Self-normalization techniques for streaming confident regression

Odalric-Ambrym Maillard

Submitted to the Bernoulli

Self-normalization techniques for streaming confident regression

ODALRIC-AMBRYM MAILLARD[1]

[1] *Team TAO, Inria Saclay - Île de France & LRI, 91190 Gif-sur-Yvette, France*
E-mail: odalricambrym.maillard@inria.fr

We consider, in a generic streaming regression setting, the problem of building a confidence interval (and distribution) on the next observation based on past observed data. The observations given to the learner are of the form (x, y) with $y = f(x) + \xi$, where x can have arbitrary dependency on the past observations, f is unknown and the noise ξ is sub-Gaussian conditionally on the past observations. Further, the observations are assumed to come from some external filtering process making the number of observations itself a random stopping time. In this challenging scenario that captures a large class of processes with non-anticipative dependencies, we study the ordinary, ridge, and kernel least-squares estimates and provide confidence intervals based on self-normalized vector-valued martingale techniques, applied to the estimation of the mean and of the variance. We then discuss how these adaptive confidence intervals can be used in order to detect a possible model mismatch as well as to estimate the future (self-information, quadratic, or transportation) loss of the learner at a next step.

Keywords: Concentration inequalities; dependent variables; regression; self-normalized; sequential prediction
AMS 2000 Mathematics Subject Classification: XXXXX.

Contents

1. Introduction

Motivation Over the past few years, the growing requirement for decision strategies that are active and sequential in a number of areas of machine learning and statistics, has made the construction of *confidence intervals* a building block of utmost importance in order to design solutions that enjoy provably near-optimal performance guarantees, whether in term of generalization, robustness or regret minimization. This includes for instance designing modern confidence-based Monte-Carlo samplers, the problem of stochastic optimization as well as solving the exploration-exploitation trade-off in several variants of multi-armed bandits or in reinforcement learning.

Challenges A typical difficulty that appears when designing confidence intervals with *streaming* data is to handle, on the one hand, the possibly strong *dependency* between the current and past observations, and on the other hand, the possible drift (aka non-stationarity) of the observation process.

Non-stationarity has been studied from various standpoints in the literature, and we focus on *modeled* non-stationarity, when a class of processes is given that models the non-stationarity (think of a parametric regression model). This classical approach enables to reduce the problem of non-stationary prediction to that of estimating a (stationary) parameter. Using the standard terminology, we are in the realizable case when the process generating the signal belongs to the given class, and we are in the unrealizable case otherwise.

The dependency between the current and past observations can be handled, on the other hand, thanks to a powerful martingale method. When applicable, this generally avoids resorting to mixing coefficients and block techniques, which are less convenient due to the difficulty to estimate mixing coefficients in a fully empirical way. A specific difficulty of sequential decision making that is rarely considered is to handle the situation when the history of observations given to the learner may itself result from an active selection process, thus generating a history whose length is a random stopping time adapted to the filtration of the past observations. This situation typically appears in multi-armed bandits, when there is one learner per arm and the bandit algorithm must decide which arm to play and thus which learner to feed with new data. We take care of this additional dependency explicitly.

Last but not least, since the high-probability confidence intervals are typically used by an active decision algorithm, we target the construction of *empirical* confidence intervals with as few parameters as possible and where most process-dependent quantities are

estimated. In a streaming setting, we also care that all quantities can be updated at each time step at small cost.

Thus, the challenges we consider are the following: To build confidence intervals for a non-stationary signal in the realizable modeled-stationary case, to detect when we are instead in the non-realizable case, to handle the dependency of current observations with the past history as well as the random length of the history; finally, to design empirical confidence bounds (in the spirit of empirical-Bernstein bounds, as opposed to Bernstein bounds) with estimates that can be maintained at a low numerical cost. Along the way, we will also discuss how the designed confidence intervals can be used in order to estimate various notion of losses.

Setting In this paper, the problem of sequential prediction of a time series based on past observations is considered. More specifically, for a sequence of observations $y_1, \dots, y_n$ taking values in an abstract space $\mathscr{Y}$ and generated by some unknown stochastic process $\rho^\star$ which in turn is a probability measure on the set of all infinite-sequences in $\mathscr{Y}^\infty$, the learner aims to estimate the conditional distribution of the next stochastic outcome[1] Y_{n+1}. We consider a streaming protocol where at every time-step n, once the $\rho^\star$-conditional probability $\rho^\star(Y_{n+1}|Y_1, \dots, Y_n)$ of the next outcome Y_{n+1} is estimated, the realized value y_{n+1} of the outcome is revealed and the predictor goes on to estimate the distribution of the next outcome. The error of predicting a distribution π_n for the output y_{n+1} is measured by some loss $\ell_{n+1}(\pi_n)$. We do not assume that $\mathscr{Y}$ is bounded (or, for a discrete set $\mathscr{Y}$, that it is finite). We are looking for a learning agent that produces a sequence of distributions $\{\pi_n\}_n$ and achieves a low cumulative loss after any N steps

$$\sum_{n=1}^{N} \ell_{n+1}\big(\pi_n\big)\,.$$

In order to be able to predict the distribution of Y_{n+1}, the learner has access at step $n+1$ to some information x_{n+1} taking values in some abstract space $\mathscr{X}$. In the idealized scenario where a learner has access to an infinite memory, the side information x_{n+1} may simply be the sequence of all past observations $y^n = (y_n, \dots, y_1)$, in which case $\mathscr{X}$ corresponds to the set $\mathscr{Y}^\star$ of all finite sequences on $\mathscr{Y}$. Thus having access to side information comes without loss of generality. In a more realistic scenario, however, x_{n+1} may only be a finite sequence $(y_n, \dots, y_{n-k+1})$ of the k most recent observations, or even simply the very crude information given by the number of past observations $x_{n+1} = n$. We consider that x_{n+1} is a possibly stochastic function of y^n, and we sometimes write $x_{n+1} = x(y^n)$ to insist on this point. We denote by $h_n = \big((x_n, y_n), \dots, (x_1, y_1)\big)$ the sequence of all observations and information received up to step n. Our goal is then at step $n+1$, having been exposed to the past history h_n (perhaps by some external algorithm) and the received information x_{n+1}, to predict the conditional distribution $\rho^\star(Y_{n+1}|Y^n)$. Note that we do not assume that h_n is available to the learner –think for

[1] We use uppercase to denote the random variable Y, and lower case y to denote one specific realization of it

instance of the case of a limited memory, but that this "feeding" history h_n is the sequence of information and observation that the learner has been exposed to, sequentially, up to this step and that she may have used in order to improve her prediction.

Classes Let us continue this section with some necessary modeling assumptions. Indeed, to be completely arbitrary, the observations may be highly dependent and no assumptions on the nature of their dependence can be made. As such the task may be completely intractable as one cannot hope to get a bounded error without any assumption. It is then standard to introduce a restricted class of predictors and compete with the best predictor in that class.

Let us consider for clarity the case $\mathscr{Y} = \mathbb{R}$ of real-valued observations. In this case, a typical way to model the signal is to consider that the observations $h_n = \{(x_i, y_i)\}_{i\in[n]}$ are of the form

$$\begin{aligned} &y_i = f^\star(x_i) + \xi_i \text{ where } f^\star \in \mathscr{F} \\ &\text{and } \forall \lambda \in \mathbb{R} \ln \mathbb{E}\left[\exp(\lambda \xi_i) \middle| h_{i-1} \right] \leqslant \psi^\star(\lambda)\,, \end{aligned} \tag{1}$$

where $\psi^\star$ is the dual of a potential function ψ (convex, non-negative, null at 0) controlling the level of noise, and $\mathscr{F}$ is some given the function space modeling the non-stationarity. A typical function space $\mathscr{F}$ can be a parametric space $\mathscr{F}_{\varphi,B} = \{f : f(x) = \langle \theta, \varphi(x)\rangle, \theta \in \mathbb{R}^d, ||\theta|| \leqslant B\}$ where $\varphi : \mathscr{X} \to \mathbb{R}^d$ is a given feature function. It can be a reproducing kernel Hilbert space with given kernel k in the non-parametric case. In both cases, the function $f^\star$ does not change with time, and the problem reduces to estimating the (stationary) parameters describing the function. For this reason, such signals can be termed *parametric-stationary, non-stationary* signals. We study in section 3 different least-squares estimates and the construction of confidence intervals based on these estimates. Note that x_n is allowed to be arbitrary dependent on the past history h_{n-1}, making such models able to capture a large class of non-stationary signals with complex dependencies. Also, each given class $(\mathscr{F}, \psi)$ corresponds to a different modeling assumption on the non-stationarity and the noise, thus leading to different predictors $\pi(h_n)$.

Likewise, in the case when $\mathscr{Y} = \{1, \dots, S\}$ and the side observation is the full history of past observations $x_i = y^{i-1} (= (y_{i-1}, \dots, y_{-\infty}))$ a standard class is that of Markov models of a given order on some finite alphabet[2]. This results in different estimates built with the frequencies of words of various length, and confidence intervals using for instance the method of types (Csiszar, 1998), extended to the dependent scenario with a random history length. In the sequel, however, we won't discuss such constructions for the sake of brevity and focus on real-valued observations with a given $(\mathscr{F}, \psi)$.

Remark 1.1 *We focus on the construction of confidence distribution for one class, assuming we are in the realizable case, that is the signal is indeed modeled by the considered*

[2]The case when S is unknown or infinite can be handled by considering that a specific symbol is used to denote a symbol not previously seen, and thus estimating transitions to this symbol.

class. When several classes are considered, the problems of model selection and model aggregation naturally appear, as well as the problem of change point from one class to an other. Since all these questions crucially depend on our ability to first handle the realizable case, we thus leave the very appealing but orthogonal question of addressing model aggregation, selection and change-point based on these confidence distributions for future work. We discuss nonetheless how confidence intervals can be trivially used in order to estimate the validity of the realizable case assumption, and to help the learner predict her future loss.

Previous work We employ a powerful martingale technique from (Peña, Lai and Shao, 2008) that was also used for instance in (Abbasi-Yadkori, Pal and Szepesvari, 2011) or (Rusmevichientong and Tsitsiklis, 2010) in the setting of regression applied to the linear multi-armed bandit problem. Indeed, the requirement of adaptive confidence intervals naturally appears in the multi-armed bandit community where the number of observations of each arm is a random stopping time (due to the action selection algorithm), and where confidence bounds often need to be fully computable from data as they are used by the algorithms. See (Chu et al., 2011, Abbasi-Yadkori, Pal and Szepesvari, 2011) or (Krause and Ong, 2011, Valko et al., 2013, Grünewälder et al., 2010) for references making explicit use of such bounds in a multi-armed bandit context. The martingale technique from (Peña, Lai and Shao, 2008) was correctly applied for the first time in (Abbasi-Yadkori, Pal and Szepesvari, 2011) for regularized least-squares estimate, where its application is direct. We show how a careful application of the same technique enables to strengthen preliminary results from (Rusmevichientong and Tsitsiklis, 2010) for the more challenging ordinary least-squares estimate.

All these linear multi-armed bandits work require that the level of noise is given to the learner, which is often not the case in practice. We make use of techniques inspired from the empirical-Bernstein bounds from (Maurer and Pontil, 2009), but adapted to self-normalized vector-valued martingales with a random stopping time in order to derive bounds agnostic to the knowledge of the noise. We show that such bounds can actually be derived easily (although the derivation is bit long) when using the right concentration tools. We also discuss an alternative "corrective" noise estimate (see Section 3.2.2) directly derived from our bounds.

When working with dependent data, it is natural to consider a mixing assumption. This has been done for instance in the context of *bounded* sequences in (Kuznetsov and Mohri, 2015), where Hoeffding-type results for φ-mixing processes were used; see also (Audiffren and Ralaivola, 2014) in a bandit setting. The old block-techniques indeed allows the extension of many concentration results to the mixing scenario. It is however often difficult to use in practice since mixing coefficients are typically unknown to the learner. In our regression setting, we benefit from the special structure of the noise that enables us to apply the martingale method that turns out to be a powerful tool to handle dependent data.

Outline and contribution In Section 2, we first discuss the assumptions regarding the noise model, as well as the loss function. We make an explicit link between the choice

of the noise model and that of the underlying loss function that we believe is interesting. We discuss specifically the quadratic loss on the observations, its corresponding transportation loss on the distributions (defined by optimal transport), and the popular self-information loss.

In Section 3, we study three least-squares estimators of the unknown mean – an ordinary least-squares, a ridge estimate, and a kernel estimate, and provide our main results regarding the construction of confidence bounds on their estimation and prediction error, see Theorem 3.4, Theorem 3.3, Theorem 3.5. These bounds depend on the level of noise and we thus analyze the concentration of a simple variance estimate in Theorem 3.9 and Theorem 3.8 in the spirit of the work on empirical Bernstein inequalities (see e.g. (Maurer and Pontil, 2009)). We also illustrate these bounds and compare them against alternatives from Abbasi-Yadkori, Pal and Szepesvari (2011) and Rusmevichientong and Tsitsiklis (2010).

Motivated by the availability of a self-normalized confidence bound for the mean estimate, we propose in section 3.2.2, as a side result illustrating the use of confidence intervals, an alternative procedure designed to take care of the unknown level of noise, that is numerically efficient and shows excellent practical performance. It is by construction a lower estimate on the actual level of noise and is thus called "optimistic". It is however challenging to analyze its properties in the full-blown dependent scenario. We provide some hints in the independent setting and leave the general analysis of this estimate aside as this is not the main focus of this article.

Section 5 provides the key component of our contributions. The core of our work is a careful use of self-normalized techniques, combined with a peeling argument and elementary statistics. We show in this section that these simple techniques can actually lead to powerful results. We provide in Section 5.1 and 5.2 a detailed step by step derivation of our results that we tried to present in a simple and clear way. Among the interesting innovations of the proof, we isolate Lemma 5.3 about concentration for real-valued random variables that we believe is interesting beyond the scope of this paper, and, in the proof technique for ordinary-least squares regression, the localization of eigenvalues of the normalization matrix used in the self-normalized technique. To avoid clutter, we choose to provide all our result for the regression setting on the real line $\mathscr{Y} = \mathbb{R}$, although similar concentration results could be derived for Markov signals over a finite alphabet, by combining the method of types together with the Laplace method and peeling techniques.

2. Preliminary remarks regarding the losses

In this section, we spend a few lines to motivate two different ways to define the loss $\ell_{n+1}(\pi)$. This enables us to introduce some important notions and relate the link between the choice of the loss and the modeling assumption. The first one is to consider the loss between the unknown conditional *distribution* of the observation and the distribution predicted by the learner; this leads to a loss measured in terms of optimal transport. The second one is to measure the loss between the *observation* y_{n+1} and the distribution predicted by the learner; this leads to the popular notion of self-information loss. Formally,

this is captured by the two following alternative formulations, that are further justified in the next sub-sections. Let us consider that $\mathscr{X}$ and $\mathscr{Y}$ are metric spaces with known respective metrics given by $\ell_{\mathscr{X}}$ and $\ell_{\mathscr{Y}}$.

Formulation 1 (Transportation loss) *The loss of a distribution π on the metric space $(\mathscr{Y}, \ell_{\mathscr{Y}})$ for the prediction of observation $Y_{n+1} \sim \rho^\star(\cdot|Y^n)$ is given by*

$$\ell_{n+1}(\pi) = \mathscr{T}_{\ell_{\mathscr{Y}}}\left(\rho^\star(\cdot|Y^n), \pi\right),$$

where $\mathscr{T}_{\ell_{\mathscr{Y}}}$ is the optimal transport loss given by definition 2.1 below.

Formulation 2 (Self-information loss) *The loss of a distribution π on $\mathscr{Y}$ for the prediction of observation y_{n+1} is given by*

$$\ell_{n+1}(\pi) = \ell_I(y_{n+1}, \pi) = -\ln\left(p(y_{n+1})\right),$$

where $p(y)$ denotes the density of distribution π with respect to the reference measure on $\mathscr{Y}$ (e.g. Lebesgue, or counting in case $\mathscr{Y}$ is discrete).

Note We can highlight two red-line examples regarding $\mathscr{Y}$:
- $\mathscr{Y} = \mathbb{R}$, and $\ell_{\mathscr{Y}}$ is the quadratic loss $\ell_2(y_1, y_2) = \frac{1}{2R^2}(y_1 - y_2)^2$ for some positive R.
- $\mathscr{Y} = \{1, \dots, S\}$ is a discrete set of symbols $S \in \mathbb{N} \cup \{\infty\}$, with the grossiere metric

$$\ell_0(y_1, y_2) = \begin{cases} 0 & \text{if } y_1 = y_2, \\ 1 & \text{else.} \end{cases}$$

2.1. Transportation loss induced by $\ell_{\mathscr{Y}}$

The loss on $\mathscr{Y}$ naturally induces a loss on measures, via optimal transport.

Definition 2.1 (Transportation loss) *The transportation loss between distributions π_1 and π_2 for the loss $\ell_{\mathscr{Y}}$ is given by*

$$\mathscr{T}_{\ell_{\mathscr{Y}}}(\pi_1, \pi_2) \stackrel{\text{def}}{=} \inf_{\gamma \in \Gamma(\pi_1, \pi_2)} \int_{\mathscr{Y} \times \mathscr{Y}} \ell_{\mathscr{Y}}(y_1, y_2) d\gamma(y_1, y_2),$$

where $\Gamma(\pi_1, \pi_2)$ denotes all couplings of distributions γ with first marginal π_1 (that is, $\pi_1 = \int_{\mathscr{Y}} d\gamma(\cdot, y_2)$) and second marginal $\pi_2 (= \int_{\mathscr{Y}} d\gamma(y_1, \cdot))$.

Remark 2.2 *A well-known example is when the loss is associated to a p-norm $\ell_p(y_1, y_2) = ||y_1 - y_2||_p^p$. In this case, we recover a Wasserstein (aka the Fréchet or Mallows or Kantorovitch) distance*

$$\mathscr{T}_{\ell_p}(\pi_1, \pi_2)^{1/p} = \mathscr{W}_p(\pi_1, \pi_2) = \left(\inf_{\gamma \in \Gamma(\pi_1, \pi_2)} \int_{\mathscr{Y} \times \mathscr{Y}} ||y_1 - y_2||_p^p d\gamma(y_1, y_2) \right)^{1/p}.$$

In general, solving the optimal transport problem is not an easy task. For our two illustrative cases, the optimal transport problem reduces to a nicer formulation. In dimension 1, the value of the optimal transport has an explicit form for convex potentials:

Proposition 2.3 *Assume that $\mathscr{Y} = \mathbb{R}$, and $\ell_{\mathscr{Y}}(y_1, y_2) = \psi(y_1 - y_2)$, where $\psi : \mathbb{R} \to \mathbb{R}^+$ is a convex potential. Then, the transportation loss induced by $\ell_{\mathscr{Y}}$ satisfies*

$$\mathscr{T}_{\ell_{\mathscr{Y}}}(\pi_1, \pi_2) = \int_0^1 \psi\Big(\Pi_1^{-1}(t) - \Pi_2^{-1}(t)\Big) dt\,,$$

where $\Pi_1 = \mathbb{P}_{X\sim\pi_1}(X \leqslant \cdot)$ is the cdf of π_1 and $\Pi_2 = \mathbb{P}_{X\sim\pi_2}(X \leqslant \cdot)$ is the cdf of π_2.

For the potential $\psi(y) = y^2$, it is not difficult to derive that

$$\mathscr{W}_2\big(\pi_1, \pi_2\big) = \left[(\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2 + 2\Big(\sigma_1\sigma_2 + \mu_1\mu_2 - \int_0^1 \Pi_1^{-1}(t)\Pi_2^{-1}(t)dt\Big)\right]^{1/2},$$

where μ_1, σ_1 (resp. μ_2, σ_2) are the mean and standard deviation of π_1 (resp. π_2). From this formula, we easily deduce the special form for Gaussian real-valued distributions[3]:

$$\mathscr{W}_2\big(\mathscr{N}(\mu_1, \sigma_1^2), \mathscr{N}(\mu_2, \sigma_2^2)\big) = \sqrt{(\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2}\,.$$

When $\mathscr{Y}$ is discrete, we recover another classical notion:

Proposition 2.4 *Assume that $\mathscr{Y} = \{1, \dots, S\}$, and $\ell_0(y_1, y_2) = \mathbb{I}\{y_1 \neq y_2\}$. Then, the transportation loss induced by ℓ_0 coincides with the total variation distance*

$$\mathscr{T}_{\ell_0}\big(\pi_1, \pi_2\big) = d_{TV}(\pi_1, \pi_2) = \frac{1}{2}\sum_{y\in\mathscr{Y}} |\pi_1(y) - \pi_2(y)|\,.$$

2.2. Self-information loss induced by $\ell_{\mathscr{Y}}$

In this section, we show that a given loss $\ell_{\mathscr{Y}}$ on $\mathscr{Y}$ induces a natural class of distributions whose tails are controlled, and vice versa, that a given class of distributions naturally induces a loss on $\mathscr{Y}$. This second point of view is especially useful when the given space is for instance discrete with no useful metric, while the first one is more natural when we have access to some informative loss. Now, the two points of view are fairly exchangeable.

