \langle\phi(x), \phi(y)\rangle$ with the inner product taken in ℓ_2. We also let $\phi_\circ(x) = (\sqrt{\alpha_1}e_1(x_2), \sqrt{\alpha_2}e_2(x_2), \ldots)$. As a reminder, we are interested in the behaviour of the ridge regression estimator when $(x_t)_{t\in[0,n]}$ is a process in $\mathcal{X}$ and $(y_t)_{t\in[0,n]}$ is the solution to the stochastic differential equation

$$\mathrm{d}y_t = \langle\theta_\star, \phi(x_t)\rangle \,\mathrm{d}t + \mathrm{d}B_t . \tag{16}$$

The covariate process is chosen using the construction in the proof of Theorem 1. Let $b_t = [t]_{2\pi}$ be a process in $[0, 2\pi]$ and $a_t = -\langle\hat{\varphi}_t, \phi_\circ(b_t)\rangle$, where $\hat{\varphi}_t = H_t^{-1}S_t$ with

$$H_t = \lambda\mathbb{1} + \int_0^t \phi_\circ(b_s)\phi_\circ(b_s)^\top \,\mathrm{d}s \qquad\qquad S_t = \int_0^t \phi_\circ(b_s)\,\mathrm{d}y_s - \int_0^t \phi_\circ(b_s)a_s \,\mathrm{d}s .$$

Let $\tau = \min\{t : a_t^2 \geq M^2\}$ with $M^2 = \frac{2}{\lambda}(8 + 2/\lambda + 5\log\det(H_0^{-1}H_n))$ and

$$x_t = \begin{cases} (a_t, b_t) & \text{if } t \leq \tau \\ (0, \pi/2) & \text{otherwise} . \end{cases} \tag{17}$$

Theorem 6 *The following hold:*
(a) The information gain for the feature map ϕ in Equation (15) *satisfies*

$$\gamma_n(\phi;\lambda) = O\left(\sqrt{n/\lambda}\log(\mathrm{poly}(n, 1/\lambda))\right) .$$

(b) When the covariate process $(x_t)_{t\in[0,n]}$ *is chosen according to Equation* (17) *and the response process* $(y_t)_{t\in[0,n]}$ *satisfies Equation* (16). *Then, provided that* $\lambda = O(n^{1/3})$,

$$\mathbb{E}_{\mathbf{0}}\left[\frac{\langle\hat\theta_n - \theta_\star, v\rangle^2}{\|v\|^2_{D_n^{-1}}}\right] = \Omega\left(\sqrt{n/\lambda}\right) ,$$

where $\hat\theta_n$ *is the ridge regression estimator in Equation* (13).

The proof is given in Appendix D with *(a)* following from an explicit calculation and *(b)* by using Corollary 5. By substituting *(a)* into Equation (14) you can see that the lower bound on the risk in *(b)* matches the upper bound obtained from the self-normalised inequality up to logarithmic factors. That is, the self-normalised inequality of Abbasi-Yadkori et al. (2011) is tight up to logarithmic factors in non-trivial kernel regimes and there can be a considerable price for adaptive design.

6. Experiments

We evaluate the ridge regression estimator in the natural discrete time approximation of the continuous time construction. Let $(\eta_t)_{t=1}^n$ be a sequence of independent standard Gaussian random variables and $(b_t)_{t=1}^n$ be a deterministic sequence of vectors in $\mathbb{R}^d$. We define a sequence of random covariates $(x_t)_{t=1}^n$ with $x_t = (a_t, b_t)$ where $a_t \in \mathbb{R}$ is random (and adaptive). For the experiments we only consider the case where $\theta_\star = \mathbf{0}$ so the response is always $y_t = \eta_t$. Define $H_t = \lambda\mathbb{1} + \sum_{s=1}^t b_s b_s^\top$ and $S_t = \sum_{s=1}^t b_s(y_s - a_s)$ where $a_t = -\langle b_t, \hat\varphi_{t-1}\rangle$ with $\hat\varphi_t = H_t^{-1}\sum_{s=1}^t b_s(y_s - a_s)$. We evaluate the performance of the ridge regression estimator with $\lambda = 1$ in two experiments. The first where $d = 1$ and $b_t = 1$ for all t. The results and more details are in Figure 1. In the second experiment we use the augmented Fourier kernel from the previous section. The results and more details are in Figure 2. All histograms are plotted using data from $N = 10^3$ independent runs. The headline summary of both experiments is that the continuous-time theory also holds in discrete time.