From loss to class For the first claim, let us consider that $\mathscr{Y} \subset \mathbb{R}^d$ and that the loss comes from a convex potential $\ell_{\mathscr{Y}}(y, y') = \psi(y - y')$, where ψ is convex, non-negative and null in 0. Now, one can always decompose $Y_n = \mu_n + \xi_n$, where we introduced the conditional mean $\mu_n = \mathbb{E}[Y_n|Y^{n-1}]$ and the conditional centered noise ξ_n. Controlling the tails of ξ_n is crucially important in order to control the loss of any algorithm. We introduce the following definition

[3] The formula for general multi-variate Gaussians, also known as the Bures metric, is given by

$$\mathscr{W}_2\big(\mathscr{N}(\mu_1, \Sigma_1), \mathscr{N}(\mu_2, \Sigma_2)\big) \quad = \quad \sqrt{||\mu_1 - \mu_2||_2^2 + \mathrm{Tr}\Big(\Sigma_1 + \Sigma_2 - 2\Big(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\Big)^{1/2}\Big)}\,.$$

Definition 2.5 (Loss-adapted noise) *The noise is adapted to the loss $\ell_{\mathscr{Y}}$ if it satisfies*

$$\forall n \in \mathbb{N}, \ln \mathbb{E}[\exp(\langle \lambda, \xi_n \rangle) | Y^{n-1}] \leqslant \psi^\star(\lambda) \quad \textit{where} \quad \psi^\star(\lambda) = \sup_{y \in \mathscr{Y} - \mathscr{Y}} \langle \lambda, y \rangle - \psi(y)\,.$$

for all λ such that the right-hand side term is finite. That is, the cumulative generative function of the noise is dominated by the Legendre-Fenchel dual of the potential loss function.

For instance on $\mathscr{Y} = \mathbb{R}$, the dual of the quadratic potential $\psi_2(z) = \frac{z^2}{2R^2}$ is $\psi_2^\star(\lambda) = \frac{\lambda^2 R^2}{2}$, and the definition reduces to that of a R-sub-Gaussian noise. Thus, a given loss function $\ell_{\mathscr{Y}}$ induces a natural class of distributions.

There are many ways to justify such a notion. First it is quite intuitive, since when the loss grows fast away from 0, this means we cannot tolerate a too large noise, while a very flat potential loss function will produce low error even for a large noise. Now, a more formal way is simply to control the loss $\ell_{\mathscr{Y}}(y_n, \mu_n)$ between the random observation and its expectation. For clarity, let us consider that $\mathscr{Y} = \mathbb{R}$ and ψ is strongly convex, non-negative with $\psi(0) = 0$, so that we can write ψ_+ the (invertible) restriction of ψ to $\mathscr{D}_\psi \cap \mathbb{R}^+$ and ψ_- to $\mathscr{D}_\psi \cap \mathbb{R}^-$. Let g be some upper envelope function on $\lambda \to \ln \mathbb{E}[\exp(\lambda \xi_n) | Y^{n-1}]$. Then the error obtained by the oracle predictor μ_n satisfies

$$\begin{aligned}
\mathbb{P}\Big(\ell_{\mathscr{Y}}(Y_n, \mu_n) \geqslant t | Y^{n-1}\Big) &= \mathbb{P}\Big(\psi(\xi_n) \geqslant t | Y^{n-1}\Big) \\
&\leqslant \mathbb{P}\Big(\xi_n \geqslant \underbrace{\psi_+^{-1}(t)}_{t_+} | Y^{n-1}\Big) + \mathbb{P}\Big(-\xi_n \geqslant \underbrace{\psi_-^{-1}(t)}_{t_-} | Y^{n-1}\Big) \\
&\leqslant \exp(-g^\star(t_+)) + \exp(-g^\star(t_-))\,,
\end{aligned}$$

where $g^\star$ is the Legendre-Fenchel conjugate of g. In particular, when g is $\psi^\star$, then $\psi^{\star\star} = \psi$ by convexity of ψ, and $g^\star(t_+) = g^\star(t_-) = t$. Thus, we obtain in this case that

$$\mathbb{P}\Big(\ell_{\mathscr{Y}}(Y_n, \mu_n) \geqslant t | Y^{n-1}\Big) \leqslant 2\exp(-t)\,.$$

Thus, the probability that the loss of the optimal predictor is larger than t decays exponentially fast with t. This desirable property does not necessary hold when the noise is not adapted to the loss. Further, if g_+ is another function such that $\forall \lambda g_+(\lambda) \geqslant g(\lambda)$, then $g_+^\star(x) \leqslant g^\star(x)$ and thus the bound on the loss of the optimal predictor for a noise ξ' with upper envelope function g_+ is larger than for g, which explains why we want a tight envelope.

From class to loss We now proceed backward, starting from a class of distributions in order to build a loss function. For an abstract space $\mathscr{Y}$, there is no necessarily natural notion of linearity giving a meaning to $y - y'$ or $\langle \lambda, y \rangle$. Think for instance of a discrete space $\mathscr{Y} = \{1, \dots, S\}$. However, one can still consider a class of distributions, and functions on $\mathscr{Y}$. Thus, given a candidate distribution π for Y_n, we consider $g(\lambda) = \ln \mathbb{E}_\pi \exp \lambda(Y)$, for any function λ that is bounded, continuous. Interpreting g as the dual $\psi^\star$ of a convex

loss, it is then natural to look at its dual $g^\star$ in order to recover the definition of the loss. Note that $g^\star$ acts on measures ν. It comes

$$\begin{aligned} g^\star(\nu) &= \sup_{\lambda \in \mathscr{C}_B(\mathscr{Y})} (\nu, \lambda) - g(\lambda) \\ &= \sup_{\lambda \in \mathscr{C}_B(\mathscr{Y})} \mathbb{E}_\nu[\lambda(Y)] - \ln \mathbb{E}_\pi \exp \lambda(Y) \\ &= KL(\nu, \pi)\,, \end{aligned}$$

where $(\cdot,\cdot)$ is the duality product, and $\mathscr{C}_B(\mathscr{Y})$ are continuous bounded functions on $\mathscr{Y}$. Thus, the loss induced by π on probability measures coincides with the Kullback-Leibler divergence. In particular interpreting an observation Y_n as a Dirac distribution at point Y_n, it comes $-\ln(p(Y_n))$ where p denotes the density of π with respect to the reference measure. This justifies the introduction of the following

Definition 2.6 (Self-information loss) *The loss of a distribution π on $\mathscr{Y}$ with density p is given by*

$$\ell_I(y, \pi) = -\ln p(y)\,.$$

Remark 2.7 *The self-information loss is a popular and standard loss in the literature on sequential prediction. Its expectation with respect to y coincides with the Kullback-Leibler of the distribution of y with π. The notion of loss-adapted noise, although less frequent, also appears in certain works. We refer to (Merhav and Feder, 1998) for an extended study of universal sequential prediction and further details.*

The self-information loss may look very appealing due to its interpretation in terms of models and its applicability to any abstract space $\mathscr{Y}$. However, one may be aware of a specific issue, which we illustrate here with a simple Gaussian model. For a Gaussian model, say with constant variance σ^2, the self-information loss at point y writes

$$\ell_I(y, \mathscr{N}(f, \sigma^2)) = \frac{(y-f)^2}{2\sigma^2} + \frac{1}{2}\ln(2\pi\sigma^2)\,.$$

When the variance is unknown and (estimated) by a procedure, a model that estimates a high variance (and thus a large bandwidth and confidence intervals) will mistakingly consider it should incur a lower loss than a model that estimates a low variance. This raises a specific issue. In the context of model selection, this makes any practical use of this loss by an algorithm tricky in case the observation noise is unknown. This issue does not appear when the noise of the process is perfectly known as it is only due to the term σ^2 that must be estimated. Now, such a phenomenon does not occur when considering, in the same Gaussian scenario, the transportation loss (see Section 2.1). Such questions should be taken into considerations when choosing the loss.

3. Least-squares strategies on the real line

In this section, we continue introducing the objects of interest, that are standard estimates, mostly in order to fix the notations. We focus on the case $\mathscr{Y} = \mathbb{R}$ when observa-

tions are real-valued. Given a parametric modeling class $(\mathscr{F}, \psi)$ with $\mathscr{F} = \{f : f(x) = \langle \theta, \varphi(x) \rangle, \theta \in \mathbb{R}^d, ||\theta|| \leqslant B\}$ it is natural to find at step $N+1$ a parameter θ_N^{ψ} that minimizes the past empirical loss, that is which solves the following problem

$$\min_{\theta \in \Theta} \sum_{n=1}^{N} \psi\Big(y_n - \theta^\top \varphi(x_n)\Big) \,. \tag{2}$$

In the sequel, we focus on the quadratic potential $\psi(x) = \frac{x^2}{2R^2}$, as it corresponds to an R-sub-Gaussian noise assumption and leads to a fully explicit solution to (2). Indeed, in this case every solution θ to (2) must satisfy

$$G_N \theta = \sum_{n=1}^{N} \varphi(x_n) y_n \,, \text{ where } \quad G_N = \sum_{n=1}^{N} \varphi(x_n) \varphi(x_n)^\top \,.$$

We refer to the problem formulation (1) using the quadratic potential as the *sub-Gaussian streaming regression* model.

Pseudo-inverse solution For convenience, we introduce $Y_N = (y_1, \dots, y_N)^\top \in \mathbb{R}^N$ and the $N \times D$ matrix $\Phi_N = (\varphi^\top(x_1), \dots, \varphi^\top(x_N))^\top$. Using these notations a solution to (2) must satisfy $G_N \theta = \Phi_N^\top Y_N$, where $G_N = \Phi_N^\top \Phi_N$. A specific solution is then $\theta_N^\dagger = G_N^\dagger \Phi_N^\top Y_N$ where $G_N^\dagger$ denotes the pseudo-inverse of G_N, and the space of solution can be described as

$$\Theta_N = \{\theta \in \Theta : G_N(\theta_N^\dagger - \theta) = 0\} = \{\theta_N^\dagger + \ker(G_N)\} \cap \Theta \,.$$

Note that when $\Theta = \mathbb{R}^d$ and G_N is invertible, $G_N^\dagger = G_N^{-1}$ and the solution reduces to the ordinary least squares solution $\theta_N^\dagger = \widehat{\theta}_N = G_N^{-1} \Phi_N^\top Y_N$. Further, it holds for all x,

$$|f^\star(x) - f_{\widehat{\theta}_N}(x)| \leqslant ||\theta^\star - \widehat{\theta}_N||_{G_N} ||\varphi(x)||_{G_N^{-1}} \,. \tag{3}$$

Equation 3 thus enables to control the prediction error at any point x. On the other hand, in the general case, for all $\theta \in \Theta_N$ it holds

$$\sum_{n=1}^{N} (f^\star(x_n) - f_\theta(x_n))^2 = \sum_{n=1}^{N} (\theta^\star - \theta)^\top \varphi(x_n) \varphi(x_n)^\top (\theta^\star - \theta) = ||\theta^\star - \theta||_{G_N}^2 \,.$$

Thus in case of over-fitting, we must have $\forall \theta \in \Theta_N$, $||\theta^\star - \theta||_{G_N} = 0$, which indicates that G_N is unable to provide information about the localization of $\theta^\star$. Also in that case, G_N is not invertible and we can no longer apply (3).

Regularized solution When G_N is not invertible, another approach is to regularize the problem by constraining the quadratic norm of the parameter. This leads to the ridge solution (aka Tikhonov–Phillips regularization, or ℓ_2-regularization) given for a regularization parameter $\lambda \in \mathbb{R}_\star^+$ by

$$\theta_{N,\lambda} = G_{N,\lambda}^{-1} \mathbf{\Phi}_{[N]}^\top \mathbf{Y}_{[N]} \,. \text{ where } G_{N,\lambda} = \mathbf{\Phi}_{[N]}^\top \mathbf{\Phi}_{[N]} + \lambda I_d \,.$$

Remark 3.1 *The regularization parameter λ is better understood from a Bayesian point of view in the Gaussian i.i.d. model. Indeed, if we pick a random function using $\theta \sim \mathcal{N}(0,\Sigma)$ (a prior which models how hard it is to generate one function), and assume that the noise is exactly i.i.d. Gaussian ($\xi_n \sim \mathcal{N}(0,\sigma^2)$ for each n) and x_n is independent on the past, then the posterior mean function is a random function following the law $\widehat{f}_N(x)|x,x_1,\dots,x_N,y_1,\dots,y_n \sim \widehat{\pi}_N(x) = \mathcal{N}(\mu_N(x),\sigma_N^2(x))$ where*

$$\begin{aligned}\mu_N(x) &= \varphi(x)^\top(\mathbf{\Phi}_{[N]}^\top\mathbf{\Phi}_{[N]}+\sigma^2\Sigma^{-1})^{-1}\mathbf{\Phi}_{[N]}^\top\mathbf{Y}_{[N]}\\ \sigma_N^2(x) &= \sigma^2\varphi(x)^\top(\mathbf{\Phi}_{[N]}^\top\mathbf{\Phi}_{[N]}+\sigma^2\Sigma^{-1})^{-1}\varphi(x)\,.\end{aligned}$$

This enables to directly interpret λ as being an a priori value of the variance term, and to recover the ℓ_2-regularized least-squares using $\Sigma = \frac{\sigma^2}{\lambda}I_d$. See appendix A for details.

Remark 3.2 *It is actually natural to combine the pseudo-inverse and ridge solution in the case of a finite-dimensional function space. Indeed, one may use the ridge estimate $\theta_{N,\lambda}$ in the first rounds as long as G_N is not invertible, and then switch to choosing the ordinary least-squares estimate $\widehat{\theta}_N$ for larger values of N.*

Kernel solution Finally, in the case when the function space is of large or infinite dimension, it is desirable to directly compute the estimate of the mean without computing the possibly infinite dimensional parameter θ. Actually, we can derive the equivalent form

$$\begin{aligned}\mu_N(x) &= \varphi(x)^\top\Sigma\mathbf{\Phi}_{[N]}^\top(\mathbf{\Phi}_{[N]}\Sigma\mathbf{\Phi}_{[N]}^\top+\sigma^2 I_N)^{-1}\mathbf{Y}_{[N]}\\ \sigma_N^2(x) &= \varphi(x)^\top\Sigma\varphi(x)-\varphi(x)^\top\Sigma\mathbf{\Phi}_{[N]}^\top(\mathbf{\Phi}_{[N]}\Sigma\mathbf{\Phi}_{[N]}^\top+\sigma^2 I_N)^{-1}\mathbf{\Phi}_{[N]}\Sigma\varphi(x)\,.\end{aligned}$$

This "functional" form is convenient as it generalizes to infinite dimensions. Indeed let us introduce $k(x,x)$ to generalize $\varphi(x)^\top\Sigma\varphi(x)$, as well as $k_N(x) = (k(x_n,x))_{n\in[N]}$ and $\mathbf{K}_N = (k(x_i,x_j))_{i,j\in[N]}$. We then get

$$\begin{aligned}\mu_N(x) &= k_N(x)(\mathbf{K}_N+\sigma^2 I_N)^{-1}\mathbf{Y}_{[N]}\\ \sigma_N^2(x) &= k(x,x)-k_N(x)^\top(\mathbf{K}_N+\sigma^2 I_N)^{-1}k_N(x)\,.\end{aligned}$$

In the next sections, we provide tight confidence bounds on the estimation error of the next observation for each of these three procedures, namely the ordinary, ridge and kernel least squares.

3.1. Mean estimates

In this section, we assume that level of noise R is known (or upper bounded by a known constant), and derive confidence intervals on the mean function. We discuss the estimation of the noise level later in section 3.2,

Let us remind that the history of observations $h_N = \{(x_1,y_1),\dots,(x_N,y_N)\}$ is generated by a random process and that the observations points x_n are allowed to be arbitrary

function of the past observations before step n. Further, N is allowed to be a random stopping time for the filtration of the past.

In this context, we obtain the following results on the estimation of the mean function at any point $x \in \mathscr{X}$. At a high level, Theorem 3.3 applies to function spaces of small dimension, or when the number of observations is large enough that G_N is invertible. Theorem 3.4 applies to function space of large but finite dimensions, when N is too small to ensure that G_N is invertible. This is the reason for using a regularization. Theorem 3.5 finally applies to generic, possibly infinite dimensional function space.

Theorem 3.3 (Ordinary Least-squares) *Assume that N is a stopping time adapted to the filtration of the past. Then in the sub-Gaussian streaming regression model it holds*

$$\mathbb{P}\left(\exists x \in \mathscr{X}\, |f^\star(x) - f_{\theta^\dagger_N}(x)| \geqslant 2R||\varphi(x)||_{G^\dagger_N}\sqrt{\ln\left(\frac{\kappa_d(e^2\Lambda_N^2)}{\delta}\right)} \cap\ \lambda_{\min}(G_N) > 0\right) \leqslant \delta.$$

where $\kappa_d(x) = \frac{2}{3}\pi^2 \ln^2(x/e)\left\lceil \frac{\ln(x)}{2}\right\rceil[(12(d+1)\sqrt{d})^d x^d + d]$ *and* $\Lambda_N = \lambda_{\max}(G_N)$.

Theorem 3.4 (Regularized Least-squares) *Assume that N is a stopping time adapted to the filtration of the past. Then in the sub-Gaussian streaming regression model it holds*

$$\mathbb{P}\Bigg(\exists x \in \mathscr{X}\, |f^\star(x) - f_{\theta_{N,\lambda}}(x)| \quad \geqslant \quad ||\varphi(x)||_{G^{-1}_{N,\lambda}}\Bigg[\frac{\lambda}{\sqrt{\lambda_{\min}(G_{N,\lambda})}}||\theta^\star||_2 + R\sqrt{2\ln\left(\frac{det(G_N+\lambda I)^{1/2}}{\delta det(\lambda I)^{1/2}}\right)}\Bigg]\Bigg) \leqslant \delta.$$

Theorem 3.5 (Kernel Least-squares) *Assume that N is a stopping time adapted to the filtration of the past. Then in the sub-Gaussian streaming regression model, for each $x \in \mathscr{X}$ it holds*

$$\mathbb{P}\Bigg(|f^\star(x) - \mu_N(x)| \geqslant |f^\star(x) - k_N(x)^\top(\mathbf{K}_N + \sigma^2 I_N)^{-1}\mathbf{f}_N| + R\sqrt{2\Big(||k_N(x)||^2_{(\mathbf{K}_N+\sigma^2 I_N)^{-2}} + 1\Big)\ln\left(\frac{\sqrt{1+||k_N(x)||^2_{(\mathbf{K}_N+\sigma^2 I_N)^{-2}}}}{\delta}\right)}\Bigg) \leqslant \delta.$$

Remark 3.6 *The proofs of the three results rely on an application of the Laplace method from (Peña, Lai and Shao, 2008), together with a non trivial peeling argument that enables to localize the eigenvalues of the matrices $G_N, G_{N,\lambda}, \mathbf{K}_N$. We detail the full technique in the dedicated Section 5.*

The result of Theorem 3.5 is not simultaneous over all $x \in \mathscr{X}$, contrary to the previous two results. The reason is that in this case, we apply the Laplace method directly to the 1×1 term $E_N^\top(\mathbf{K}_N + \sigma^2 I_N)^{-1}k_N(x)$ (where E_N is the noise vector) that has a deterministic dimension. It would be more natural to apply the method to $E_N^\top(\mathbf{K}_N + \sigma^2 I_N)^{-1}$, but the dimension, N, of this vector is a random stopping time and the method may no

longer apply in this case. Now, when N is deterministic, the result of Theorem 3.5 holds simultaneously over all $x \in \mathscr{X}$.

The quantity $|f^\star(x) - k_N(x)^\top (\mathbf{K}_N + \sigma^2 I_N)^{-1} \mathbf{f}_N|$ of Theorem 3.5 is a direct analogue of the quantity $\frac{\lambda}{\sqrt{\lambda_{\min}(G_{N,\lambda})}} ||\theta^\star||_2$ of Theorem 3.4, and depends on the smoothness of the function $f^\star$ in the RKHS.

Discussion The result of Theorem 3.4 is essentially the same as that of (Abbasi-Yadkori, Pal and Szepesvari, 2011), only with a minor refinement that can be also done in their case. Using a regularization is important in case the considered function space is large with respect to the number of observations (in the sense that G_N is not invertible). For smaller function spaces or large enough N, one may want however to stop regularizing and put λ to 0. Despite being a great result, Theorem 3.4 is in this sense unsatisfactory with respect to the dependency with the regularization λ, since the bound becomes trivial as $\lambda \to 0$.

The result of Theorem 3.3 is novel and interesting in this respect. Indeed, compared to the regularized approach, we managed to get rid of the regularization parameter λ, that is intuitively not required when G_N is invertible. The result also improves on the approach used in (Rusmevichientong and Tsitsiklis, 2010), that was the best result so far in this case. From their work, it is possible, after some careful rewriting of their proof (simply gluing the different terms they control separately), to obtain the following result:

Theorem 3.7 (Least-squares reconstruction for finite-dimension spaces) *Let us consider the sub-Gaussian streaming regression model. Assume that G_N is $d \times d$ invertible with smallest eigenvalue $\lambda_{\min}(G_N) \geqslant \lambda_0 > 0$ bounded away from 0 and that $\forall n, ||\varphi(x_n)|| \leqslant R_{\mathscr{X}}$ for a deterministic $R_{\mathscr{X}}$. Then, if $N \geqslant \frac{\lambda_0}{12 R_{\mathscr{X}}^2}$ it holds for all $\delta \in [0,1]$ and $x \in \mathscr{X}$, with probability higher than $1-\delta$,*

$$|f^\star(x) - f_{\theta^\star}(x)| \leqslant R ||\varphi(x)||_{G_N^{-1}} \sqrt{\tilde{D}_N(\delta)}\,,$$

where we introduced the quantity

$$\tilde{D}_N(\delta) = 16 \left[1 + \ln\left(1 + \frac{36 R_{\mathscr{X}}^2}{\lambda_0}\right)\right] \left[d \ln\left(\frac{36 R_{\mathscr{X}}^2}{\lambda_0} N\right) + \ln(1/\delta)\right] \ln(N)\,.$$

The quantity $\tilde{D}_N(\delta)$ obtained in their result should be compared with $\ln\left(\frac{\kappa_d (e^2 \Lambda_N^2)}{\delta}\right)$ from Theorem 3.3, where we managed to move the $O(\ln(N))$ factor inside the ln term and provide a result that is valid for any random stopping time N, without restricting N to be large enough. Also, we do not require a deterministic bound $R_{\mathscr{X}}$ but can work directly with the largest eigenvalue Λ_N computed from the data. The gap between the two bounds can actually be significant as we illustrate in Figure 1 on a toy problem.

The result of Theorem 3.5 regarding the kernel regression setting is also novel and directly follows from our proof technique and intermediate results for the two other theorems.

3.2. Variance estimates

We now turn to the more involved problem of estimating the unknown noise level R, that is generally not known in practice. We consider for this purpose two approaches. The first approach simply consists in further specifying the noise model and studying a standard empirical variance estimate. The second approach consists in making explicit use of the results derived in Theorems 3.3, 3.4, 3.5 and taking advantage of our streaming setting in order to provide a "corrective" estimate of the variance that adjusts its value at each time step when a mismatch is detected. We first proceed with the variance estimation procedure.

3.2.1. Estimation

The next theorems provide a control on the accuracy of the following (slightly biased) estimates

$$\widehat{\sigma}^2_N = \frac{1}{N}\sum_{n=1}^{N}(y_n - \langle \theta^\dagger_N, \varphi(x_n)\rangle)^2 \quad \text{and} \quad \widehat{\sigma}^2_{N,\lambda} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \langle \theta_{N,\lambda}, \varphi(x_n)\rangle)^2 \,.$$

Assumption 1 (Strongly sub-Gaussian noise) *The noise sequence is not-only conditionally sub-Gaussian but further strongly sub-Gaussian, in the sense that*

$$\forall i, \forall \lambda < \frac{1}{2R^2} \quad \ln \mathbb{E}\left[\exp(\lambda \xi_i^2)\middle| h_{i-1}\right] \leqslant -\frac{1}{2}\ln\left(1 - 2\lambda R^2\right).$$

This assumption naturally holds for Gaussian variables, which explains the name.

Theorem 3.8 (Ordinary variance estimate) *Under Assumption 1, in the sub-Gaussian streaming regression model, then for any random stopping time N adapted to the filtration of the past, with probability higher than $1-3\delta$, that either $\lambda_{\min}(G_N) < \lambda_0$ or*

$$R\left(1 - \sqrt{\frac{C_N(\delta)}{N}} - \sqrt{\frac{C_N(\delta)}{N} + \frac{D_N(\delta)}{N}}\right) \leqslant \sqrt{\widehat{\sigma}^2_N} \leqslant R\left(1 + \sqrt{\frac{2C_N(\delta)}{N}}\right).$$

where $D_N(\delta) = 4\ln(\kappa_d(e\Lambda_N/\lambda_0)/\delta)$ and $C_N(\delta) = \ln(e/\delta)\big[1 + \ln(\pi^2\ln(N)/6)/\ln(1/\delta)\big]$.

Theorem 3.9 (Regularized variance estimate) *Under Assumption 1, in the sub-Gaussian streaming regression model, for any random stopping time N for the filtration of the past, with probability higher than $1-3\delta$, it holds*

$$\begin{aligned}
\sqrt{\widehat{\sigma}^2_{N,\lambda}} &\leqslant R\left[1 + \sqrt{\frac{2C_N(\delta)}{N}}\right] + ||\theta^\star||_2\sqrt{\frac{\lambda}{N}}\sqrt{1 - \frac{\lambda}{\lambda_{\max}(G_{N,\lambda})}} \\
\sqrt{\widehat{\sigma}^2_{N,\lambda}} &\geqslant R\left[1 - \sqrt{\frac{C_N(\delta)}{N}} - \sqrt{\frac{C_N(\delta) + 2D_{N,\lambda}(\delta)}{N}}\right] - \sqrt{\frac{2\lambda^{1/2}R||\theta^\star||_2\sqrt{D_{N,\lambda}(\delta)}}{N}}\,.
\end{aligned}$$

where $D_{N,\lambda}(\delta) = 2\ln\left(\frac{det(G_N+\lambda I)^{1/2}}{\delta\lambda^{d/2}}\right)$.

Remark 3.10 *We observe from these two theorems that the lower bound on the noise is a little different from the upper bound. Also, the estimate is potentially better when one does not need to regularize, since in this case we do not need to bound the unknown parameter* $||\theta_\star||_2$.

Remark 3.11 *Since* λ *can be thought as a proxy for* R^2*, it may be tempting to use the variance estimate in order to tune the parameter* λ*. However, this procedure does not straightforwardly leads to theoretical guarantees because all our proof techniques are based on version of the Laplace method that requires* λ *to be independent from the observations.*

Corollary 3.12 (Bounds on the variance) *Under Assumption 1, in the sub-Gaussian streaming regression model, for any random stopping time* N *for the filtration of the past, with probability higher than* $1-3\delta$*, it holds that either* $\lambda_{\min}(G_N) < \lambda_0$ *or*

$$\sqrt{\widehat{\sigma}^2_N}\left(1+\sqrt{\frac{2C_N(\delta)}{N}}\right)^{-1} \leqslant R \leqslant \sqrt{\widehat{\sigma}^2_N}\left(1-\sqrt{\frac{C_N(\delta)}{N}}-\sqrt{\frac{C_N(\delta)}{N}+\frac{D_N(\delta)}{N}}\right)_+^{-1},$$

where $(x)_+ = \max(x,0)$*, and using the convention that* $(x)_+^{-1} = 0$ *if* $x \leqslant 0$*. Now, if an upper bound* $R^+ \geqslant R$ *is known to the learner, one can derive the following inequalities that hold with probability higher than* $1-3\delta$

$$\sqrt{\widehat{\sigma}^2_N} - R^+\sqrt{\frac{2C_N(\delta)}{N}} \leqslant R \leqslant \sqrt{\widehat{\sigma}^2_N} + R^+\left(\sqrt{\frac{C_N(\delta)}{N}}+\sqrt{\frac{C_N(\delta)+D_N(\delta)}{N}}\right).$$

Likewise, for the regularized estimate, for any $\lambda > 0$*, with probability higher than* $1-3\delta$*, it holds that*

$$\begin{aligned} R &\leqslant \frac{1}{\alpha^2}\left(\left[\sqrt{\widehat{\sigma}^2_{N,\lambda}}\alpha + \frac{\lambda^2||\theta^\star||_2\sqrt{D_{N,\lambda}(\delta)}}{2N\lambda_{\min^{3/2}(G_{N,\lambda})}}\right]^{1/2} + \left[\frac{\lambda^2||\theta^\star||_2\sqrt{D_{N,\lambda}(\delta)}}{2N\lambda_{\min^{3/2}(G_{N,\lambda})}}\right]^{1/2}\right)^2 \\ R &\geqslant \left[\sqrt{\widehat{\sigma}^2_{N,\lambda}} - ||\theta^\star||_2\sqrt{\frac{\lambda}{N}\left(1-\frac{\lambda}{\lambda_{\max}(G_{N,\lambda})}\right)}\right]\left(1+\sqrt{\frac{2C_N(\delta)}{N}}\right)^{-1}, \end{aligned}$$

where we introduced

$$\alpha = \left(1-\sqrt{\frac{C_N(\delta)}{N}}-\sqrt{\frac{C_N(\delta)+D_{N,\lambda}(\delta)\left(1+\frac{\lambda}{\lambda_{\min(G_{N,\lambda})}}\right)}{N}}\right)_+.$$

Further, if an upper bound $R^+ \geqslant R$ *is known to the learner, one can derive the following inequalities that hold with probability higher than* $1-3\delta$*,*

$$\begin{aligned} R &\leqslant \sqrt{\widehat{\sigma}^2_{N,\lambda}} + R^+\left(\sqrt{\frac{C_N(\delta)}{N}}+\sqrt{\frac{C_N(\delta)+2D_{N,\lambda}(\delta)}{N}}\right) + \sqrt{\frac{2\lambda^{1/2}R^+||\theta^\star||_2\sqrt{D_{N,\lambda}(\delta)}}{N}} \\ R &\geqslant \sqrt{\widehat{\sigma}^2_{N,\lambda}} - R^+\sqrt{\frac{2C_N(\delta)}{N}} - ||\theta^\star||_2\sqrt{\frac{\lambda}{N}}\sqrt{1-\frac{\lambda}{\lambda_{\max}(G_{N,\lambda})}}. \end{aligned}$$

3.2.2. Correction

The standard estimates of the variance presented in the previous section have some drawbacks. The main one is the behavior in case of over-fitting, since for instance the variance of the ordinary least-squares estimate becomes 0 in this case, and is thus completely non informative about the true variance. For this reason, it is desirable to consider a different approach that remains valid in the case of over-fitting (and that requires only $o(n)$ updating complexity at step n). We design such an estimate, in a very natural way, taking advantage of the confidence bounds on the mean and of the sequential setting.

At high level, the *corrective* procedure that we introduce compares the high probability confidence intervals and the actual observations, and searches for the smallest variance parameter that makes all past and current confidence intervals valid (that is, the confidence interval contains the observed value). This procedure guarantees that the resulting corrected variance estimate is indeed a high probability lower bound on the actual variance as we show below.

Lemma 3.13 (Corrective variance estimates) *Let us consider the sub-Gaussian streaming regression model and define the following instantaneous and cumulative variance estimates*

$$\widehat{r}_{N+1}(\delta) = \frac{\max(|y_{N+1} - f_{\widehat{\theta}_N}(x_{N+1})| - c_N, 0)}{b_N(\delta/3) + \sqrt{2\ln(3/\delta)}}\,, \qquad \widehat{R}_{N+1}(\delta) = \max_{n\in[N]} \widehat{r}_{n+1}(6\delta/(\pi n)^2)\,.$$

where the function b and constant c are such that

$$\mathbb{P}\left(\exists x \in \mathscr{X} : |f^\star(x) - f_{\widehat{\theta}_N}(x)| \geqslant c_N + Rb_N(\delta)\right) \leqslant \delta\,.$$

Then, under assumption 1, $\widehat{r}_{N+1}(\delta)$ satisfies with probability higher than $1-\delta$ the inequality $\widehat{r}_{N+1}(\delta) \leqslant R$. Likewise, $\widehat{R}_{N+1}(\delta) \leqslant R$ with probability higher than $1-\delta$.

One advantage of these quantities is to remain meaningful even when there is over-fitting, contrary to the classical variance estimate that become 0. The splitting of the data into training and testing is done in a very simple (perhaps naive) way by testing only on the next observation, which avoids the combinatorial complexity of combining all possible 2-set partitions of the data.

Note that Assumption 1 is not required in order to lower bound the value R. This assumption (or a related one) is required in order to show the estimate is tight. This is intuitive, since otherwise nothing requires the R value of the sub-Gaussian assumption to be minimal.

Proof of Lemma 3.13:

Indeed, it suffices to remark that by construction of the confidence interval, it holds with probability higher than $1-\delta$, for all x,

$$|f^\star(x) - f_{\widehat{\theta}_N}(x)| \leqslant c_N + Rb_N(\delta)\,.$$

Thus, under a R-sub-Gaussian assumption, the observation $y_N = y(x)$ at point $x = x_N$ is such that $y(x) - f^\star(x)$ is R-sub-Gaussian and thus with probability higher than $1-3\delta$, by a triangular inequality followed by a union bound it must hold

$$|y(x) - f_{\theta_{N,\lambda}}(x)| \leqslant c_N + Rb_N(\delta) + R\sqrt{2\ln(1/\delta)}\,.$$

We simply conclude by noting that $\widehat{r}_{N+1}(\delta)$ is actually the smallest value R such that the high probability bound is satisfied with probability $1-\delta$

$$\begin{aligned}\widehat{r}_{N+1}(\delta) &= \min\{R : |y_{N+1} - f_{\widehat{\theta}_N}(x_{N+1})| \leqslant c_N + Rb_N(\delta/3) + R\sqrt{2\ln(3/\delta)}\} \\ &= \frac{\max(|y_{N+1} - f_{\widehat{\theta}_N}(x_{N+1})| - c_N, 0)}{b_N(\delta/3) + \sqrt{2\ln(3/\delta)}}\,.\end{aligned}$$

A natural way to define the corrective estimate is then to take the maximum $\widehat{r}_n(\delta)$ at each time step. Let $\delta_n = 6\delta/(\pi n)^2$. Using a simple union bound, we define

$$\begin{aligned}\widehat{R}_{N+1}(\delta) &= \min\{R : \forall n \in [N], |y_{n+1} - f_{\widehat{\theta}_n}(x_{n+1})| \leqslant c_n + Rb_n(\delta_n/3) + R\sqrt{2\ln(3/\delta_n)}\} \\ &= \max_{n\in[N]} \frac{\max(|y_{n+1} - f_{\widehat{\theta}_n}(x_{n+1})| - c_n, 0)}{b_n(\delta_n/3) + \sqrt{2\ln(3/\delta_n)}} = \max_{n\in[N]} \widehat{r}_{n+1}(\delta_n)\,.\end{aligned}$$

It satisfies with probability higher than $1-\delta$ the inequality $\widehat{R}_{N+1}(\delta) \leqslant R$. □

The previous lemma shows that it is possible to build a high probability lower bound on the noise level R. It is interesting to investigate to which extent the corrective estimate is loose. Due to the dependency between the random variables, this is in general a challenging question. We provide below a hint on a quantity related to $\widehat{R}_{N+1}(\delta)$ in the simplified situation when both x_{n+1} and ξ_{n+1} are independent on the past history h_n, N is deterministic and when ξ_{n+1} is moreover exactly Gaussian. We also restrict our illustration to the ordinary least-squares estimate $\widehat{\theta}_N$.

In this case, we first remark that by construction of the ordinary least-squares estimate it comes

$$\begin{aligned}\max_{n\in[N]} |y(x_{n+1}) - f_{\widehat{\theta}_n}(x_{n+1})| &= \max_{n\in[N]} |(\theta^\star - \widehat{\theta}_n)^\top \varphi(x_{n+1}) + \xi_{n+1}| \\ &= \max_{n\in[N]} \left| E_n^\top \mathbf{\Phi}_n G_n^{-1} \varphi(x_{n+1}) + \xi_{n+1} \right| \\ &= \max_{n\in[N]} \left| \sum_{m=1}^{n} \underbrace{\varphi(x_m)^\top G_n^{-1} \varphi(x_{n+1})}_{\lambda_{n,m}} \xi_m + \xi_{n+1} \right|.\end{aligned}$$

In this form, we see that one needs to control the maximum of the absolute value of a sum of Gaussian random variables. Indeed, using the assumption that each x_n is independent on the past observations before n, then so is the matrix G_n and thus $\lambda_{n,m}$.

Now, using the assumption that ξ_n is independent on the past, and thus of ξ_m, $m < n$, we remark that $\sum_{m=1}^n \lambda_{n,m}\xi_m + \xi_{n+1}$ is a centered Gaussian with standard deviation $\tilde{R} = R(\varphi(x_{n+1})^\top G_n^{-1}\varphi(x_{n+1}) + 1)$. Thus, we deduce that the following quantity

$$Z_N = \max_{n\in[N]} \frac{|y(x_{n+1}) - f_{\theta_n}(x_{n+1})|}{\sqrt{||\varphi(x_{n+1})||^2_{G_n^{-1}} + 1}} = \max_{n\in[N]} |g_n|\,,$$

is the maximum of the absolute value of N many R-Gaussian random variables $\{g_n\}_{n\in[N]}$.

At this point, let us remind that in the case when $\tilde{Z}_K = \max_{n\in[K]} \tilde{g}_n$ is the maximum of K *i.i.d.* centered *Gaussian* variables $\{\tilde{g}_n\}_{n\in[K]}$ with variance R^2, then the following holds

$$R\sqrt{\frac{2\ln(K)}{2\pi\ln(2)}} \leqslant \mathbb{E}[\tilde{Z}_K] \leqslant R\sqrt{2\ln(K)}\,.$$

We further know that $\tilde{Z}_K$ is asymptotically of order $R\sqrt{2\ln(K)}$ up to fluctuations of order $O(R\ln\ln(K)/\sqrt{\ln(K)})$. Indeed, it holds for these independent variables (see e.g. (Leadbetter, Lindgren and Rootzén, 2012, Theorem 1.5.3)) that

$$\mathbb{P}\left(\frac{\max_{n\in[K]}\tilde{g}_n}{\sqrt{2\ln(K)}} - R \leqslant -\frac{R}{4\ln(K)}\left[\ln\ln K + \ln(4\pi) + 2\ln\ln(1/\delta)\right]\right) \overset{K\to\infty}{\to} \delta\,. \tag{4}$$

The proof of this result actually follows from a simple study of the cdf of the Gaussian distribution. This control, even though it is asymptotic, gives a hint about the behavior of $\tilde{Z}_K$, and shows that with high probability, $\tilde{Z}_K/\sqrt{2\ln(K)}$ cannot be a too loose lower-bound on the level of noise R.

However, even in our simplified case, we need to control Z_N that is the maximum of $2N$ many correlated R-Gaussian random variables (since $|g_n| = \max\{g_n, -g_n\}$, and because g_n depends on g_{n-1}). By the Slepian inequality, we can relate Z_N to $\tilde{Z}_{2N}$: Indeed, for all (deterministic) z, the inequality $\mathbb{P}(Z_N \geqslant z) \leqslant \mathbb{P}(\tilde{Z}_{2N} \geqslant z)$ holds true. On the other hand, by a simple union bound over the K random variables, it holds

$$\mathbb{P}\left(\tilde{Z}_K \geqslant R\sqrt{2\ln(1/\delta)}\right) \leqslant K\delta\,.$$

Thus, for $K = 2N$, we deduce that with probability higher than $1-\delta$, then

$$Z_N \leqslant R\sqrt{2\ln(2N/\delta)})\,.$$

However, this only provides an *upper* bound on the quantity Z_K. Slepian's inequality does not able to control the reverse inequality that would be required in order to understand how loose is the *lower* bound estimate.

We leave as an (partially) open question the generalization of (4) to the case when the $\{\tilde{g}_n\}_{n\in[N]}$ are the partial sums of i.i.d Gaussian variables $\{\tilde{\zeta}_n\}_{n\in[N]}$ (that is $\tilde{g}_n = \sum_{i=1}^n \zeta_i$)

as well as that of getting a fully non-asymptotic result valid for all N (possibly a random stopping time). Such a result would complement that of Lemma 3.13 in order to derive a confidence interval for corrective estimate of the form $R - c_{N+1} \leqslant \widehat{R}_{N+1}(\delta) \leqslant R.$ for some $c_{N+1} \overset{N\to\infty}{\to} 0$ yet to be defined.