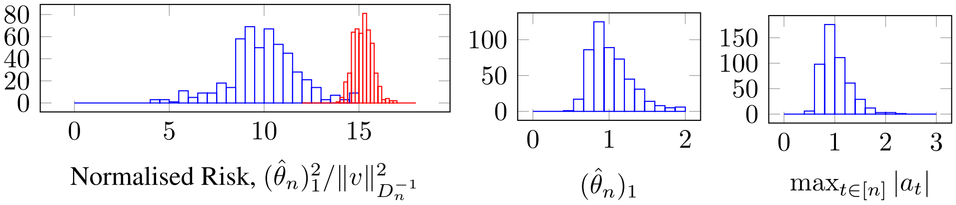

Figure 1: The histograms summarise the performance of the least-squares estimator on the simple 2-dimensional problem with $d = 1$ and $b_t = 1$ for all t and $n = 10^5$. On the left is the histogram of the normalised risk (blue) and the bound on the normalised risk obtained from Equation (2) with $\lambda = 1$ and $\delta = 1/2$ (red). The middle plot shows the histogram of the first coordinate of the estimated parameter, which is is concentrated around 1 (actual value 0). The last histogram shows the maximum magnitude of the adaptive part of the covariate, which just shows there is no funny business going on.

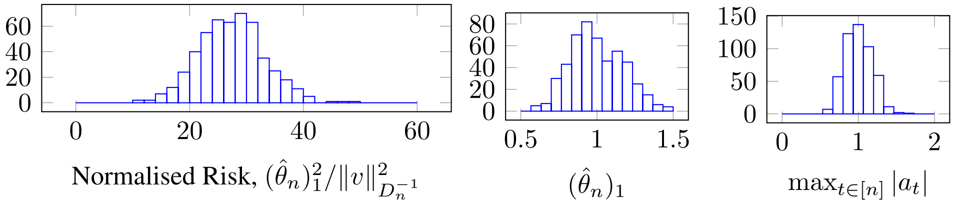

Figure 2: The plots summarise the performance of the least-squares estimator on the truncation of the kernel setting explained before. Let $n = 10^3$ and $(z_t)_{t=1}^n$ be a sequence of random variables uniformly distributed in $[0, 2\pi]$ and $b_t = (\sqrt{\alpha_1}e_1(z_t), \sqrt{\alpha_2}e_2(z_t), \ldots, \sqrt{\alpha_d}e_d(z_t))$, where the orthonormal functions $(e_m)_{m=1}^d$ were introduced in Section 5 and $d = 50$. The upper bound given by the self-normalised bound is very tightly concentrated about 60 for this data. On the middle plot you can see the histogram of the first coordinate of the estimated parameter. As in Figure 1, the estimator is not concentrating about the truth. And also like in Figure 1, the maximum magnitude of the adaptive component of the covariate is well concentrated and small.

7. Discussion

Our results show that there can be a considerable price for handling adaptive design when constructing confidence intervals. In particular, the well-loved upper bounds based on self-normalised martingale methods are not improvable in general. We finish with a few remarks.

Discrete time The use of continuous time allowed for a rather clean analysis of the counter-example with an exact optimal risk calculation. We believe all calculations will hold approximately in discrete time, which is supported by the experimental evidence. Regrettably we were not able to find a standard result in the literature that would yield the desired result as a corollary of our analysis. Though it must be admitted that such a result may exist and even be well known to experts on numerical approximation of stochastic differential equations.

Implications for bandits Our results suggest that optimal regret may not be obtainable with a completely decoupled analysis of LinUCB. For the non-contextual case we suspect that the design matrix may by close to deterministic and by carefully controlling the sample path of the algorithm it might be possible to prove optimal regret. The contextual setting is more challenging, especially when contexts are adversarial. In that case we are more hesitant to speculate about whether or not LinUCB is optimal at all. Perhaps a lower bound can be established showing that some mechanism for introducing independence is indeed essential.

Regularisation Although it is a little orthogonal to our work, staring at the confidence bound for kernel linear regression, the regularisation term appears both in the effective dimension and in front of the parameter norm. As far as we know there has been very little work trying to balance these terms. In the (low-dimensional) linear bandit there is little to be gained from any optimisation because the effective dimension in that case depends only weakly on the regularisation. As we have seen, however, in the kernel setting the effective dimension can heavily depend on the level of regularisation. A deeper investigation may be justified.