3.3. Streaming updates and complexity

In the case of streaming data, recomputing the ordinary least-squares solution from scratch at each step may not be computationally efficient, due to the linear scaling with the length of the history. Instead, it is desirable to use the solution computed at time n in order to help computing that at time $n+1$. Such incremental updates have been studied in the literature. We provide for the sake of completeness the following folklore result that is a direct application of the Shermann-Morrison formula:

Lemma 3.14 (Online updates to the least-squares solution) *The solution $\theta_{n+1,\lambda}$ can be computed at each time step, by maintaining only a $d \times d$ matrix $G_{n,\lambda}^{-1}$, and the vector $\Phi_n^\top Y_n$ with updates formulas given by*

$$\begin{aligned} G_{n+1,\lambda}^{-1} &= G_{n,\lambda}^{-1} - \frac{G_{n,\lambda}^{-1}\varphi(x_{n+1})\varphi(x_{n+1})^\top G_{n,\lambda}^{-1}}{1+\varphi(x_{n+1})^\top G_{n,\lambda}^{-1}\varphi(x_{n+1})} \qquad \textit{with } G_0^{-1} = \lambda^{-1} I_{d\times d} \\ \Phi_{n+1}^\top Y_{n+1} &= \Phi_n^\top Y_n + \varphi(X_{n+1}) y_{n+1} \qquad \textit{with } \Phi_0^\top Y_0 = 0_d \,. \end{aligned}$$

provided that $1+\varphi(x_{n+1})^\top G_{n,\lambda}^{-1}\varphi(x_{n+1}) \neq 0\,.$

A similar but more complicated update formula exists for the pseudo-inverse matrix $G_n^\dagger$. It can be found for instance in (Campbell and Meyer, 2009, Th.3.1.3, p.47), and we refer the reader to this reference for further details. Thus, both $\theta_N^\dagger$ and $\theta_{n,\lambda}$ can be computed efficiently, with an update complexity that is at most $O(d^2)$ at each step n.

The variance estimates can also be updated sequentially in an efficient way. Indeed, it suffices to rewrite the variance estimate as

$$\begin{aligned} \widehat{\sigma}_{N,\lambda}^2 &= \frac{1}{N}\sum_{n=1}^N (y_n - \langle \theta_{N,\lambda}, \varphi(x_n)\rangle)^2 \\ &= \underbrace{\frac{1}{N}\sum_{n=1}^N y_n^2}_{q_N} + \frac{1}{N}\theta_{N,\lambda}{}^\top \underbrace{\Big(\sum_{n=1}^N \varphi(x_n)\varphi(x_n)^\top\Big)}_{G_N} \widehat{\theta}_{N,\lambda} - \frac{2}{N}\theta_{N,\lambda}{}^\top \underbrace{\sum_{n=1}^N \varphi(x_n)\, y_n}_{s_N} . \end{aligned}$$

Since q_N is a scalar, $\theta_N^\dagger$ is a vector of dimension d and G_N is a $d\times d$ matrix, they update in a constant time with respect to n. A similar decomposition holds for $\widehat{\sigma}_N^2$.

The computation of the kernel estimates is trickier, since one must update the kernel matrix, that is $N \times N$ (this should be compared with the $d \times d$ feature matrix G_N). Thus, even a rank 1 update requires $O(N)$ steps in general for these infinite dimension

models, and can thus be costly for large N. One may want to resort to approximations such as sketching in order to avoid a numerical complexity growing linearly with N, but one needs to control the approximation error introduced in that case.

3.4. Some numerical illustration

In this section, we provide a few numerical experiments that enable to visualize the empirical confidence intervals that we build, and compare them to other constructions.

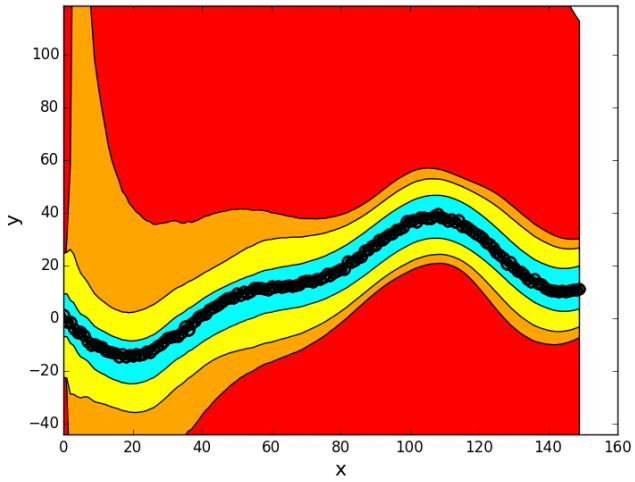

Figure 1. Confidence intervals in the realizable case, with given bound on the noise level $R = 3.$, for various methods. Red: Ordinary least-squares estimate from (Rusmevichientong and Tsitsiklis, 2010) (Theorem 3.7) Orange: Ordinary least-squares estimate using the improved Theorem 3.3. Yellow: Regularized least-squares estimate with $\lambda = 1$, using Theorem 3.4. Cyan: High-confidence interval corresponding to a Gaussian $\mathcal{N}(f_{\widehat{\theta}_n}(\cdot), R)$.

In Figure 1, we have plotted the confidence intervals over the observations y_n built from various methods, for some simple signal in a linear function space of low dimension d, with noise level less than $R = 3$. In this first series of plots, R is given to the learners.

For the ordinary least-squares, we have plotted the confidence interval[4] resulting from (Rusmevichientong and Tsitsiklis, 2010) (in Red), and the one resulting from our improvement (in Orange), see Theorem 3.3. For the regularized estimate, we plotted the bound from Theorem 3.4 (in Yellow, see also (Abbasi-Yadkori, Pal and Szepesvari, 2011)), with parameter $\lambda = 1$ and $||\theta^\star|| \leqslant \sqrt{d}$. Note that large values of λ (compared to the actual level of noise) may lead to loose confidence bound. Finally, we plotted in cyan the

[4]They are only valid after the number of observations is large enough so that the G_n matrix has large enough lowest eigenvalue.

high probability confidence interval of a Gaussian centered at the estimated value $f_{\widehat{\theta}_n}$, with variance R^2, that is $[f_{\widehat{\theta}_n}(x) \pm R\sqrt{2\log(1/\delta)}]$. This naive interval (we refer to it as the in-model confidence interval) is however not in general a valid high-probability confidence interval as it does not take into account the estimation error of $\widehat{\theta}_n$.

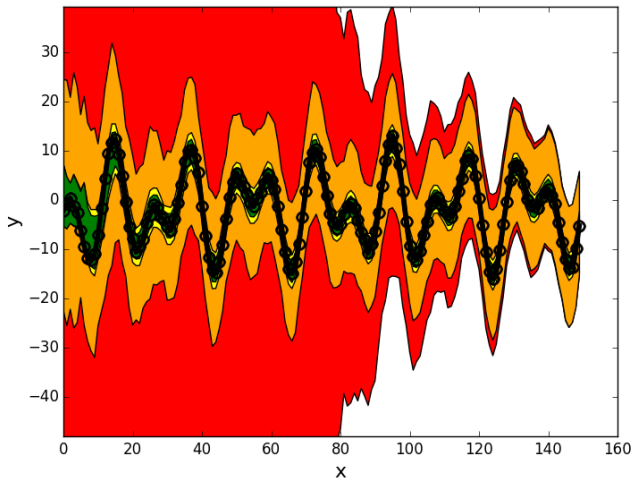

Figure 2. Confidence intervals from Theorem 3.4, in the realizable case, for the regularized least-squares estimate with $\lambda = 1$ and various methods to handle the noise. Red: Upper-bound from Corollary 3.12 without any knowledge of the noise. Orange: Bound R given to the learner. Yellow: Lower-bound from Corollary 3.12 without any knowledge of the noise. Green: Corrective noise estimate.

In Figure 2, we consider the confidence intervals computed from Theorem 3.4, and study the influence of the estimation of the noise parameter R on the resulting bounds. We plot in Orange the confidence interval when a bound on R is given (Here $R = 3$, and the actual noise level has been chosen uniformly in $[0, 3]$). We plot in Red (respectively Yellow) the intervals using for R the upper bound (respectively lower bound) from Corollary 3.12, when no bound on R is known. Despite the fat the noise level is completely unknown, the confidence interval shrinks reasonably fast. Finally, the Green intervals correspond to using the corrective noise estimate from Lemma 3.13 for the parameter R. Similar figures can be obtained when studying the estimation of R in the context of *ordinary* rather than regularized least-squares estimates.

Remark 3.15 *To avoid clutter, we haven't plot the interval using a given bound R^+, but it should be in between the orange and yellow intervals. Note also that the yellow intervals in Figure 2 are not valid high probability confidence intervals in general. Both the yellow and green intervals can be qualified as being "optimistic".*

4. Confidence, Model adequacy and Predictive loss

In this section, we provide some key applications of the confidence intervals built in the previous section. On the one hand, we focus on the notion of *model adequacy* that enables to measure empirically the adequacy of the modeling assumptions to capture the observed signal, and thus to raise an alert when the adequacy is too low. On the other hand, we provide an estimate of the *loss* of the learner at a next observation step, for various notions of losses. This is especially relevant to the task of model selection or aggregation in the context of sequential prediction.

In the following, we abstract the previous models, for the sake of generality. Let $\mathscr{P}$ be a set of stochastic processes indexed by $\mathscr{X}$, with value in $\mathscr{Y}$ (that is for $P \in \mathscr{P}$, and $x \in \mathscr{X}$, $y \sim P(x)$ belongs to $\mathscr{Y}$). We call a prediction strategy $\mathcal{P}$ over $\mathscr{P}$ a learning algorithm that outputs, for a training history h, one specific process denoted $P(h) \in \mathscr{P}$.

Definition 4.1 (Confidence set) *A confidence set at point $x \in \mathscr{X}$ for the prediction strategy $\mathcal{P}$ over $\mathscr{P}$ trained on history h, is a subset of $\mathscr{Y}$ indexed by the confidence level $\delta \in [0,1]$, such that for any $P \in \mathscr{P}$ such that h has been generated by P (in the sense that $y \sim P(x)$ for each $(x,y) \in h$), then*

$$\forall \delta \in [0,1], \quad \mathbb{P}_{Y \sim P(x)}\Big(Y \in \mathscr{C}(x, \delta, \mathcal{P}(h))\Big) \geqslant 1 - \delta \,.$$

Likewise, for a given history h, the confidence distribution set at point $x \in \mathscr{X}$ is

$$\mathscr{D}(x, h, \mathcal{P}) = \left\{ P \in \mathscr{P} : \forall \delta \in [0,1], \quad \mathbb{P}_{Y \sim P(x)}\Big(Y \in \mathscr{C}(x, \delta, \mathcal{P}(h))\Big) \geqslant 1 - \delta \right\}.$$

Remark 4.2 *It is relevant to illustrate the difference between the confidence set that is considered here and a perhaps more naive one. Consider a Gaussian model where $y \sim \mathcal{N}(f^\star(x), \sigma^2)$ with known σ^2. Once a regression parameter θ_n is computed, it is tempting to consider the distribution $\mathcal{N}(f_{\theta_n}(x), \sigma^2)$, and corresponding interval $\tilde{\mathscr{C}}(x,\delta) = [f_{\theta_n}(x) \pm \sigma\sqrt{2\log(1/\delta)}]$. This "in-model" confidence interval (and distribution) is quite different from that of Definition 4.1, since the latter takes into account the uncertainty about the random variable θ_n, contrary to $\tilde{\mathscr{C}}(x,\delta)$.*

Example: From Theorem 3.3, we deduce that for the ordinary least-squares solution with $\widehat{f}_h = f_{\widehat{\theta}_n}$, it holds (when R is known)

$$\mathscr{C}(x, \delta, \mathcal{P}(h)) = \{y : |y - \widehat{f}_h(x)| \leqslant b_h + c_h\sqrt{2\ln(1/\delta)}\}\,,$$

with $b_h = 2R||\varphi(x)||_{G_N^\dagger}\sqrt{\ln\Big(3\kappa_d(e^2\Lambda_N^2)\Big)} + \sqrt{2\ln(3)}$ and $c_h = \sqrt{2}R||\varphi(x)||_{G_N^\dagger} + 1$. We get a similar form for the regularized least-squares solution from Theorem 3.4, respectively with $b_h = ||\varphi(x)||_{G_{N,\lambda}^{-1}}\left[\frac{\lambda}{\sqrt{\lambda_{\min}(G_{N,\lambda})}}||\theta^\star||_2 + R\sqrt{2\ln\left(\frac{3det(G_N+\lambda I)^{1/2}}{det(\lambda I)^{1/2}}\right)}\right] + \sqrt{2\ln(3)}$ and $c_h = R||\varphi(x)||_{G_{N,\lambda}^{-1}} + 1$.

Definition 4.3 (Model Adequacy) *We define the model adequacy for an observation y at point x of the prediction strategy $\mathcal{P}$ trained on history h by*

$$\alpha(y,x,\mathcal{P}(h)) = \sup\{\delta : \delta \in \mathrm{Argmin}\{\inf\{\ell_{\mathscr{Y}}(y,y') : y' \in \mathscr{C}(x,\delta,\mathcal{P}(h))\}, \delta \in [0,1]\}\,.$$

A value of 1 indicates no mismatch, while a value of 0 indicates a strong mismatch.

Examples: In the *Markov* setting on a finite alphabet, the model adequacy simply coincides with the probability assigned to each observation y the prediction strategy. We show see this Section 4.1 below. In the sub-Gaussian streaming regression model, it takes another form as we show in Section 4.1; For a typical confidence set given by

$$\mathscr{C}(x,\delta,\mathcal{P}(h)) = \{y : |y - \widehat{f}_h(x)| \leqslant b_h + c_h\sqrt{2\ln(1/\delta)}\}\,,$$

and for the quadratic loss $\ell_{\mathscr{Y}}(y,y') = (y-y')^2$, we show in Section 4.1 that the model adequacy has the following form

$$\alpha(y,x,\mathcal{P}(h)) = \exp\left(-\frac{(\widehat{f}_h(x) - b_h - y)_+^2}{2c_h^2} - \frac{(y - \widehat{f}_h(x) - b_h)_+^2}{2c_h^2}\right),$$

where $(z)_+ = \max\{z,0\}$ denotes the positive part of z. Note that $\alpha(y,x,\mathcal{P}(h)) = 1$ for all y such that $|y - \widehat{f}_h(x)| \leqslant b_h$. Further, the model adequacy integrates to $\int_{\mathscr{Y}} \alpha(y,x,\mathcal{P}(h))dy = \sqrt{2\pi c_h^2} + 2b_h$, so that the normalized model adequacy coincides with the density of a quasi-Gaussian distribution that is sometimes called a *possibilist distribution* due to the fact it not only corresponds to a model estimate (such that $\mathscr{N}(f_{\widehat{\theta}_N}(x),\sigma^2)$), but also accounts for the uncertainty about $\widehat{\theta}_N$, thus effectively combining the distributions $\mathscr{N}(f_\theta(x),\sigma^2)$ for all plausible parameters θ.

The next three definitions introduce the predictive loss, that is an estimate of the high-probable loss that one may incur at a (next) observation point x. Thus, the predictive loss is typically used, at step n with history h_n, to predict the possible loss at point $x = x_{n+1}$, assuming the observed signal obeys our modeling assumptions.

Definition 4.4 (Predictive Quadratic Loss) *The predictive quadratic loss at point $x \in \mathscr{X}$ for the prediction strategy $\mathcal{P}$ trained on history h, with level $\delta_0 \in [0,1]$ is defined by*

$$\overline{\ell}_{\mathscr{Y}}(x,\delta_0,\mathcal{P}(h)) = \sup\{\ell_{\mathscr{Y}}(y,y'), y, y' \text{ such that } \alpha(y,x,\mathcal{P}(h)) \geqslant \delta_0\}\,.$$

Definition 4.5 (Predictive Self-information Loss) *The predictive self-information loss at point $x \in \mathscr{X}$ for the prediction strategy $\mathcal{P}$ trained on history h, with level $\delta_0 \in [0,1]$ is defined by*

$$\overline{\ell}_I(x,\delta_0,\mathcal{P}(h)) = \sup\{\ell_I(y,\rho(\cdot|x,\mathcal{P}(h)), y \text{ such that } \alpha(y,x,\mathcal{P}(h)) \geqslant \delta_0\}\,.$$

Definition 4.6 (Predictive Transportation Loss) *The predictive transportation loss at point $x \in \mathscr{X}$ for the prediction strategy $\mathcal{P}$ trained on history h, with level $\delta_0 \in [0,1]$ is defined by*

$$\begin{aligned}\overline{\mathscr{T}}(x, \delta_0, \mathcal{P}(h)) &= \sup\{\mathscr{T}_{\ell_{\mathscr{Y}}}(P(x), \rho(\cdot|x, \mathcal{P}(h)), P \in \mathcal{P} \\ &\qquad \textit{such that } \mathbb{P}_{Y\sim P(x)}[\alpha(Y, x, \mathcal{P}(h)) < \delta_0] \leqslant \delta_0\}\,.\end{aligned}$$

4.1. Model adequacy

In this section, we provide explicit computations for the model adequacy introduced in Definition 4.3.

Example 1 When $\mathscr{Y}$ is discrete and $\ell(y, y') = \mathbb{I}\{y \neq y'\}$, let $p(\cdot|x, h, \mathscr{M})$ be the prediction distribution on the value of the observation at point x for model $\mathscr{M}$ with history h, it is natural to define the corresponding confidence set

$$\mathscr{C}(x, \delta, h, \mathscr{M}) = \bigcup\{y_{k_1}, \dots, y_{k_K} \in \mathscr{Y} : \sum_{i=1}^{K} p(y_{k_i}|x, h, \mathscr{M}) \geqslant 1-\delta \text{ and } K \text{ is minimal}\}\,.$$

Now, for ℓ being the grossiere metric, it can be checked that the model adequacy satisfies $\alpha(Y; x, h, \mathscr{M}) = p(Y|x, h, \mathscr{M})$, and thus

$$\tilde{p}(y|x, h, \mathscr{M}) = \frac{p(y|x, h, \mathscr{M})}{\sum_{y\in\mathscr{Y}} p(y|x, h, \mathscr{M})} = p(y|x, h, \mathscr{M})\,.$$

Indeed, in that case, we get by construction

$$\begin{aligned}&\inf\{\ell(Y, y) : y \in \mathscr{C}(x, \delta, h, \mathscr{M})\} = \mathbb{I}\{Y \notin \mathscr{C}(x, \delta, h, \mathscr{M})\} \\ &\quad\leqslant \mathbb{I}\{p(Y|x, h, \mathscr{M}) \leqslant 1 - \sum_{y\in\mathscr{C}(x,\delta,h,\mathscr{M})} p(y|x, h, \mathscr{M})\}\,. \\ &\quad\leqslant \mathbb{I}\{p(Y|x, h, \mathscr{M}) \leqslant 1 - \max\{\sum_{i=1}^{K} p(y_{k_i}|x, h, \mathscr{M}) : \sum_{i=1}^{K} p(y_{k_i}|x, h, \mathscr{M}) \geqslant 1-\delta\}\}\,. \\ &\quad\leqslant \mathbb{I}\{p(Y|x, h, \mathscr{M}) \leqslant \delta\}\,.\end{aligned}$$

Since this last quantity is 0 for all $\delta < p(Y|x, h, \mathscr{M})$, we thus get the desired conclusion.

Example 2 When $\mathscr{Y} = \mathbb{R}$ and $\ell(y, y') = (y - y')^2$, we get some interesting result. Indeed, for a regression model with σ^2-sub-Gaussian noise, the confidence set is of the form

$$\begin{aligned}\mathscr{C}(x, \delta, h, \mathscr{M}) &= \Big[f_{h,\mathscr{M}}(x) - \lambda B - \sqrt{2\sigma^2_{x,h,\mathscr{M}} \ln(C_{h,\mathscr{M}}/\delta)} - \sqrt{2\sigma^2 \ln(1/\delta)}, \\ &\qquad f_{h,\mathscr{M}}(x) + \lambda B + \sqrt{2\sigma^2_{x,h,\mathscr{M}} \ln(C_{h,\mathscr{M}}/\delta)} + \sqrt{2\sigma^2 \ln(1/\delta)}\Big] \\ &\subset \Big[a - c\sqrt{2\ln(1/\delta)}, b + c\sqrt{2\ln(1/\delta)}\Big]\,,\end{aligned}$$

where we introduced $a = f_{h,\mathscr{M}}(x) - \lambda B - \sqrt{2\sigma^2_{x,h,\mathscr{M}} \ln(C_{h,\mathscr{M}})}$, $b = f_{h,\mathscr{M}}(x) + \lambda B + \sqrt{2\sigma^2_{x,h,\mathscr{M}} \ln(C_{h,\mathscr{M}})}$, and finally $c = \sqrt{\sigma^2_{x,h,\mathscr{M}} + \sigma^2}$ for convenience. Now, at the price of enlarging the confidence set,

$$\begin{aligned}\inf\{\ell(Y,y) : y \in \mathscr{C}(x,\delta,h,\mathscr{M})\} &\geqslant \inf\{(Y-y)^2 : y \in \big[a - c\sqrt{2\ln(1/\delta)}, b + c\sqrt{2\ln(1/\delta)}\big]\} \\ &= \mathbb{I}\{Y < a - c\sqrt{2\ln(1/\delta)}\}(Y - a + c\sqrt{2\ln(1/\delta)})^2 \\ &\quad + \mathbb{I}\{Y > b + c\sqrt{2\ln(1/\delta)}\}(Y - b - c\sqrt{2\ln(1/\delta)})^2 .\end{aligned}$$

The critical values for δ are $e^{-\frac{(Y-a)^2}{2c^2}}$ if $Y < a$ and $e^{-\frac{(Y-b)^2}{2c^2}}$ if $Y > b$, that is we find that $\alpha(Y; x, h, \mathscr{M}) = e^{-\frac{(a-Y)_+^2}{2c^2} - \frac{(Y-b)_+^2}{2c^2}}$. Further, we remark that

$$\begin{aligned}\int_{\mathscr{Y}} \alpha(y; x, h, \mathscr{M}) dy &= \int_{-\infty}^{a} e^{-\frac{(a-y)^2}{2c^2}} dy + |b-a| + \int_b^{\infty} e^{-\frac{(y-b)^2}{2c^2}} dy \\ &= \sqrt{2\pi c^2} + |b-a| ,\end{aligned}$$

which leads to the (possibilistic) distribution induced by the enlargement of $\mathscr{C}(x,\delta,h,\mathscr{M})$ defined by

$$\tilde{p}(dy|x,h,\mathscr{M}) = \frac{1}{\sqrt{2\pi c^2} + |b-a|} e^{-\frac{(a-Y)_+^2}{2c^2} - \frac{(Y-b)_+^2}{2c^2}} \lambda(dy) .$$

4.2. Loss estimates

The benefit of confidence intervals is to be able to easily derive an upper bound on the loss of a prediction algorithm at the next input point (or any other). We illustrate this point below in the Gaussian setting by providing a list of estimates for the quadratic, self-information and transportation losses defined in the previous section, that directly derives from our previous results.

We provide our illustrative examples in the case of Gaussian regression only, for the purpose of clarity.

Theorem 4.7 (Predictive quadratic loss) *The predictive quadratic loss of the model that predicts $\mathscr{N}(\widehat{f}_h(x), \sigma_h)$ from history h, in the sub-Gaussian streaming regression model is given by*

$$\overline{\ell}_{\mathscr{Y}}(x, \delta_0, \mathcal{P}(h)) = 4\Big(b_h + c_h\sqrt{2\ln(1/\delta_0)}\Big)^2 ,$$

where b_h and c_h are such that

$$\mathscr{C}(x,\delta,\mathcal{P}(h)) = \{y : |y - \widehat{f}_h(x)| \leqslant b_h + c_h\sqrt{2\ln(1/\delta)}\} .$$

Proof of Theorem 4.7:

Specifying $\mathcal{P}(h)$ and $\alpha(y, x, \mathcal{P}(h))$ in the Gaussian example, we derive from Definition 4.4 the equalities

$$\begin{aligned}
\overline{\ell}_{\mathscr{Y}}(x, \delta_0, \mathcal{P}(h)) &= \sup\Big\{(y-y')^2 : \\
&\qquad y, y' \text{ such that } \exp\Big(-\frac{(\widehat{f}_h(x) - b_h - y)_+^2}{2c_h^2} - \frac{(y - \widehat{f}_h(x) - b_h)_+^2}{2c_h^2}\Big) \geqslant \delta_0\Big\} \\
&= \max\Big\{(y-y')^2 : y = \widehat{f}_h(x) - b_h - \sqrt{2c_h^2 \ln(1/\delta_0)} \\
&\qquad \text{or } y = \widehat{f}_h(x) + b_h + c_h\sqrt{2\ln(1/\delta_0)}\Big\} \\
&= 4\Big(b_h + \sqrt{2c_h^2 \ln(1/\delta_0)}\Big)^2 .
\end{aligned}$$

□

Theorem 4.8 (Predictive self-information loss) *The predictive self-information loss of the model that predicts $\mathcal{N}(\widehat{f}_h(x), \sigma_h)$ from history h in the sub-Gaussian streaming regression model is given by*

$$\overline{\ell}_I(x, \delta_0, \mathcal{P}(h)) = \frac{1}{2}\ln(2\pi\sigma_h^2) + \frac{1}{2\sigma_h^2}(b_h + c_h\sqrt{2\ln(1/\delta_0)})^2 ,$$

where b_h and c_h are such that

$$\mathscr{C}(x, \delta, \mathcal{P}(h)) = \{y : |y - \widehat{f}_h(x)| \leqslant b_h + c_h\sqrt{2\ln(1/\delta)}\} .$$

Proof of Theorem 4.8:

We consider a confidence set of the form

$$\mathscr{C}(x, \delta, \mathcal{P}(h)) = \{y : |y - \widehat{f}_h(x)| \leqslant b_h + c_h\sqrt{2\ln(1/\delta)}\} ,$$

Specifying $\mathcal{P}(h)$ and $\alpha(y, x, \mathcal{P}(h))$ in the Gaussian example, we derive from Defini-

tion 4.5

$$\begin{aligned}\overline{\ell}_I(x,\delta_0,\mathcal{P}(h)) &= \sup\Bigg\{\frac{(y-\widehat{f}_h(x))^2}{2\sigma_h^2}+\frac{1}{2}\ln(2\pi\sigma_h^2): \\ &\quad y \text{ such that } \exp\Bigg(-\frac{(\widehat{f}_h(x)-b_h-y)_+^2}{2c_h^2}-\frac{(y-\widehat{f}_h(x)-b_h)_+^2}{2c_h^2}\Bigg)\geqslant\delta_0\Bigg\} \\ &= \max\Bigg\{\frac{(y-\widehat{f}_h(x))^2}{2\sigma_h^2}+\frac{1}{2}\ln(2\pi\sigma_h^2): \\ &\quad y=\widehat{f}_h(x)-b_h-\sqrt{2c_h^2\ln(1/\delta_0)} \text{ or } y=\widehat{f}_h(x)+b_h+\sqrt{2c_h^2\ln(1/\delta_0)}\Bigg\} \\ &= \frac{1}{2}\ln(2\pi\sigma_h^2)+\frac{1}{2\sigma_h^2}(b_h+c_h\sqrt{2\ln(1/\delta_0)})^2\,.\end{aligned}$$

□

Theorem 4.9 (Transportation and predictive Transportation loss) *In the streaming regression model when the noise is (exactly) Gaussian conditionally on the past, then, with probability higher than* $1-4\delta$ *it holds*

$$\mathscr{T}\Big(\mathscr{N}(f^\star(x),R^2),\mathscr{N}(f_{\theta_N^\dagger}(x),\widehat{\sigma}_N^2)\Big)\leqslant$$
$$R^2\Bigg[4||\varphi(x)||^2_{G_N^\dagger}\ln\Big(\frac{\kappa_d(e^2\Lambda_N^2)}{\delta}\Big)+\max\Bigg\{\sqrt{\frac{C_N(\delta)}{N}}+\sqrt{\frac{C_N(\delta)+D_N(\delta)}{N}},\sqrt{\frac{2C_N(\delta)}{N}}\Bigg\}^2\Bigg].$$

Further, one can replace R *with any upper bound* $\widehat{R}_n$ *for instance from Corollary 3.12 in order to get a fully empirical bound on the transportation loss:*

$$\widehat{R}_n=\sqrt{\widehat{\sigma}_N^2}+R^+\Bigg(\sqrt{\frac{C_N(\delta)}{N}}+\sqrt{\frac{C_N(\delta)+D_N(\delta)}{N}}\Bigg).$$

If $\mathcal{P}_{\mathscr{N}}$ *denotes the model assuming that the noise is exactly conditionally Gaussian, then the predictive transportation loss for that model model is given by*

$$\begin{aligned}\overline{\mathscr{T}}(x,\delta_0,\mathcal{P}_{\mathscr{N}}(h)) &= R^2\Bigg[8||\varphi(x)||^2_{G_N^\dagger}\ln\Big(\frac{\kappa_d(e^2\Lambda_N^2)}{\delta_0}\Big) \\ &\quad +2\max\Bigg\{\sqrt{\frac{C_N(\delta_0)}{N}}+\sqrt{\frac{C_N(\delta)+D_N(\delta_0)}{N}},\sqrt{\frac{2C_N(\delta_0)}{N}}\Bigg\}^2\Bigg].\end{aligned}$$

Proof of Theorem 4.9:

We simply remark that in the specific case of Gaussian distributions, the transportation loss can be written in a fully explicit way as

$$\mathscr{T}\Big(\mathscr{N}(f^\star(x),R^2),\mathscr{N}(f_{\theta_N^\dagger}(x),\widehat{\sigma}_N^2)\Big)=|f^\star(x)-f_{\theta_N^\dagger}(x)|^2+|R-\widehat{\sigma}_N|^2.$$

Then, combining Theorem 3.3 and Theorem 3.8 together with a union bound, it comes with probability higher than $1-4\delta$

$$|f^\star(x)-f_{\theta_N^\dagger}(x)|^2+|R-\widehat{\sigma}_N|^2\leqslant R^2\Bigg[4||\varphi(x)||^2_{G_N^\dagger}\ln\Big(\frac{\kappa_d(e^2\Lambda_N^2)}{\delta}\Big)$$
$$+\max\left\{\sqrt{\frac{C_N(\delta)}{N}}+\sqrt{\frac{C_N(\delta)+D_N(\delta)}{N}},\sqrt{\frac{2C_N(\delta)}{N}}\right\}^2\Bigg].$$

□

It is interesting to remark from these results that the self-information and transportation loss may behave very differently. In particular, using the self-information loss criterion appears to be dangerous in the context of model selection when one must choose between different models, when the noise level is unknown and must be estimate. Indeed in this case, models estimating a large variance will be favored due to the σ_h^2 appearing in the denominator. This situation does not seem to appear when using the transportation loss criterion. Dealing with the transportation is however not easy beyond the Gaussian case, for which everything is explicit (see Section 2.1).

Remark 4.10 (Mismatch detection) *An application of the previous tools is the possibility to detect a mismatch between the considered model of observations and the actual observed data. A simple way to detect such a mismatch is to use the model adequacy, and test whether $\alpha(y_{n+1},x_{n+1},\mathcal{P}(h_n))$ is less than some alert threshold τ. A complementary way is to track large changes in the predictive loss, such as a large increase. We do not detail this more in order to avoid deviating the reader's focus from the main topic of this paper.*

5. Self-normalized techniques for Sub-Gaussian Regression

In this section, we provide the proofs of the main theorems regarding the mean and variance estimates of the considered least-squares solution. The proofs are a little long but are actually not difficult as they make use of standard tools in statistics. One method is the Laplace method, another one is peeling, yet another one is the martingale method. Although we make use of a sort of localization technique, thanks to the concentration of confidence sets, we do not need to resort to symmetrization (and Rademacher complexities) for the construction of confidence sets.

We first recall a simple yet powerful result argument for randomly stopped vector valued processes. An example of proof of this result is given in (Abbasi-Yadkori, Pal and Szepesvari, 2011) and we briefly reexplained it below.

Lemma 5.1 (Vector-valued Martingale Control) *Assume that the noise sequence* $\{\xi_n\}_{n=0}^{\infty}$ *is conditionally* R*-sub-Gaussian*

$$\forall n \in \mathbb{N}, \forall \lambda \in \mathbb{R}, \quad \ln \mathbb{E}[\exp(\lambda \xi_n)|h_{n-1}] \leqslant \frac{\lambda^2 R^2}{2} .$$

Let N *be a stopping time with respect to the filtration* $\{\mathscr{F}_n\}_{n=0}^{\infty}$ *generated by the variables* $\{x_n, \xi_n\}_{n=0}^{\infty}$*. Let us denote*

$$M_m^q = \exp\left(q^\top \sum_{n=1}^{m} \varphi(x_n)\xi_n - \frac{R^2}{2} q^\top \Phi_m^\top \Phi_m q \right)$$

where $\Phi_m = (\varphi(x_1), \dots, \varphi(x_m))^\top$ *is* $N \times d$*. Then, for all* $q \in \mathbb{R}^d$ *the quantity* M_N^q *is well defined and satisfies*

$$\ln \mathbb{E}[M_N^q] \leqslant 0 .$$

The only difficulty in the proof is to handle the stopping time. Indeed, for all $m \in \mathbb{N}$, thanks to the conditional R-sub-Gaussian property, it is immediate to show that $\{M_m^q\}_{q=0}^{\infty}$ is a non-negative super-martingale and actually satisfies $\ln \mathbb{E}[M_m^q] \leqslant 0$.

By the convergence theorem for nonnegative super-martingales, $M_\infty^q = \lim_{m\to\infty} M_m^q$ is almost surely well-defined, and thus M_N^q is well-defined (whether $N < \infty$ or not) as well. In order to show that $\ln \mathbb{E}[M_N^q] \leqslant 0$, we introduce a stopped version $Q_m^q = M_{\min\{N,m\}}^q$ of $\{M_m^q\}_m$. Now $\mathbb{E}[M_N^q] = \mathbb{E}[\liminf_{m\to\infty} Q_m^q] \leqslant \liminf_{m\to\infty} \mathbb{E}[Q_m^q] \leqslant 1$ by Fatou's lemma, which concludes the proof. We refer to (Abbasi-Yadkori, Pal and Szepesvari, 2011) for further details.

We continue this section with a known and powerful self-normalized result for vector-valued martingales (see e.g. (Peña, Lai and Shao, 2008)) of which we make extensive use in the sequel:

Lemma 5.2 (Laplace method) *Let* A *be a* d*-dimensional random vector and* B *a* $d \times d$ *random matrix. Assume that for all* $\lambda \in \mathbb{R}^d$*,* $\mathbb{E}[\exp(\lambda^\top A - \frac{1}{2}\lambda^\top B \lambda)] \leqslant 1$*. Then for any deterministic* $d \times d$ *matrix* V*, it holds*

$$\mathbb{P}\left(||A||_{(B+V)^{-1}} \geqslant R\sqrt{2 \ln\left(\frac{\det(B+V)^{1/2}}{\delta \det(V)^{1/2}} \right)} \right) \leqslant \delta .$$

Finally, we provide the following result that is useful when dealing with generic real-valued distributions, and is possibly interesting beyond the scope of this paper. The proof technique combines a simple peeling argument together with the martingale method, and is inspired from techniques from (Cappé, Garivier and Maillard, 2013).

Lemma 5.3 (Self-normalized Concentration inequality) *Let $\{Z_i\}_{i=1}^{\infty}$ be a sequence of random variables. Let $\varphi : \mathbb{R} \to \mathbb{R}^+$ be a convex upper-enveloppe of the cumulant generative function of the conditional distributions with $\varphi(0) = 0$, and $\varphi_\star$ its Legendre-Fenchel transform, that is:*

$$\begin{aligned} \forall \lambda \in \mathscr{D}, \forall i, \qquad & \ln \mathbb{E}\Big[\exp\Big(\lambda Z_i\Big) \Big| \mathscr{H}_{i-1} \Big] \leqslant \varphi(\lambda)\,, \\ \forall x \in \mathbb{R} \qquad & \varphi_\star(x) = \sup_{\lambda \in \mathbb{R}} \big(\lambda x - \varphi(\lambda)\big), \end{aligned}$$

where $\mathscr{D} = \{\lambda \in \mathbb{R} : \forall i, \ln \mathbb{E}\Big[\exp\Big(\lambda Z_i\Big) \Big| \mathscr{H}_{i-1} \Big] < \infty\}$. Assume that $\mathscr{D}$ contains an open neighborhood of 0. Then, $\forall c \in \mathbb{R}^+$, there exists a unique x_c such that for all i, $x_c > \mathbb{E}\Big[Z_i \Big| \mathscr{H}_{i-1}\Big]$, and $\varphi_\star(x_c) = c$, and a unique x'_c such that for all i, $x'_c < \mathbb{E}\Big[Z_i \Big| \mathscr{H}_{i-1}\Big]$ and $\varphi_\star(x'_c) = c$. We define $\varphi^{-1}_{\star,+} : c \mapsto x_c$, $\varphi^{-1}_{\star,-} : c \mapsto x'_c$. Then $\varphi^{-1}_{\star,+}$ is not decreasing and $\varphi^{-1}_{\star,-}$ is not increasing.

Let N_n be a random stopping time (for the filtration generated by $\{Z_i\}_{i=1}^{\infty}$) a.s. bounded by n. Then

$$\begin{aligned} \forall \delta \in (0,1)\, \mathbb{P}\left[\frac{1}{N_n} \sum_{i=1}^{N_n} Z_i \geqslant \varphi^{-1}_{\star,+}\left(\frac{1 + \ln(1/\delta)}{N_n} \right) \right] \quad &\leqslant \quad \Big(\lceil \ln(n) \ln(e/\delta) \rceil \Big) \delta \\ \forall \delta \in (0,1)\, \mathbb{P}\left[\frac{1}{N_n} \sum_{i=1}^{N_n} Z_i \leqslant \varphi^{-1}_{\star,-}\left(\frac{1 + \ln(1/\delta)}{N_n} \right) \right] \quad &\leqslant \quad \Big(\lceil \ln(n) \ln(e/\delta) \rceil \Big) \delta \end{aligned}$$

Now, if N is a (possibly unbounded) random stopping time for the filtration generated by $\{Z_i\}_{i=1}^{\infty}$, it holds

$$\begin{aligned} \mathbb{P}\left(\frac{1}{N} \sum_{i=1}^{N} Z_i \geqslant \varphi^{-1}_{\star,+}\left(\frac{1 + \ln(1/\delta)}{N} \left[1 + \frac{2}{\ln(1/\delta)} \ln\left(\frac{\pi \ln(N) \ln(1/\delta)}{\sqrt{6}(1 + \ln(1/\delta))} \right) \right] \right) \right) \quad &\leqslant \quad \delta \\ \mathbb{P}\left(\frac{1}{N} \sum_{i=1}^{N} Z_i \leqslant \varphi^{-1}_{\star,-}\left(\frac{1 + \ln(1/\delta)}{N} \left[1 + \frac{2}{\ln(1/\delta)} \ln\left(\frac{\pi \ln(N) \ln(1/\delta)}{\sqrt{6}(1 + \ln(1/\delta))} \right) \right] \right) \right) \quad &\leqslant \quad \delta\,. \end{aligned}$$

Example: For R-Sub-Gaussian distributions (e.g. for $Z = X - \mathbb{E}[X]$ with $X \in [0,1]$, then $R = 1/2$):

$$\varphi(\lambda) = \frac{\lambda^2 R^2}{2}, \qquad \varphi^\star(x) = \frac{x^2}{2R^2}, \qquad \varphi^{-1}_{\star,+}(c) = \sqrt{2R^2 c}, \qquad \varphi^{-1}_{\star,-}(c) = -\sqrt{2R^2 c}\,.$$

Proof of Lemma 5.3:

First, one easily derives (or recalls) the following properties, from those of the Legendre-Fenchel transform.

- $\varphi_\star(0)=0$, $\varphi_\star(x) \overset{x\to+\infty}{\to} \infty$, $\varphi_\star$ is convex, increasing on $\mathbb{R}^+$.
- $\forall x$ such that $\varphi_\star(x) < \infty$, there exists a unique $\lambda_x \in \mathscr{D}_\nu$ such that $\varphi_\star(x) = \lambda_x x - \varphi(\lambda_x)$.
- $\forall c \in \mathbb{R}^+$, there exists a unique $x_c > \mathbb{E}[Z]$ such that $\varphi_\star(x_c)=c$. We write it $\varphi^{-1}_{\star,+}(c)$. $\varphi^{-1}_{\star,+}$ is not decreasing.

---

1. A peeling argument We start with a peeling argument. Let us choose some $\eta>0$ and define $t_k=(1+\eta)^k$, for $k=0,\dots,K$, with $K=\lceil \frac{\ln(n)}{\ln(1+\eta)} \rceil$ (thus $n \leqslant t_K$). Let $\varepsilon_t \in \mathbb{R}^+$ be a sequence that is non-increasing in t.

$$\begin{aligned}
&\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varepsilon_{N_n}\Big) \\
&\quad\leqslant\ \mathbb{P}\Big(\bigcup_{k=1}^{K}\{t_{k-1}<N_n\leqslant t_k\}\cap\{\sum_{i=1}^{N_n} Z_i \geqslant N_n\varepsilon_{N_n}\}\Big) \\
&\quad\leqslant\ \sum_{k=1}^{K}\mathbb{P}\Big(\exists t\in(t_{k-1},t_k]:\sum_{i=1}^{t} Z_i \geqslant t\varepsilon_t\Big)
\end{aligned}$$

Let $\lambda_k>0$, for $k=1,\dots,K$.

$$\begin{aligned}
&\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varepsilon_{N_n}\Big) \\
&\quad\leqslant\ \sum_{k=1}^{K}\mathbb{P}\Big(\exists t\in(t_{k-1},t_k]:\sum_{i=1}^{t} Z_i \geqslant t\varepsilon_t\Big) \\
&\quad\leqslant\ \sum_{k=1}^{K}\mathbb{P}\Big(\exists t\in(t_{k-1},t_k]:\exp\Big(\lambda_k\Big(\sum_{i=1}^{t} Z_i\Big)\Big)\geqslant\exp(\lambda_k t\varepsilon_t)\Big) \\
&\quad\leqslant\ \sum_{k=1}^{K}\mathbb{P}\Big(\exists t\in(t_{k-1},t_k]:\underbrace{\exp\Big(\lambda_k\Big(\sum_{i=1}^{t} Z_i\Big)-t\varphi(\lambda_k)\Big)}_{W_{k,t}} \\
&\qquad\qquad\qquad\qquad\geqslant\exp\Big(t\big(\lambda_k\varepsilon_t-\varphi(\lambda_k)\big)\Big)\Big) \\
&\quad\leqslant\ \sum_{k=1}^{K}\mathbb{P}\Big(\exists t\in(t_{k-1},t_k]:W_{k,t}\geqslant\exp\Big(t\big(\lambda_k\varepsilon_{t_k}-\varphi(\lambda_k)\big)\Big)\Big).
\end{aligned}$$

Since $\varepsilon_{t_k}>0$, we can choose a $\lambda_k>0$ such that $\varphi^\star(\varepsilon_{t_k})=\lambda_k\varepsilon_{t_k}-\varphi(\lambda_k)$.