References

Y. Abbasi-Yadkori. *Online learning for linearly parametrized control problems*. PhD thesis, 2013.

Y. Abbasi-Yadkori, D. Pál, and Cs. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320. Curran Associates, Inc., 2011.

P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

R. Degenne, H. Shao, and W. Koolen. Structure adaptive algorithms for stochastic bandits. 07 2020.

I. Karatzas and S. Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 2012.

K. Khamaru, Y. Deshpande, L. Mackey, and M. Wainwright. Near-optimal inference in adaptive linear regression. *arXiv preprint arXiv:2107.02266*, 2021.

T.L. Lai and C. Z. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

T. Lattimore and Cs. Szepesvári. The end of optimism? An asymptotic analysis of finite-armed linear bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 728–737, Fort Lauderdale, FL, USA, 2017. JMLR.org.

T. Lattimore and Cs. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Y. Li, Y. Wang, and Y. Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. In *Proceedings of the 32nd Conference on Learning Theory*, pages 2173–2174, Phoenix, USA, 2019. JMLR.org.

R. Liptser and V. Spokoiny. Deviation probability bound for martingales with applications to statistical estimation. *Statistics & probability letters*, 46(4):347–357, 2000.

A. Tirinzoni, M. Pirotta, M. Restelli, and A. Lazaric. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. *Advances in Neural Information Processing Systems*, 33:1417–1427, 2020.

S. Vakili, K. Khezeli, and V. Picheny. On information gain and regret bounds in gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 82–90. PMLR, 2021a.

S. Vakili, J. Scarlett, and T. Javidi. Open problem: Tight online confidence intervals for rkhs elements. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4647–4652. PMLR, 15–19 Aug 2021b.

M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini. Finite-time analysis of kernelised contextual bandits. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pages 654–663, Arlington, VA, USA, 2013. AUAI Press.

K. Zhang, L. Janson, and S. Murphy. Statistical inference with M-estimators on adaptively collected data. *Advances in Neural Information Processing Systems*, 34, 2021.

Appendix A. Technical Lemmas

Fact 7 *Suppose that G is a square matrix of the form*

$$G = \begin{bmatrix} A & B \\ C & D \end{bmatrix} ,$$

with D and $A - BD^{-1}C$ both invertible. The following hold:
(a) $G^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}$.
(b) $\log\det G = \log\det A + \log\det(D - CA^{-1}B)$.

Lemma 8 *Suppose that X is a random variable with $\mathbb{P}(|X| \geq 2\log(1/\delta)) \leq \delta$. Then, $\mathbb{E}[X^2] \leq 8$.*

Proof $\mathbb{E}[X^2] = \int_0^\infty \mathbb{P}(X^2 \geq t)\,\mathrm{d}t = \int_0^\infty \mathbb{P}(X \geq \sqrt{t})\,\mathrm{d}t \leq \int_0^\infty \exp\left(-\frac{\sqrt{t}}{2}\right)\mathrm{d}t = 8\,.$ ■

Appendix B. Proof of Lemma 2

By the formula for the determinant of a block matrix, the analysis in Step 3 of the proof of Theorem 1 and Jensen's inequality,

$$\begin{aligned}
\mathbb{E}[\log\det D_0^{-1} D_n] &= \log\det H_0^{-1} H_n + \mathbb{E}\left[\log\left(1 + \frac{\int_0^n a_t^2\,\mathrm{d}t - \left\|\int_0^n a_t b_t\right\|_{H_n^{-1}}^2 \mathrm{d}t}{\lambda}\right)\right] \\
&\leq \log\det H_0^{-1} H_n + \log\left(1 + \frac{1}{\lambda}\mathbb{E}\left[\int_0^n a_t^2\,\mathrm{d}t - \left\|\int_0^n a_t b_t\right\|_{H_n^{-1}}^2 \mathrm{d}t\right]\right) \\
&= \log\det H_0^{-1} H_n + \log\left(1 + \frac{1}{\lambda}\operatorname{Breg}(H_0, H_n)\right) \\
&\leq \log\det H_0^{-1} H_n + \log\left(1 + \frac{1}{\lambda}\log\det(H_0^{-1} H_n)\right) \\
&\leq \left(1 + \frac{1}{\lambda}\right)\log\det(H_0^{-1} H_n)\,.
\end{aligned}$$

Appendix C. Proof of Theorem 4

We start with a simple proposition:

Proposition 9 *With probability at least* $1 - \delta$*,*

$$\mathbb{P}\left(\textit{exists } t \in [0,n] : a_t^2 \geq \frac{2\|b_t\|^2}{\lambda}\left(\log(1/\delta) + \log\det(H_0^{-1}H_t)\right)\right) \leq \delta\,.$$

Proof By the second step of the proof of Theorem 1, the measure of $(\hat\varphi_t)$ under $\mathbb{P}$ and $\mathbb{P}_{\mathbf{0}}$ are the same. Under $\mathbb{P}_{\mathbf{0}}$, we have

$$S_t = \int_0^s b_s(\mathrm{d}y_s - a_s\,\mathrm{d}s) = \int_0^s b_s\,\mathrm{d}B_s + \int_0^s H_s^{-1}S_s\,\mathrm{d}s\,.$$

Therefore, $\mathrm{d}S_t = b_t\,\mathrm{d}B_t + S_t\,\mathrm{d}t$, which has a strong solution of $S_t = H_t\int_0^s H_s^{-1}b_s\,\mathrm{d}B_s$ and

$$\hat\varphi_t = \int_0^t H_s^{-1}b_s\,\mathrm{d}B_s\,.$$

Let $Q_t = H_0^{-1} + \int_0^t H_s^{-1}b_sb_s^\top H_s^{-1}\,\mathrm{d}s = 2H_0^{-1} - H_t^{-1}$. By the self-normalised inequality, with probability at least $1-\delta$ for all t it holds that

$$\|\hat\varphi_t\|^2_{Q_t^{-1}} \leq 2\log\left(\frac{1}{\delta}\right) + \log\det\left(2\mathbb{1} - H_0H_t^{-1}\right)\,.$$

On the event that the above inequality holds for all t,

$$a_t^2 = \langle b_t, \varphi_t\rangle^2 \leq \|b_t\|^2_{Q_t}\|\varphi_t\|^2_{Q_t^{-1}} \leq \frac{2\|b_t\|^2}{\lambda}\left(2\log\left(\frac{1}{\delta}\right) + \log\det\left(2\mathbb{1} - H_0H_t^{-1}\right)\right)\,.$$

The result follows because

$$\begin{aligned}\log\det(2\mathbb{1} - H_0H_t^{-1}) &\leq \operatorname{tr}(\mathbb{1} - H_0H_t^{-1}) = \operatorname{tr}(H_t^{-1}(H_t - H_0)) \\ &= -\operatorname{Breg}(H_t, H_0) + \log\det(H_0^{-1}H_t) \leq \log\det(H_0^{-1}H_t)\,.\end{aligned}$$

■

Moving now to the proof of Theorem 4. Note that $\hat\theta_n = \hat\theta_\tau$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\frac{\langle\hat\theta_\tau - \theta_\star, v\rangle^2}{\|v\|^2_{D_\tau^{-1}}}\right] &\geq \mathbb{E}\left[\mathbf{1}_{\tau=n}\frac{\langle\hat\theta_n - \theta_\star, v\rangle^2}{\|v\|^2_{D_n^{-1}}}\right] \\ &\geq \mathbb{E}\left[\frac{\langle\hat\theta_n - \theta_\star, v\rangle^2}{\|v\|^2_{D_n^{-1}}}\right] - \mathbb{E}\left[\mathbf{1}_{\tau<n}\frac{\langle\hat\theta_n - \theta_\star, v\rangle^2}{\|v\|^2_{D_n^{-1}}}\right] \\ &\geq \frac{\operatorname{Breg}(H_0, H_n)^2}{\lambda + \operatorname{Breg}(H_0, H_n)} - \mathbb{E}\left[\mathbf{1}_{\tau<n}\frac{\langle\hat\theta_n - \theta_\star, v\rangle^2}{\|v\|^2_{D_n^{-1}}}\right]\,.\end{aligned}$$

The negative term is upper bounded by

$$\mathbb{E}\left[\mathbf{1}_{\tau<n}\frac{\langle\hat{\theta}_n-\theta_\star,v\rangle^2}{\|v\|^2_{D_n^{-1}}}\right]\leq\sqrt{\mathbb{P}(\tau<n)\mathbb{E}\left[\frac{\langle\hat{\theta}_n-\theta_\star,v\rangle^4}{\|v\|^4_{D_n^{-1}}}\right]}\,.$$