2. Doob's maximal inequality At this, point, we show that the sequence $\{W_{k,t}\}_t$ is a non-negative super-martingale, where $W_{k,t} = \exp\Big(\lambda_k\big(\sum_{i=1}^t Z_i\big) - t\varphi(\lambda_k)\Big)$. Indeed, note that:

$$\begin{aligned}\mathbb{E}[W_{k,t+1}|\mathscr{F}_t] &= W_{k,t}\mathbb{E}[\exp\big(\lambda_k Z_{t+1}\big)|\mathscr{F}_t]\exp(-\varphi(\lambda_k))\\ &\leqslant W_{k,t}\,.\end{aligned}$$

Thus, using that $t_{k-1} \geqslant t_k/(1+\eta)$, we find

$$\begin{aligned}&\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varepsilon_{N_n}\Big)\\ &\leqslant \sum_{k=1}^K \mathbb{P}\Big(\exists t \in (t_{k-1}, t_k] W_{k,t} \geqslant \exp\Big(t\varphi^\star(\varepsilon_{t_k})\Big)\Big)\\ &\leqslant \sum_{k=1}^K \mathbb{P}\Big(\max_{t\in(t_{k-1},t_k]} W_{k,t} \geqslant \exp\Big(\frac{t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big)\Big)\\ &\overset{(a)}{\leqslant} \sum_{k=1}^K \exp\Big(-\frac{t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big)\,,\end{aligned}$$

where (a) holds by application of Doob's maximal inequality for non-negative super-martingales, using that $\max_{t\in(t_{k-1},t_k]} W_{k,t} \leqslant \max_{t\in(0,t_k]} W_{k,t}$ and $W_{k,0} \leqslant 1$.

3. Parameter tuning for bounded N_n Now, let us choose ε_t such that $t\varphi_\star(\varepsilon_t) = c > 1$ is a constant, that is $\varepsilon_t = \varphi_{\star,+}^{-1}(c/t)$ (non increasing with t). Thus, we get for all $\eta \in (0, n-1)$:

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varepsilon_{N_n}\Big) &\leqslant \sum_{k=1}^{\lceil \frac{\ln(n)}{\ln(1+\eta)}\rceil} \exp\Big(-\frac{t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big)\\ &\leqslant \lceil\frac{\ln(n)}{\ln(1+\eta)}\rceil \exp\Big(-\frac{c}{1+\eta}\Big)\end{aligned}$$

For $\eta = 1/(c-1)$ (which is not optimal in general) and $c = \ln(e/\delta) > 1$ we get for $\delta < 1$ that

$$\begin{aligned}&\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varphi_{\star,+}^{-1}(\ln(e/\delta)/N_n)\Big)\\ &\leqslant \lceil\ln(n)c\rceil e\exp\Big(-c\Big)\\ &= \lceil\ln(n)\ln(e/\delta)\rceil\delta\,.\end{aligned}$$

Another way to tune the parameters is to make use of the Lambert W function.

---

4. Parameter tuning for unbounded N_n
We restart from

$$\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varepsilon_{N_n}\Big) \quad \leqslant \quad \sum_{k=1}^{K} \exp\Big(-\frac{t_k \varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big)\,,$$

where $t_k = (1+\eta)^k$ and $K = \lceil \frac{\ln(n)}{\ln(1+\eta)} \rceil$, and choose a different tuning for ε_t in order to handle an infinite sum (with $K=\infty$). Let us choose ε_t that satisfies $t\varphi_\star(\varepsilon_t) = c + f(t)$, where $f(t)$ is chosen such that

$$\sum_{k=1}^{\infty} \exp\Big(-\frac{f(t_k)}{1+\eta}\Big) < \infty\,.$$

Choosing $f(t_k) = (1+\eta)\ln(k^2\pi^2/6) = 2(1+\eta)\ln\Big(\frac{\pi \ln(t_k)}{6^{1/2}(1+\eta)}\Big)$ and $\eta = 1/(c-1)$, it comes

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varepsilon_{N_n}\Big) \quad &\leqslant \quad \exp\Big(-\frac{c}{1+\eta}\Big)\sum_{k=1}^{\infty}\frac{1}{k^2}\frac{6}{\pi^2} \\ &= \quad \exp\Big(-\frac{c}{1+\eta}\Big) \\ &= \quad \exp(-c)e\,.\end{aligned}$$

Thus, $f(t_k) = \frac{2c}{c-1}\ln\Big(\frac{\pi\ln(t_k)(c-1)}{6^{1/2}c}\Big)$, and for $c = \ln(e/\delta) > 1$, it comes

$$\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varphi_{\star,+}^{-1}\Big(\frac{c}{N_n}\Big[1+\frac{2}{c-1}\ln\Big(\frac{\pi\ln(N_n)(c-1)}{6^{1/2}c}\Big)\Big]\Big)\Big) \leqslant \delta$$

Thus, we obtain

$$\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varphi_{\star,+}^{-1}\Big(\frac{1+\ln(1/\delta)}{N_n}\Big[1+\frac{2}{\ln(1/\delta)}\ln\Big(\frac{\pi\ln(N_n)\ln(1/\delta)}{6^{1/2}(1+\ln(1/\delta))}\Big)\Big]\Big)\Big) \leqslant \delta$$

---

5. Reverse bounds. We now provide a similar result for the reverse bound. Let

$\varepsilon_t \in \mathbb{R}$ be a sequence that is non-decreasing with t, and $\lambda_k > 0$, for $k = 1, \dots, K$. Then

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \leqslant \varepsilon_{N_n}\Big) &\leqslant \sum_{k=1}^{K} \mathbb{P}\Big(\exists t \in (t_{k-1}, t_k] \exp\Big(-\lambda_k\Big(\sum_{i=1}^{t} Z_i\Big) - t\varphi(-\lambda_k)\Big) \\ &\qquad \geqslant \exp\Big(t(-\lambda_k\varepsilon_t - \varphi(-\lambda_k))\Big)\Big) \\ &\leqslant \sum_{k=1}^{K} \mathbb{P}\Big(\exists t \in (t_{k-1}, t_k], W_{k,t} \geqslant \exp\Big(t(-\lambda_k\varepsilon_{t_k} - \varphi(-\lambda_k))\Big)\Big)\end{aligned}$$

If $\varepsilon_{t_k} < \mathbb{E}[Z_{t_k}]$, we can choose $\lambda_k = \lambda_{\varepsilon_{t_k}} > 0$ such that $\varphi^\star(\varepsilon_{t_k}) = -\lambda_k\varepsilon_{t_k} - \varphi(-\lambda_k) \geqslant 0$. Thus, using that $t_{k-1} > t_k/(1+\eta)$, it comes

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \leqslant \varepsilon_{N_n}\Big) &\leqslant \sum_{k=1}^{K} \mathbb{P}\Big(\exists t \in (t_{k-1}, t_k], W_{k,t} \geqslant \exp\Big(t\varphi^\star(\varepsilon_{t_k})\Big)\Big) \\ &\leqslant \sum_{k=1}^{K} \mathbb{P}\Big(\max_{t\in(t_{k-1},t_k]} W_{k,t} \geqslant \exp\Big(\frac{t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big)\Big) \\ &\leqslant \sum_{k=1}^{K} \exp\Big(\frac{-t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big)\end{aligned}$$

Now, let us choose $\varepsilon_t < \mathbb{E}[Z_t]$ such that $t\varphi_\star(\varepsilon_t) = c > 1$, that is $\varepsilon_t = \varphi_{\star,-}^{-1}(c/t)$ (non decreasing with t). For $\eta = 1/(c-1)$ and $c = \ln(e/\delta)$, we obtain

$$\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \leqslant \varphi_{\star,-}^{-1}(\ln(e/\delta)/N_n)\Big) \leqslant \lceil \ln(n)\ln(e/\delta)\rceil\delta\,.$$

□

5.1. Mean estimates

Lemma 5.1 enables to prove powerful concentration results. We apply it first to the setting of ordinary least-squares regression and then to ridge-regression (aka Tikhonov–Phillips regularization, or ℓ_2-regularization).

Theorem 5.4 (Self-normalized vector concentration for random stopping time) *Let assume that the observations $\{(x_n, y_n)\}_{n\in[N]}$ come from a regression model $y_n = f_{\theta^\star}(x_n) + \xi_n$ with $\theta^\star \in \Theta$, and that N is a random stopping time for the filtration of the*

past. Let $\lambda > 0$ be a deterministic constant. The following result holds in the case when a deterministic bound Λ is known on $\Lambda_N = \lambda_{\max}(G_N)$:

$$\mathbb{P}\left(\Lambda_N \leqslant \Lambda \cap \sup_{p:||p||^2_{G_N}\geqslant\lambda} \frac{|E_N^\top \Phi_N p|^2}{||p||^2_{G_N}} > 4R^2 \ln\left[4\lceil\ln(\Lambda/\lambda)\rceil \left\lceil \frac{\ln(e\Lambda_N/\lambda)}{2}\right\rceil \frac{K_d(\frac{e\Lambda_N}{\lambda})^d + d}{\delta}\right]\right) \leqslant \delta,$$

where $K_d = (12(d+1)\sqrt{d})^d$ and $\Lambda_N = \lambda_{\max}(G_N)$ is the maximal eigenvalue of G_N. Note that the dependency with Λ only appears in the first ln *term. In general, it holds:*

$$\mathbb{P}\left(\sup_{p:||p||^2_{G_N}\geqslant\lambda} \frac{|E_N^\top \Phi_N p|^2}{||p||^2_{G_N}} > 4R^2 \ln\left[2\pi^2 \ln^2(\Lambda_N/\lambda)\left\lceil \frac{\ln(e\Lambda_N/\lambda)}{2}\right\rceil \frac{K_d(\frac{e\Lambda_N}{\lambda})^d + d}{3\delta}\right]\right) \leqslant \delta\,, \quad (5)$$

Likewise, an alternative bound is given by

$$\mathbb{P}\left(\sup_{p:||p||^2_{G_N}\geqslant\Lambda_N^{-1}} \frac{|E_N^\top \Phi_N p|^2}{||p||^2_{G_N}} > 4R^2 \ln\left[8\pi^2 \ln^2(\Lambda_N)\left\lceil \ln(e\Lambda_N)\right\rceil \frac{K_d(e\Lambda_N)^{2d} + d}{3\delta}\right]\right) \leqslant \delta\,. \quad (6)$$

Further, for any deterministic semi-definite positive matrix V it holds

$$\mathbb{P}\left(||\Phi_N^\top E_N||_{(G_N+V)^{-1}} \geqslant R\sqrt{2\ln\left(\frac{det(G_N+V)^{1/2}}{\delta det(V)^{1/2}}\right)}\right) \leqslant \delta\,.$$

The first result is better understood for ordinary least-squares regression (without regularization). Indeed, let $\theta \in \Theta_N$ be one minimizer of the least-squares error. We first note that if A is self-adjoint, then $||\theta^\star - \theta||_A = \sup_{p\in\mathscr{S}^D} \frac{|p^\top A(\theta^\star-\theta)|}{||p||_A}$, where $\mathscr{S}^D$ is the unit sphere of $\mathbb{R}^D$. Further, when $A = G_N$, we have the property that

$$G_N(\theta^\star - \theta) = G_N\theta^\star - \Phi_N^\top(\Phi_N\theta^\star + E_N) = \Phi_N^\top E_N\,.$$

Thus, $\sup_{p\in\mathscr{S}^D} \frac{p^\top \Phi_N^\top E_N}{||p||_{G_N}} = ||\theta^\star - \theta||_{G_N}$. A similar result appeared in (Rusmevichientong and Tsitsiklis, 2010). We derive a slightly different version here, by combining Lemma 5.1 with the Laplace method from (Peña, Lai and Shao, 2008) (actually a variant of it is used in (Rusmevichientong and Tsitsiklis, 2010)) and a covering argument in a slightly different way.

The last result is for ℓ_2-regularized least-squares regression, and was shown in (Abbasi-Yadkori, Pal and Szepesvari, 2011) from a combination of Lemma 5.1 with the Laplace method from (Peña, Lai and Shao, 2008). We do not reproduce this proof here and refer to (Abbasi-Yadkori, Pal and Szepesvari, 2011) for details.

Before continuing, note that if we introduce the following quantity,

$$\kappa_d(x) \quad = \quad \frac{2}{3}\pi^2 \ln^2(x/e)\left\lceil \frac{\ln(x)}{2}\right\rceil [(12(d+1)\sqrt{d})^d x^d + d]$$

we immediately deduce from (5) and (6) that

$$\mathbb{P}\left(\sup_{p:||p||^2_{G_N} \geqslant \lambda} \frac{|E_N^\top \Phi_N p|^2}{||p||^2_{G_N}} > 4R^2 \ln\left[\frac{\kappa_d(e\Lambda_N/\lambda)}{\delta}\right]\right) \leqslant \delta$$
$$\mathbb{P}\left(\sup_{p:||p||^2_{G_N} \geqslant \Lambda_N^{-1}} \frac{|E_N^\top \Phi_N p|^2}{||p||^2_{G_N}} > 4R^2 \ln\left[\frac{\kappa_d(e^2\Lambda_N^2)}{\delta}\right]\right) \leqslant \delta\,.$$

Proof of Theorem 5.4:

Let us first notice that

$$\Phi_N^\top E_N = \sum_{n=1}^N \varphi(x_n)\xi_n\,,$$

is a sum of noise terms, and that due to normalization, one could consider $\mathbb{R}^D$ instead of $\mathscr{S}^D$, or a unit sphere in A-metric in case the corresponding covering number is easy to handle. or any unit sphere for any norm. We consider in the sequel a covering of $\mathscr{S}^D$ in Euclidean norm as it applies to all cases and is simple to handle. We also denote $\mathscr{H}_{n-1}$ the filtration generated by the past history of observations up to step $n-1$.

1. A peeling argument In order to handle the concentration of the term we want to control, let us first introduce a sequence of values $\lambda_k = (1+\eta)^k\lambda$, for some $\eta > 0$ and $\lambda > 0$ to be defined later. We use this in order to localize the quantity $||p||_{G_N}$. Thus, we use the decomposition

$$\mathbb{P}\left(\sup_{p:||p||_{G_N} \geqslant \lambda} \frac{|E_N^\top \Phi_N p|^2}{||p||^2_{G_N}} > \varepsilon_N\right) \leqslant \sum_{k=0}^\infty \mathbb{P}\left(\forall p \text{ s.t.} ||p||^2_{G_N} \in [\lambda_k, \lambda_{k+1}) \quad \frac{|E_N^\top \Phi_N p|^2}{||p||^2_{G_N}} > \varepsilon_N\right)$$

2. Pointwise concentration We now provide a point-wise concentration result that holds for a single $p \in \mathscr{S}^d$. We use the fact that for all $\lambda \in \mathbb{R}^+$ such that $\lambda p^\top \varphi(x_n) \in \mathscr{D}_{\xi_n}$, which we denote for convenience $\lambda \in \mathbb{R}^+ \cap \frac{1}{p^\top\varphi(x_n)}\mathscr{D}_{\xi_n}$,

$$\ln \mathbb{E}[\exp(\lambda \xi_n p^\top \varphi(x_n))|h_{n-1}] \leqslant \psi(\lambda p^\top \varphi(x_n))\,.$$

Applying this inequality to the random variable $Z = \sum_{n=1}^m \varphi(x_n)\xi_n$, for a deterministic $m \in \mathbb{N}$ it holds

$$\ln \mathbb{E}\left[\exp(\lambda p^\top Z - \sum_{n=1}^m \psi(\lambda p^\top \varphi(x_n))\right] \leqslant 0\,.$$

In the special case when ξ_n is R-sub-Gaussian conditionally on $\mathscr{H}_{n-1}$, then $\psi(\lambda) = \frac{\lambda^2 R^2}{2}$ and $\mathscr{D}_{\xi_n} = \mathbb{R}$. In that case, the previous inequality reduces to

$$\ln \mathbb{E}\bigg[\exp(\lambda p^\top Z - \frac{\lambda^2 R^2}{2} ||p||^2_{G_m}) \bigg] \leqslant 0\,.$$

Now, the same inequality holds for m being replaced by the stopping time N, using a standard stopping time argument. However, we cannot deduce that for all $\delta \in (0,1)$, $x \in \mathscr{S}^d$,

$$\mathbb{P}\bigg(p^\top \Phi_N^\top E_N \geqslant \inf_{\lambda \in \mathbb{R}^+} \bigg\{ \frac{\lambda R^2}{2} p^\top \sum_{n=1}^N \varphi_n \varphi_n^\top p + \frac{\ln(1/\delta)}{\lambda} \bigg\} \bigg) \not\leqslant \delta\,,$$

which would lead to the desirable

$$\mathbb{P}\bigg(p^\top \Phi_N^\top E_N \geqslant \sqrt{2R^2 \ln(1/\delta)} ||p||_{G_N} \bigg) \not\leqslant \delta\,.$$

The reason is that the λ achieving the optimum in the above expression explicitly depends on $\mathscr{H}_N$ via G_N and the control of the log-Laplace transform only holds for λ adapted to the filtration (and thus in particular should be $\mathscr{H}_0$-measurable).

To overcome this difficulty, we resort to the Laplace method, for the 1-dimensional variable $A = p^\top \Phi_N^\top E_N \in \mathbb{R}$ and the 1×1 random matrix $B = R^2 ||p||^2_{G_N}$. It thus holds for any $c \in \mathbb{R}^+_\star$, $\delta \in [0,1]$,

$$\mathbb{P}\bigg(\frac{|p^\top \Phi_N^\top E_N|^2}{||p||^2_{G_N} + c} \geqslant 2R^2 \ln \left(\frac{\sqrt{1 + ||p||^2_{G_N}/c}}{\delta} \right) \bigg) \leqslant \delta\,.$$

At this point, we use the localization and focus on the event $||p||^2_{G_N} \in [\lambda_k, \lambda_{k+1})$. For each such p, choosing $c = \kappa \lambda_k \leqslant \kappa ||p||^2_{G_N}$ for some $\kappa > 0$ leads to the bound

$$\mathbb{P}\bigg(|p^\top \Phi_N^\top E_N|^2 \geqslant 2R^2 \ln \left(\frac{\sqrt{1 + \frac{\lambda_{k+1}}{\lambda_k \kappa}}}{\delta} \right) (1+\kappa) ||p||^2_{G_N} \bigg) \leqslant \delta\,.$$

Making the expression $\lambda_k = (1+\eta)^k \lambda$ explicit, and using $\kappa = 1/2$, it comes

$$\mathbb{P}\bigg(||p||^2_{G_N} \in [\lambda_k, \lambda_{k+1}) \cap \frac{|p^\top \Phi_N^\top E_N|^2}{||p||^2_{G_N}} \geqslant 3R^2 \ln \left(\frac{\sqrt{3+2\eta}}{\delta} \right) \bigg) \leqslant \delta\,.$$

---

3. Parameter tuning for bounded $||p||_{G_N}$. We thus deduce a first result: Let $\Lambda > \lambda$ and $K = \lceil \frac{\ln(\Lambda/\lambda)}{\ln(1+\eta)} \rceil$. Then using a union bound, we deduce that

$$\mathbb{P}\bigg(||p||^2_{G_N} \in [\lambda, \Lambda) \cap \frac{|p^\top \Phi_N^\top E_N|^2}{||p||^2_{G_N}} \geqslant 3R^2 \ln \left(\frac{K\sqrt{3+2\eta}}{\delta} \right) \bigg) \leqslant \delta\,.$$

We can now optimize the quantity $K\sqrt{3+2\eta}$ in η. We choose $\eta = e^2 - 1$ for which $\sqrt{3+2\eta} \leqslant 4$ and obtain that for all $\delta \in [0,1]$, it holds

$$\mathbb{P}\left(||p||^2_{G_N} \in [\lambda, \Lambda) \cap \frac{|p^\top \Phi_N^\top E_N|^2}{||p||^2_{G_N}} \geqslant 3R^2 \ln\left(\frac{4\lceil \ln(\Lambda/\lambda)/2 \rceil}{\delta}\right)\right) \leqslant \delta\,. \tag{7}$$

3b. Parameter tuning for unbounded $||p||_{G_N}$. We now consider the situation when no a priori upper bound is known on $||p||_{G_N}$. In that case, we can choose $\delta_k = \delta \frac{6}{\pi^2 k^2}$ for the confidence associated to the k^{th} localization slice $||p||^2_{G_N} \in [\lambda_k, \lambda_{k+1})$ We find that $k \leqslant k_p = \frac{\ln(||p||^2_{G_N}/\lambda)}{\ln(1+\eta)} \leqslant k_p + 1$, and thus we obtain,

$$\mathbb{P}\left(||p||^2_{G_N} \geqslant \lambda \cap \frac{|p^\top \Phi_N^\top E_N|^2}{||p||^2_{G_N}} \geqslant 3R^2 \ln\left(\frac{\pi^2 k_p^2 \sqrt{3+2\eta}}{6\delta}\right)\right) \leqslant \sum_{k=0}^{\infty} \delta_k = \delta\,,$$

where we used the fact that $\sum_{k=0}^{\infty} \delta_k = \delta$. Replacing k_p with its expression, it comes

$$\mathbb{P}\left(||p||^2_{G_N} \geqslant \lambda \cap \frac{|p^\top \Phi_N^\top E_N|^2}{||p||^2_{G_N}} \geqslant 3R^2 \ln\left(\frac{\pi^2 \ln(||p||^2_{G_N}/\lambda)^2 \sqrt{3+2\eta}}{6\delta \ln(1+\eta)^2}\right)\right) \leqslant \delta\,.$$

Choosing $\eta = 55.5$ that approximately minimizes $\frac{\sqrt{3+2\eta}}{6\ln(1+\eta)^2} \leqslant 0.11$, it comes $\forall \delta \in [0,1]$

$$\mathbb{P}\left(||p||^2_{G_N} \geqslant \lambda \cap \frac{|p^\top \Phi_N^\top E_N|^2}{||p||^2_{G_N}} \geqslant 3R^2 \ln\left(\frac{0.11\pi^2 \ln(||p||^2_{G_N}/\lambda)^2}{\delta}\right)\right) \leqslant \delta\,. \tag{8}$$

---

4. A covering argument The next step is to move from concentration results for each separate p to result holding simultaneously for all p. To this end, we resort to a covering argument. Let us consider a covering $\mathscr{C}$ of $\mathscr{S}^d$ for the Euclidean norm, such that $\forall p \in \mathscr{S}^d, \exists q_p \in \mathscr{C} \subset \mathscr{S}^d : ||p - q_p|| \leqslant h$. Let M_h denotes the size of the cover. It can be shown that $M_h \leqslant \left(\frac{2\sqrt{d}}{h}\right)^d$.

We study the quadratic form $g(p) = p^\top G_N p$, that satisfies for points $p, q \in \mathscr{S}$ such that $||p - q|| \leqslant h$ the following inequality

$$g(p) - g(q) = (p-q)^\top G_N (p+q) \leqslant 2h\lambda_{\max}(G_N)\,.$$

Now the quadratic form $f(p) = p^\top \Phi_N^\top E_N E_N^\top \Phi_N p$ satisfies

$$\begin{aligned} f(p) - f(q_p) &= (p-q_p)^\top \Phi_N^\top E_N E_N^\top \Phi_N (p+q_p) \\ &\leqslant ||p-q_p||||p+q_p||||\Phi_N^\top E_N||^2 \\ &\leqslant 2h||\Phi_N^\top E_N||^2 \\ &= 2h\sum_{i=1}^{d} f(e_i)\,, \end{aligned}$$

where e_i denotes the i^{th} unit canonical vector of $\mathbb{R}^d$. Using these notations, for each $q \in \mathscr{C} \cup \{e_1, \dots, e_d\}$ it has been proved in part 1 that for deterministic λ, Λ,

$$\mathbb{P}\Big(g(q) \in [\lambda, \Lambda] \cap f(q) > \varepsilon_{\lambda,\Lambda,\delta} g(q) \Big) \leqslant \delta \,,$$

for any δ where $\varepsilon_{\lambda,\Lambda,\delta} = 3R^2 \ln\left(\frac{2\ln(\Lambda/\lambda)+5}{\delta}\right)$. Then we deduce that with probability at least $1-(M_h+d)\delta$, it holds $\forall p \in \mathscr{S}^d$

$$\begin{aligned} f(p) &\leqslant f(q_p) + 2h||\Phi_N^\top E_N||^2 \\ &\leqslant \varepsilon_{\lambda,\Lambda,\delta} g(q_p) + 2h \sum_{i=1}^d \varepsilon_{\lambda,\Lambda,\delta} g(e_i) \\ &\leqslant \varepsilon_{\lambda,\Lambda,\delta} \Big(g(q_p) + 2hd\lambda_{\max}(G_N) \Big) \\ &\leqslant \varepsilon_{\lambda,\Lambda,\delta} \Big(g(p) + 2h\lambda_{\max}(G_N)(d+1) \Big) \,. \end{aligned}$$

Now, using that $g(p) \leqslant \lambda$ and considering that $\lambda_{\max}(G_N) \leqslant \Lambda$, we thus choose $h = \frac{\alpha\lambda}{2\Lambda(d+1)}$. In this case we deduce that simultaneously for all $p \in \mathscr{S}^d$, then $f(p) \leqslant \varepsilon_{\lambda,\Lambda,\delta}(1+\alpha) g(p)$ with probability $1-(M_h+d)\delta$, and thus

$$\begin{aligned} \mathbb{P}\Bigg(\lambda_{\max}(G_N) \leqslant \Lambda \cap \sup_{p:||p||^2_{G_N} \in [\lambda,\Lambda]} \frac{|E_N^\top \Phi_N p|^2}{||p||_A^2} > 3(1+\alpha) R^2 \ln\left(\frac{4\lceil \ln(\Lambda/\lambda)/2 \rceil}{\delta} \right) \Bigg) \\ \leqslant (M_h+d)\delta \leqslant \Big((\frac{4(d+1)\sqrt{d}}{\alpha} \frac{\Lambda}{\lambda})^d + d \Big) \delta \,. \end{aligned}$$

Choosing $\alpha = 1/3$ and introducing $K_d = (12(d+1)\sqrt{d})^d$, we obtain the following bound

$$\mathbb{P}\Bigg(\lambda_{\max}(G_N) \leqslant \Lambda \cap \sup_{p:||p||^2_{G_N} \in [\lambda,\Lambda]} \frac{|E_N^\top \Phi_N p|^2}{||p||_{G_N}^2} > 4R^2 \ln\left(\frac{(K_d(\frac{\Lambda}{\lambda})^d + d) 4\lceil \ln(\Lambda/\lambda)/2 \rceil}{\delta} \right) \Bigg) \leqslant \delta \,.$$

---

5. A peeling argument for the eigenvalues

We now localize $\lambda_{\max}(G_N) \in [\Lambda_j, \Lambda_{j+1})$, with $\Lambda_j = \lambda(1+\kappa)^j$. We first consider the case when $\lambda_{\max}(G_N) \leqslant \Lambda$ almost surely, and thus set $J = \lceil \frac{\ln(\Lambda/\lambda)}{\ln(1+\kappa)} \rceil$.

Using a union bound over $j = 0..J-1$ and the fact that on the event $\lambda_{\max}(G_N) \in [\Lambda_j, \Lambda_{j+1})$, then it holds $\Lambda_{j+1} = (1+\kappa)\Lambda_j \leqslant (1+\kappa)\lambda_{\max}(G_N)$, and thus obtain

$$\begin{aligned} \mathbb{P}\Bigg(\sup_{p:||p||^2_{G_N} \geqslant \lambda} \frac{|E_N^\top \Phi_N p|^2}{||p||_{G_N}^2} > \\ 4R^2 \ln\left(\frac{(K_d(\frac{(1+\kappa)\lambda_{\max}(G_N)}{\lambda})^d + d) 4\lceil \ln((1+\kappa)\lambda_{\max}(G_N)/\lambda)/2 \rceil}{\delta} \right) \Bigg) \leqslant \lceil \frac{\ln(\Lambda_J/\lambda)}{\ln(1+\kappa)} \rceil \delta \,. \end{aligned}$$

We choose $\kappa = e-1$ for simplicity and obtain

$$\mathbb{P}\Bigg(\sup_{p:||p||^2_{G_N} \geqslant \lambda} \frac{|E_N^\top \Phi_N p|^2}{||p||^2_{G_N}} > 4R^2 \ln\left(\frac{\lceil \ln(\Lambda_J/\lambda)\rceil (K_d(\frac{e\lambda_{\max}(G_N)}{\lambda})^d + d) 4\lceil \ln(e\lambda_{\max}(G_N)/\lambda)/2\rceil}{\delta} \right) \Bigg) \leqslant \delta\,.$$

Now if no bound is known on $\lambda_{\max}(G_N)$, we set $\delta_j = \delta \frac{6}{\pi^2 j^2}$ for the confidence level associated to $\lambda_{\max}(G_N) \in [\Lambda_j, \Lambda_{j+1})$. Thus, by using a union bound over the peeling events, and the fact that $j \leqslant j_N \overset{\text{def}}{=} \frac{\ln(\lambda_{\max}(G_N)/\lambda)}{\ln(1+\kappa)}$ on each of these events, we obtain

$$\mathbb{P}\Bigg(\sup_{p:||p||^2_{G_N} \geqslant \lambda} \frac{|E_N^\top \Phi_N p|^2}{||p||^2_{G_N}} > 4R^2 \ln\left(\frac{\pi^2 j_N^2 (K_d(\frac{(1+\kappa)\lambda_{\max}(G_N)}{\lambda})^d + d) 4\lceil \ln((1+\kappa)\lambda_{\max}(G_N)/\lambda)/2\rceil)}{6\delta} \right) \Bigg) \leqslant \delta\,.$$

Thus, choosing $\kappa = e-1$, it comes

$$\mathbb{P}\Bigg(\sup_{p:||p||^2_{G_N} \geqslant \lambda} \frac{|E_N^\top \Phi_N p|^2}{||p||^2_{G_N}} > 4R^2 \ln\left(\frac{2\pi^2 \ln(\lambda_{\max}(G_N)/\lambda)^2 \lceil \ln(e\lambda_{\max}(G_N)/\lambda)/2\rceil (K_d(\frac{e\lambda_{\max}(G_N)}{\lambda})^d + d)}{3\delta} \right) \Bigg) \leqslant \delta\,.$$

□

Corollary 5.5 (Ordinary Least-squares) *Let us assume that N is a stopping time adapted to the filtration of the past. Then it holds*

$$\mathbb{P}\left(\exists x \in \mathscr{X}\ |f^\star(x) - f_{\theta_N^\dagger}(x)| \geqslant 2R||\varphi(x)||_{G_N^\dagger} \sqrt{\ln\left(\frac{\kappa_d(e^2\Lambda_N^2)}{\delta}\right)} \cap\ \lambda_{\min}(G_N) > 0 \right) \leqslant \delta.$$

Proof of Corollary 5.5:

Indeed, let $\theta_N^\dagger$ denote the specific-pseudo-inverse solution. Then it holds

$$\begin{aligned} |f^\star(x) - f_{\theta_N^\dagger}(x)| &= |(\theta^\star - \theta_N^\dagger)^\top \varphi(x)| \\ &\leqslant |(\theta^\star - G_N^\dagger \Phi_N^\top Y_N)^\top \varphi(x)| \\ &\leqslant |((I - G_N^\dagger G_N)\theta^\star - G_N^\dagger \Phi_N^\top E_N)^\top \varphi(x)| \\ &\leqslant |\theta^{\star\top}(I - G_N^\dagger G_N)^\top \varphi(x)| + |E_N^\top \Phi_N G_N^\dagger \varphi(x)|\,. \end{aligned}$$

Then, we control the first term in the bound by Cauchy-Schwarz's inequality

$$|\theta^{\star\top}(I - G_N^\dagger G_N)^\top \varphi(x)| \leqslant ||\theta^\star|| ||(I - G_N^\dagger G_N)^\top \varphi(x)||_2 \,,$$

Now on the event $\lambda_{\min}(G_N) > 0$, it holds $I = G_N^\dagger G_N$.

For the second term, using that $G_N^\dagger G_N G_N^\dagger = G_N^\dagger$, we obtain

$$\begin{aligned} |E_N^\top \Phi_N G_N^\dagger \varphi(x)| &\leqslant \sup_{p\in\mathbb{R}^D} \frac{p^\top \Phi_N^\top E_N}{||p||_{G_N}} ||G_N^\dagger \varphi(x)||_{G_N} \\ &= \sup_{p\in\mathscr{S}^D} \frac{p^\top \Phi_N^\top E_N}{||p||_{G_N}} ||\varphi(x)||_{G_N^\dagger} \,. \end{aligned}$$

Now, note that if $\lambda_{\min}(G_N) > 0$, then it holds $||\varphi(x)||^2_{G_N^\dagger} \geqslant ||\varphi(x)||^2 \lambda_{\min}(G_N^\dagger) \geqslant \frac{||\varphi(x)||^2}{\Lambda_N}$. We can thus apply Theorem 5.4, and more precisely

$$\mathbb{P}\left(\sup_{p:||p||^2_{G_N} \geqslant \Lambda_N^{-1}} \frac{|E_N^\top \Phi_N p|^2}{||p||^2_{G_N}} > 4R^2 \ln\left[\frac{\kappa_d(e^2\Lambda_N^2)}{\delta}\right]\right) \leqslant \delta \,,$$

in order to deduce that

$$\mathbb{P}\left(\exists x \in \mathscr{X} \, |f^\star(x) - f_{\theta_N^\dagger}(x)| \geqslant 2R||\varphi(x)||_{G_N^\dagger} \sqrt{\ln\left(\frac{\kappa_d(e^2\Lambda_N^2)}{\delta}\right)} \cap \ \lambda_{\min}(G_N) > 0 \right) \leqslant \delta.$$

□

Corollary 5.6 (Regularized Least-squares) *Let us assume that $||\theta^\star||_2 \leqslant B$, and that N is a stopping time adapted to the filtration of the past. Then it holds for all deterministic $\lambda > 0$*

$$\mathbb{P}\left(\exists x \in \mathscr{X} \, |f^\star(x) - f_{\theta_{N,\lambda}}(x)| \geqslant ||\varphi(x)||_{G_{N,\lambda}^{-1}} \left[R\sqrt{2\ln\left(\frac{det(G_N+\lambda I)^{1/2}}{\delta det(\lambda I)^{1/2}}\right)} + \frac{\lambda B}{\sqrt{\lambda_{\min}(G_{N,\lambda})}} \right] \right) \leqslant \delta.$$

Proof of Corollary 5.6:

Indeed, we first use the decomposition

$$\begin{aligned} |f^\star(x) - f_{\theta_{N,\lambda}}(x)| &= |(\theta^\star - \theta_{N,\lambda})^\top \varphi(x)| \\ &= |(\theta^\star - G_{N,\lambda}^{-1} \Phi_N^\top Y_N)^\top \varphi(x)| \\ &= |((I - G_{N,\lambda}^{-1} G_N)\theta^\star - G_{N,\lambda}^{-1} \Phi_N^\top E_N)^\top \varphi(x)| \\ &\leqslant |((I - G_{N,\lambda}^{-1} G_N)\theta^\star)^\top \varphi(x)| + |E_N^\top \Phi_N G_{N,\lambda}^{-1} \varphi(x)| \,. \end{aligned}$$

Now, we handle the first term in the right-hand side of the previous inequality by

$$\begin{aligned} |((I - G_{N,\lambda}^{-1} G_N)\theta^\star)^\top \varphi(x)| &= |\lambda \theta^{\star\top} G_{N,\lambda}^{-1} \varphi(x)| \\ &\leqslant \frac{\lambda}{\sqrt{\lambda_{\min}(G_{N,\lambda})}} ||\theta^\star||_2 ||\varphi(x)||_{G_{N,\lambda}^{-1}} \end{aligned}$$

We then turn to the second term, which we control by using the inequality

$$|E_N^\top \Phi_N G_{N,\lambda}^{-1} \varphi(x)| \leqslant ||\Phi_N^\top E_N||_{G_{N,\lambda}^{-1}} ||\varphi(x)||_{G_{N,\lambda}^{-1}} ,$$

and then by applying the last part of Theorem 5.4. □

Corollary 5.7 (Kernel Least-squares) *Let us assume that N is a stopping time adapted to the filtration of the past. Then for all $x \in \mathcal{X}$ it holds*

$$\mathbb{P}\Bigg(|f^\star(x) - \mu_N(x)| \geqslant |f^\star(x) - k_N(x)^\top (\mathbf{K}_N + \sigma^2 I_N)^{-1} \mathbf{f}_N| + R\sqrt{2\Big(||k_N(x)||^2_{(\mathbf{K}_N + \sigma^2 I_N)^{-2}} + 1\Big) \ln\Big(\frac{\sqrt{1 + ||k_N(x)||^2_{(\mathbf{K}_N+\sigma^2 I_N)^{-2}}}}{\delta}\Big)} \Bigg) \leqslant \delta.$$

Proof of Corollary 5.7:

Indeed, we first use the decomposition

$$|f^\star(x) - \mu_N(x)| \leqslant |f^\star(x) - k_N(x)^\top (\mathbf{K}_N + \sigma^2 I_N)^{-1} \mathbf{f}_N| + |k_N(x)^\top (\mathbf{K}_N + \sigma^2 I_N)^{-1} E_N| .$$

where $\mathbf{f}_N = (f^\star(x_1), \dots f^\star(x_N))$. Now, if we denote $p_N = (\mathbf{K}_N + \sigma^2 I_N)^{-1} k_N(x)$, and introduce for $\lambda \in \mathbb{R}$ and a deterministic $m \in \mathbb{N}$ the quantity $M_{m,\lambda} = \exp\Big(\lambda p_m^\top E_m - \frac{\lambda^2 R^2}{2} p_m^\top p_m\Big)$, then it turns out that $\{M_{m,\lambda}\}_{m=0}^\infty$ is a non-negative supermartingale. Indeed $p_m^\top E_m = \sum_{i=1}^{m-1} p_{m,i}\xi_i + p_{m,m}\xi_m$, with $p_{m,m}$ being a measurable function of the observations before step m. Thus,

$$\begin{aligned} \mathbb{E}\Big[\exp\Big(\lambda p_m^\top E_m - \frac{\lambda^2 R^2}{2} p_m^\top p_m\Big)\Big| h_{m-1}\Big] &\leqslant M_{m-1,\lambda} \mathbb{E}\Big[\lambda p_{m,m}\xi_m - \lambda^2 R^2 p_{m,m}^2/2\Big| h_{m-1}\Big] \\ &\leqslant M_{m-1,\lambda} . \end{aligned}$$

Using a standard stopping time argument, we thus have $\ln \mathbb{E}[M_{N,\lambda}] \leqslant 0$. We can then apply the Laplace method for the control of the concentrations. We deduce that for all $c \in \mathbb{R}_\star^+$, and $\delta \in [0,1]$, then

$$\mathbb{P}\Bigg(|p_N^\top E_N| \geqslant R\sqrt{2\Big(||p_N||^2 + c^2\Big) \ln\Big(\frac{\sqrt{c^2 + ||p_N||^2}}{\delta c}\Big)} \Bigg) \leqslant \delta .$$

□

5.2. Variance estimates

We now turn to the estimate of the variance. It is tempting to apply the self-normalized concentration result for stopping time of vector-valued martingales in order to control $||E_N||_2$. Indeed, if we introduce a $N \times N$ matrix $\Phi_N = I_N$, and $V = \lambda I$, then $||\Phi^\top E_N||_{\Phi_N^\top \Phi_N + V}] = \sqrt{1+\lambda}||E_N||_2$. However, Lemma 5.1 does not apply for two reasons. First, it is written for a dimension D that is deterministic, not a random stopping time, thus it cannot be applied directly. More importantly, the supermartingale property that holds does not satisfy the sub-Gaussian shape used for the result to apply. We thus proceed differently, and apply a more general result for real-valued distributions, given by Lemma 5.3.

Let us first recall a classical concentration result, slightly extended beyond the iid case.

Lemma 5.8 *Let us assume that the noise sequence $\{\xi_i\}_{i=1}^n$ (where n is deterministic) is conditionally strongly R-subGaussian, in the sense that*

$$\forall \lambda < 1/2R^2 \qquad \ln \mathbb{E}[\exp(\lambda \xi_i^2)|h_{i-1}] \leqslant -\frac{1}{2}\ln(1-2R^2\lambda)\,. \tag{9}$$

Then for all $\delta \in (0,1]$, it holds

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \xi_i^2 \geqslant R^2 + 2R^2\sqrt{\frac{2\ln(1/\delta)}{n}} + 2R^2\frac{\ln(1/\delta)}{n}\right) \leqslant \delta\,.$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \xi_i^2 \leqslant R^2 - 2R^2\sqrt{\frac{\ln(1/\delta)}{n}}\right) \leqslant \delta\,.$$

Further, for all $\delta \in (e^{-n}, 1]$, it holds

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \xi_i^2 \leqslant R^2 - 2R^2\sqrt{\frac{\ln(1/\delta)}{n}} + R^2\frac{\ln(1/\delta)}{n}\right) \leqslant \delta\,.$$

Note that inequality (9) becomes an equality in the case of Gaussian random variables. The first part of this result, in the i.i.d. case is known as a Birgé-Massart lemma.