Let $\Delta^2=\langle\hat{\theta}_n,v\rangle^2/\|v\|^2_{D_n^{-1}}$. By the self-normalised inequality Equation (2),

$$\mathbb{P}\left(\Delta^2-\log\det(D_0^{-1}D_n)\geq 2\log(1/\delta)\right)\leq\delta\,.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\Delta^4\right]&=\mathbb{E}\left[(\Delta^2-\log\det(D_0^{-1}D_n)+\log\det(D_0^{-1}D_n))^2\right]\\
&\leq 2\mathbb{E}\left[(\Delta^2-\log\det(D_0^{-1}D_n))^2+\log\det(D_0^{-1}D_n)^2\right]\\
&\leq 16+2\mathbb{E}\left[\log\det(D_0^{-1}D_n)^2\right] && \text{Lemma 8}\\
&=16+2\mathbb{E}\left[\left(\log\det(H_0^{-1}H_n)+\log\left(1+\frac{1}{\lambda}\left(\int_0^n a_t^2-\left\|\int_0^n a_tb_t\right\|^2_{H_n^{-1}}\right)\right)\right)^2\right]\\
&\leq 16+4\left(\log\det(H_0^{-1}H_n)\right)^2+2\mathbb{E}\left[\left(\log\left(1+\frac{1}{\lambda}\left(\int_0^n a_t^2-\left\|\int_0^n a_tb_t\right\|^2_{H_n^{-1}}\right)\right)\right)^2\right]\\
&\leq 16+4\left(\log\det(H_0^{-1}H_n)\right)^2+2\log\left(3+\frac{1}{\lambda}\mathbb{E}\left[\int_0^n a_t^2-\left\|\int_0^n a_tb_t\right\|^2_{H_n^{-1}}\right]\right)^2\\
&=16+4\left(\log\det(H_0^{-1}H_n)\right)^2+2\log\left(3+\frac{1}{\lambda}\operatorname{Breg}(H_0,H_n)\right)^2\\
&\leq 24+4\left(1+\frac{1}{\lambda}\right)\left(\log\det(H_0^{-1}H_n)\right)^2\,.
\end{aligned}$$

Combining everything shows that

$$\begin{aligned}
\mathbb{E}[\Delta_\tau]&\geq\frac{\operatorname{Breg}(H_0,H_n)^2}{\lambda+\operatorname{Breg}(H_0,H_n)}-\sqrt{\mathbb{P}(\tau<n)\left(24+4\left(1+1/\lambda\right)\left(\log\det(H_0^{-1}H_n)\right)^2\right)}\\
&\geq\frac{\operatorname{Breg}(H_0,H_n)^2}{\lambda+\operatorname{Breg}(H_0,H_n)}-1\,.
\end{aligned}$$

Appendix D. Proof of Theorem 6

There are two steps. In the first step we apply Corollary 5 to bound the risk in terms of the Bregman divergences. In the second step we control the Bregman divergence and information gain.

Step 1: Bounding the risk in terms of Bregman divergences By Corollary 5 and the choice of M,

$$\mathbb{E}_{\mathbf{0}}\left[\frac{\langle\hat{\theta}_n-\theta_\star,v\rangle^2}{\|v\|^2_{D_n^{-1}}}\right]\geq\frac{\operatorname{Breg}(H_0,H_n)^2}{\lambda+\operatorname{Breg}(H_0,H_n)}-1\,.$$

Step 2: Bounding the information gain The information gain is

$$\gamma_n(\phi; \lambda) = \sup_{(x_t)_{t \in [0,n]}} \log\left(\det\left(\mathbb{1} + \frac{1}{\lambda}\int_0^n \phi(x_t)\phi(x_t)^\top\right)\right) .$$

Let $(x_t) \in \mathcal{X}^{[0,n]}$ be arbitrary and $(\lambda_m)_{m=1}^d$ be a decreasing sequence of the eigenvalues of $\int_0^n \phi(x_t)\phi(x_t)^\top \, \mathrm{d}t$ and A_n and B_n be matrices such that

$$D_0^{-1} D_n = \begin{bmatrix} \mathbb{1} + \frac{1}{\lambda} A_n & B_n^\top \\ B_n & \mathbb{1} + \frac{1}{\lambda} C_n \end{bmatrix} ,$$

where $A_n \in \mathbb{R}^{(1+p)\times(1+p)}$ for some p to be tuned later. By the Courant-Fischer-Weyl min-max principle,

$$\sum_{m=p+1}^\infty \lambda_m \le \operatorname{tr}(C_n) = \int_0^n \sum_{m=p}^\infty \alpha_m e_m(x_t)^2 \, \mathrm{d}t \le n \sum_{m=p}^\infty \alpha_m = O(n/p) .$$

On the other hand,

$$\begin{aligned}
\sum_{m=1}^p \log\left(1 + \frac{\lambda_m}{\lambda}\right) &\le p \log\left(1 + \sum_{m=1}^p \frac{\lambda_m}{p\lambda}\right) \\
&\le p \log\left(1 + \frac{\operatorname{tr}(A_n)}{p\lambda}\right) \\
&\le p \log\left(1 + \frac{nM^2 + n\sum_{m=1}^\infty \alpha_m}{p\lambda}\right) \\
&= O\left(p \log\left(1 + \frac{nM^2}{p\lambda}\right)\right) .
\end{aligned}$$

Therefore, letting $p = \left\lceil \sqrt{n/\lambda} \right\rceil$, and using the fact that $\log(1 + x) \le x$,

$$\begin{aligned}
\log\det(D_0^{-1} D_n) &= \sum_{m=1}^\infty \log\left(1 + \frac{\lambda_m}{\lambda}\right) \\
&\le \sum_{m=1}^p \log\left(1 + \frac{\lambda_m}{\lambda}\right) + \sum_{m=p+1}^\infty \frac{\lambda_m}{\lambda} \\
&= O\left(p \log\left(1 + \frac{nM^2}{p\lambda}\right) + \frac{n}{p\lambda}\right) \\
&= O\left(\sqrt{n/\lambda}\left(1 + \log\left(1 + M^2\sqrt{\frac{n}{\lambda}}\right)\right)\right) .
\end{aligned}$$

Since this holds for any $(x_t) \in \mathcal{X}^{[0,n]}$, it follows that

$$\gamma_n(\phi; \lambda) = O\left(\sqrt{n/\lambda}\left(1 + \log\left(1 + M^2\sqrt{\frac{n}{\lambda}}\right)\right)\right) .$$

The Bregman divergence satisfies

$$\begin{aligned}\operatorname{Breg}(H_0, H_n) &= \log\det(H_0^{-1} H_n) + \operatorname{tr}(H_n^{-1}(H_0 - H_n)) \\ &= \sum_{m=1}^{\infty} \left(\log\left(1 + \frac{n\alpha_m}{2\pi\lambda}\right) - 1 + \frac{1}{1 + \frac{n\alpha_m}{2\pi\lambda}} \right) \\ &= \Omega(\sqrt{n/\lambda})\,,\end{aligned}$$

where we used the fact that $x \mapsto \log(1+x) - 1 + 1/(1+x)$ is non-negative for $x \geq 0$ and larger than 0.19 for $x \geq 1$. Therefore, using that $x \mapsto x^2/(\lambda + x)$ is increasing for positive x and λ,

$$\mathbb{E}\left[\frac{\langle \hat{\theta} - \theta_\star, v\rangle^2}{\|v\|^2_{D_n^{-1}}}\right] \geq \frac{\operatorname{Breg}(H_0, H_n)^2}{\lambda + \operatorname{Breg}(H_0, H_n)} - 1 = \Omega\left(\frac{n/\lambda}{\lambda + \sqrt{n/\lambda}}\right)$$

So, by Theorem 3, whenever $\lambda = O(n^{1/3})$ the (normalised) risk of the ridge regression estimator satisfies for $\theta_\star = \mathbf{0}$,

$$\mathbb{E}\left[\frac{\langle \hat{\theta} - \theta_\star, v\rangle^2}{\|v\|^2_{D_n^{-1}}}\right] = \Omega(\sqrt{n/\lambda}) = \Omega\left(\frac{\gamma_n(\phi;\lambda)}{\log\operatorname{poly}(n, 1/\lambda)}\right)\,.$$

Appendix E. Covering arguments

Let us explain briefly the errors in the covering arguments by Lattimore and Szepesvári (2017) and Tirinzoni et al. (2020). Both works erroneously assume that if $x, y \in \mathbb{R}^d$ satisfy $|x| \leq |y|$ coordinate-wise, then $\|x\|_V \leq \|y\|_V$ for some positive-definite matrix V. Lattimore and Szepesvári (2017) use this implicitly when they argue the existence of a particular cover in the proof of their Theorem 8. Tirinzoni et al. (2020) use it explicitly in inequality (d) on page 49.