Proof of Lemma 5.8:

Indeed, first, an immediate result is the following

Lemma 5.9 *Let us consider n random variables ξ_i, $i \in [n]$ satisfying (9). Then for all $\delta \in (0,1)$, it holds*

$$\mathbb{P}\left(\sum_{i=1}^n \xi_i^2 \geqslant \inf_{\lambda \in (0, \frac{1}{2R^2})} -\frac{n}{2\lambda}\ln(1-2R^2\lambda) + \frac{\ln(1/\delta)}{\lambda}\right) \leqslant \delta\,.$$

$$\mathbb{P}\left(\sum_{i=1}^n \xi_i^2 \leqslant \sup_{\lambda > 0} \frac{n}{2\lambda}\ln(1+2R^2\lambda) - \frac{\ln(1/\delta)}{\lambda}\right) \leqslant \delta\,.$$

Note that this result holds for a deterministic n, and does not extend trivially to the case of a random stopping time.

In order to get a more explicit result, let us look at the two optimization problems. We aim at having an upper bound on the infimum term and a lower bound on the supremum term. To this end, we use the following properties of the logarithm

$$\forall x \in (0,1), \qquad \frac{x}{1-x} \overset{(1)}{\geqslant} x + \frac{x^2}{2(1-x)} \overset{(2)}{\geqslant} -\ln(1-x) \quad \left(\geqslant \frac{x}{1-x/2} \right)$$
$$\forall x > 0, \qquad \ln(1+x) \overset{(3)}{\geqslant} \frac{x}{1+x/2} \overset{(4)}{\geqslant} x - x^2/2\,.$$

More precisely, by making use of inequality (1) and (3), we respectively get to solve the problems

$$\inf_{\lambda \in (0, \frac{1}{2R^2})} \frac{nR^2}{1-2R^2\lambda} + \frac{\ln(1/\delta)}{\lambda} \quad \text{and} \quad \sup_{\lambda > 0} \frac{nR^2}{1+R^2\lambda} - \frac{\ln(1/\delta)}{\lambda}\,.$$

Instead of using (1), one could use the slightly tighter inequality (2). This however leads to more complex approximations. Now, solving the infimum in λ gives $\lambda = \frac{1}{2R^2(1+\sqrt{\frac{n}{2\ln(1/\delta)}})} \in (0, \frac{1}{2R^2})$ and then

$$\frac{nR^2}{1-2R^2\lambda} + \frac{\ln(1/\delta)}{\lambda} = nR^2 + 2R^2\sqrt{2n\ln(1/\delta)} + 2R^2\ln(1/\delta)\,.$$

Solving the supremum in λ gives $\lambda = \frac{1}{R^2(\sqrt{\frac{n}{\ln(1/\delta)}}-1)}$, which is however positive only if $\delta > e^{-n}$. The corresponding value is

$$\frac{nR^2}{1+R^2\lambda} - \frac{\ln(1/\delta)}{\lambda} \quad = \quad R^2 n - 2R^2\sqrt{n\ln(1/\delta)} + R^2\ln(1/\delta)\,.$$

If we instead consider the (slightly sub-optimal) value $\lambda = \frac{1}{R^2}\sqrt{\frac{\ln(1/\delta)}{n}}$, and use the inequality $\frac{1}{1+x} \geqslant 1-x$ for $x \in \mathbb{R}^+$ we get

$$\frac{R^2 n}{1+\sqrt{\frac{\ln(1/\delta)}{n}}} - R^2\sqrt{n\ln(1/\delta)} \geqslant R^2 n - 2R^2\sqrt{n\ln(1/\delta)}\,.$$

□

In order to be able to handle a random stopping time, we apply a slightly different construction, and obtain the following result, at the price of loosing a $\ln\ln$ factor compared to the result without stopping time. A similar construction was given in Cappé, Garivier and Maillard (2013) for a related but different problem.

Lemma 5.10 *Assume that N_n is a random stopping time that satisfies $N_n \leqslant n$ almost surely, then it holds*

$$\mathbb{P}\Big[\frac{1}{N_n}\sum_{i=1}^{N_n}\xi_i^2 \geqslant R^2 + 2R^2\sqrt{\frac{2\ln(e/\delta)}{N_n}} + 2R^2\frac{\ln(e/\delta)}{N_n}\Big] \leqslant \Big(\lceil \ln(n)\ln(e/\delta)\rceil\Big)\delta$$

$$\mathbb{P}\Big[\frac{1}{N_n}\sum_{i=1}^{N_n}\xi_i^2 \leqslant R^2 - 2R^2\sqrt{\frac{\ln(e/\delta)}{N_n}}\Big] \leqslant \Big(\lceil \ln(n)\ln(e/\delta)\rceil\Big)\delta$$

Further, for a random stopping time N, and if we introduce $c_N = \ln(\pi^2\ln^2(N)/6)$, then it holds

$$\mathbb{P}\Big[\frac{1}{N}\sum_{i=1}^{N}\xi_i^2 \geqslant R^2 + 2R^2\sqrt{\frac{2\ln(e/\delta)(1+c_N/\ln(1/\delta))}{N}} + 2R^2\frac{\ln(e/\delta)(1+c_N/\ln(1/\delta))}{N}\Big] \leqslant \delta$$

$$\mathbb{P}\Big[\frac{1}{N}\sum_{i=1}^{N}\xi_i^2 \leqslant R^2 - 2R^2\sqrt{\frac{\ln(e/\delta)(1+c_N/\ln(1/\delta))}{N}}\Big] \leqslant \delta$$

Proof of Lemma 5.10:

According to Lemma 5.3 applied to $Z_i = \xi_i^2$, all we have to do is to compute an upper bound on the quantity $\varphi_{\star,+}^{-1}(c)$, first for the value $c = \frac{\ln(e/\delta)}{N_n}$, then for $c = \frac{\ln(e/\delta)}{N}(1 + \frac{2}{\ln(1/\delta)}\ln\Big(\frac{\pi\ln(N)\ln(1/\delta)}{6^{1/2}(1+\ln(1/\delta))}\Big)) \leqslant \frac{\ln(e/\delta)}{N}(1 + c_N/\ln(1/\delta))$. We proceed in the following way. First, the envelope function is given by

$$\varphi(\lambda) = -\frac{1}{2}\ln(1-2\lambda R^2) \leqslant \frac{\lambda R^2}{1-2\lambda R^2}\,.$$

Thus, $\varphi^\star(x) \geqslant \sup_\lambda[\lambda x - \frac{\lambda R^2}{1-2\lambda R^2}]$. Solving this optimization by differentiating over λ, the supremum is reached for $\lambda = (1-\frac{R}{\sqrt{x}})\frac{1}{2R^2} < \frac{1}{2R^2}$, with corresponding value given by

$$\begin{aligned}\tilde{\varphi}^\star(x) &= (1-\frac{R}{\sqrt{x}})\frac{x}{2R^2} - (1-\frac{R}{\sqrt{x}})\frac{\sqrt{x}}{2R}\\ &= \frac{x}{2R^2} - \frac{\sqrt{x}}{R} + \frac{1}{2}\,.\end{aligned}$$

Now, for $c > 0$, it is easily checked that $\tilde{\varphi}^\star(x) = c$ holds for $x_c = R^2(1+\sqrt{2c})^2$. As a result, we deduce that $\varphi_{\star,+}^{-1}(c) \leqslant R^2(1+\sqrt{2c})^2 = R^2 + 2R^2c + 2R^2\sqrt{2c}$.

Now, for the reverse inequality, we have to compute a lower bound on the quantity $\varphi_{\star,-}^{-1}(c)$, first for $c = \frac{\ln(e/\delta)}{N_n}$, then for $c = \frac{\ln(e/\delta)}{N}(1 + \frac{2}{\ln(1/\delta)}\ln\Big(\frac{\pi\ln(N)\ln(1/\delta)}{6^{1/2}(1+\ln(1/\delta))}\Big)) \leqslant$

$\frac{\ln(e/\delta)}{N}(1 + c_N/\ln(1/\delta))$. We proceed in the following way. First, the envelope function is given for $\lambda > 0$ by

$$\varphi(-\lambda) = -\frac{1}{2}\ln(1 + 2\lambda R^2) \geqslant -\frac{\lambda R^2}{1 + \lambda R^2}\,.$$

Thus, for $0 < x < R^2$ it holds $\varphi^\star(x) \geqslant \sup_{\lambda>0}[-\lambda x + \frac{\lambda R^2}{1+\lambda R^2}] = 1 + \sup_{\lambda>0}[-\lambda x - \frac{1}{1+\lambda R^2}]$. Solving this optimization by differentiating over λ, the supremum is reached for $\lambda = \frac{1}{R^2}(\frac{R}{\sqrt{x}} - 1) > 0$ with corresponding value given by

$$\begin{aligned}\tilde{\varphi}^\star(x) &= 1 - \frac{x}{R^2}\big(\frac{R}{\sqrt{x}} - 1\big) - \frac{\sqrt{x}}{R} \\ &= \frac{x}{R^2} - 2R\frac{\sqrt{x}}{R} + 1\,.\end{aligned}$$

Now, for $c > 0$, it is easily checked that $\tilde{\varphi}^\star(x) = c$ holds for $x_c = R^2(1 - \sqrt{c})^2$, and $x_c < R^2$ if $c < 1$. As a result, we deduce that if $c \in (0, 1)$, then $\varphi^{-1}_{\star,-}(c) \geqslant R^2 - 2R^2\sqrt{c} + R^2 c$. On the other hand, for all $c > 0$, choosing $\lambda = \frac{1}{R^2}\sqrt{c}$, and using the inequality $\frac{1}{1+v} \geqslant 1 - v$ for $v > 0$, then

$$\begin{aligned}\varphi^\star(x) &\geqslant -\frac{x}{R^2}\sqrt{c} + 1 - \frac{1}{1 + \sqrt{c}} = \sqrt{c}\big(-\frac{x}{R^2} + \frac{1}{1 + \sqrt{c}}\big) \\ &\geqslant \tilde{\varphi}^\star(x) \stackrel{\text{def}}{=} \sqrt{c}\big(-\frac{x}{R^2} + 1 - \sqrt{c}\big)\end{aligned}$$

Thus, $\tilde{\varphi}^\star(x) = c$ for $x_c = R^2 - 2R^2\sqrt{c} < R^2$. As a result, we deduce that if $c > 0$, then $\varphi^{-1}_{\star,-}(c) \geqslant R^2 - 2R^2\sqrt{c}$. □

We now proceed with the concentration results for the variance estimates in the regression setting with ordinary least-squares and regularized least squares.

Theorem 3.8 (Ordinary variance estimate) *Under the conditionally strongly-sub-Gaussian assumption, for any random stopping time N with respect to the filtration of the past, then, with probability higher than $1 - 3\delta$, either $\lambda_{\min}(G_N) < \lambda_0$ or*

$$R\bigg(1 - \sqrt{\frac{C_N(\delta)}{N}} - \sqrt{\frac{C_N(\delta)}{N} + \frac{D_N(\delta)}{N}}\bigg) \leqslant \sqrt{\widehat{\sigma}^2_N} \leqslant R\bigg(1 + \sqrt{\frac{2C_N(\delta)}{N}}\bigg)\,.$$

where $D_N(\delta) = 4\ln(\kappa_d(e\Lambda_N/\lambda_0)/\delta)$ and $C_N(\delta) = \ln(e/\delta)\big[1 + \ln(\pi^2\ln(N)/6)/\ln(1/\delta)\big]$.

Proof of Theorem 3.8:

Let $E_N \in \mathbb{R}^N$ be the vector with components $(\varepsilon_1, \dots, \varepsilon_N)$. It holds that

$$\begin{aligned}\sum_{n=1}^N (y_n - \langle \theta_N, \varphi(x_n)\rangle)^2 &= \sum_{n=1}^N (\langle \theta^\star - \theta_N, \varphi(x_n)\rangle + \varepsilon_n)^2 \\ &= (\theta^\star - \theta_N)^\top \sum_{n=1}^N \varphi(x_n)\varphi(x_n)^\top (\theta^\star - \theta_N) \\ &\quad + \sum_{n=1}^N \varepsilon_n^2 + 2(\theta^\star - \theta_N)^\top \sum_{n=1}^N \varphi(x_n)\varepsilon_n \\ &= (\theta^\star - \theta_N)^\top G_N (\theta^\star - \theta_N) + ||E_N||^2 + 2(\theta^\star - \theta_N)^\top \Phi_N^\top E_N \,.\end{aligned}$$

Now, we notice that this can be rewritten, as

$$\sum_{n=1}^N (y_n - \langle \theta_N^\dagger, \varphi(x_n)\rangle)^2 = ||E_N||^2 - ||E_N^\top \Phi_N||^2_{G_N^\dagger} + 2\theta^{\star\top}(I - G_N^\dagger G_N)^\top \Phi_N^\top E_N \,.$$

Indeed, on the one hand, it holds that

$$\begin{aligned}&(\theta^\star - \theta_N^\dagger)^\top G_N (\theta^\star - \theta_N^\dagger) \\ &= [(I - G_N^\dagger G_N)\theta^\star - G_N^\dagger \Phi_N^\top E_N]^\top G_N [(I - G_N^\dagger G_N)\theta^\star - G_N^\dagger \Phi_N^\top E_N] \\ &= \theta^{\star\top}(I - G_N^\dagger G_N)^\top G_N (I - G_N^\dagger G_N)\theta^\star \\ &\quad + E_N^\top \Phi_N G_N^\dagger G_N G_N^\dagger \Phi_N^\top E_N \\ &\quad - 2E_N^\top \Phi_N G_N^\dagger G_N (I - G_N^\dagger G_N)\theta^\star \\ &= ||E_N^\top \Phi_N||^2_{G_N^\dagger} \,,\end{aligned}$$

where we used the fact that $G_N G_N^\dagger G_N = G_N$ twice. On the other hand, we have

$$\begin{aligned}2(\theta^\star - \theta_N^\dagger)^\top \Phi_N^\top E_N &= 2[(I - G_N^\dagger G_N)\theta^\star - G_N^\dagger \Phi_N^\top E_N]^\top \Phi_N^\top E_N \\ &= 2\theta^{\star\top}(I - G_N^\dagger G_N)^\top \Phi_N^\top E_N - 2||E_N^\top \Phi_N||^2_{G_N^\dagger} \,.\end{aligned}$$

Note that $p = 2(I - G_N G_N^\dagger)\theta^\star$ satisfies $||p||_{G_N} = 0$ and thus we cannot directly apply the result of Theorem 5.4 to control $p^\top \Phi_N^\top E_N$ However, if $I \neq G_N G_N^\dagger$ it holds

$$\begin{aligned}\theta^{\star\top}(I - G_N^\dagger G_N)^\top \Phi_N^\top E_N &\leqslant ||E_N^\top \Phi_N||_{G_N^\dagger} ||\theta^\star||_{G_N} - \theta^{\star\top} G_N G_N^\dagger \Phi_N^\top E_N \\ &\leqslant ||E_N^\top \Phi_N||_{G_N^\dagger} ||\theta^\star||_{G_N} + ||G_N \theta^\star||_{G_N^\dagger} ||E_N^\top \Phi_N||_{G_N^\dagger} \\ &\leqslant 2||E_N^\top \Phi_N||_{G_N^\dagger} ||\theta^\star||_{G_N} \,.\end{aligned}$$

On the other hand, we can control the term $||E_N^\top \Phi_N||_{G_N^\dagger}$ via Theorem 5.4. Indeed,

it holds on an event Ω_1 of probability higher than $1-\delta$

$$\begin{aligned} 0 \leqslant ||E_N^\top \Phi_N||^2_{G_N^\dagger} &= E_N^\top \Phi_N G_N^\dagger \Phi_N^\top E_N \\ &\leqslant \sup_{p:||p||_{G_N}\neq 0} \frac{p^\top \Phi_N^\top E_N}{||p||_{G_N}} ||G_N^\dagger E_N^\top \Phi_N||_{G_N} \\ &= \sup_{p\in\mathscr{S}^d} \frac{p^\top \Phi_N^\top E_N}{||p||_{G_N}} ||E_N^\top \Phi_N||_{G_N^\dagger} \\ &\leqslant R^2 \underbrace{4\ln(\kappa_d(e\Lambda_N/\lambda_0)/\delta)}_{D_N(\delta)} . \end{aligned}$$

On this event, we thus get

$$\begin{aligned} \sum_{n=1}^N (y_n - \langle \theta_N, \varphi(x_n)\rangle)^2 &\leqslant ||E_N||^2 + 4R||\theta^\star||_{G_N}\sqrt{D_N(\delta)}\mathbb{I}\{I \neq G_N G_N^\dagger\} \\ &\geqslant ||E_N||^2 - 4R||\theta^\star||_{G_N}\sqrt{D_N(\delta)}\mathbb{I}\{I \neq G_N G_N^\dagger\} - R^2 D_N(\delta) , \end{aligned}$$

where $\mathbb{I}\{I \neq G_N G_N^\dagger\} = 0$ because of $\lambda_{\min}(G_N) \geqslant \lambda_0$. Finally, we use the result of Lemma 5.10 to get that with probability higher than $1-2\delta$,

$$\begin{aligned} ||E_N||^2 &\leqslant NR^2 + 2R^2\sqrt{2N\ln(e/\delta)(1+c_N/\ln(1/\delta))} + 2R^2\ln(e/\delta)(1+c_N/\ln(1/\delta)) \\ ||E_N||^2 &\geqslant NR^2 - 2R^2\sqrt{N\ln(e/\delta)(1+c_N/\ln(1/\delta))} . \end{aligned}$$

Thus, combining these two results with a union bound, we deduce that with probability higher than $1-3\delta$ it holds

$$\begin{aligned} \widehat{\sigma}_N^2 &\leqslant R^2 + 2R^2\sqrt{\frac{2\ln(e/\delta)(1+c_N/\ln(1/\delta))}{N}} + \frac{2R^2\ln(e/\delta)(1+c_N/\ln(1/\delta))}{N} \\ \widehat{\sigma}_N^2 &\geqslant R^2 - 2R^2\sqrt{\frac{\ln(e/\delta)(1+c_N/\ln(1/\delta))}{N}} - \frac{R^2 D_N(\delta)}{N} , \end{aligned}$$

We can now get a bound on $\sqrt{\widehat{\sigma}_N^2}$. Indeed

$$\begin{aligned} \widehat{\sigma}_N^2 &\leqslant (R + \sqrt{\frac{2R^2\ln(e/\delta)(1+c_N/\ln(1/\delta))}{N}})^2 \\ \widehat{\sigma}_N^2 &\geqslant (R - \sqrt{R^2\frac{\ln(e/\delta)(1+c_N/\ln(1/\delta))}{N}})^2 - R^2\frac{\ln(e/\delta)(1+c_N/\ln(1/\delta))}{N} \\ &\quad - \frac{R^2 D_N(\delta)}{N} . \end{aligned}$$

Thus, using the inequality $\sqrt{a+b} \leqslant \sqrt{a}+\sqrt{b}$, on both inequalities, we obtain

$$\begin{aligned}\sqrt{\widehat{\sigma}_N^2} &\leqslant R + R\sqrt{\frac{2\ln(e/\delta)(1+c_N/\ln(1/\delta))}{N}} \\ \sqrt{\widehat{\sigma}_N^2} &\geqslant R - R\sqrt{\frac{\ln(e/\delta)(1+c_N/\ln(1/\delta))}{N}} \\ &\quad -R\sqrt{\frac{\ln(e/\delta)(1+c_N/\ln(1/\delta))}{N} + \frac{D_N(\delta)}{N}}.\end{aligned}$$

□

Theorem 3.9 (Regularized variance estimate) *Under the conditionally strongly-sub-Gaussian assumption, for any random stopping time N for the filtration of the past, with probability higher than $1-3\delta$, it holds that*

$$\sqrt{\widehat{\sigma}_{N,\lambda}^2} \leqslant R\left[1+\sqrt{\frac{2C_N(\delta)}{N}}\right] + \frac{\lambda||\theta^\star||_2}{\sqrt{N\lambda_{\min(G_{N,\lambda})}}}\sqrt{1-\frac{\lambda}{\lambda_{\max}(G_{N,\lambda})}}$$

$$\sqrt{\widehat{\sigma}_{N,\lambda}^2} \geqslant R\left[1-\sqrt{\frac{C_N(\delta)}{N}} - \sqrt{\frac{C_N(\delta)+D_{N,\lambda}(\delta)\left(1+\frac{\lambda}{\lambda_{\min(G_{N,\lambda})}}\right)}{N}}\right] - \lambda\sqrt{\frac{2R||\theta^\star||_2\sqrt{D_{N,\lambda}(\delta)}}{N\lambda_{\min^{3/2}(G_{N,\lambda})}}}.$$

where $D_{N,\lambda}(\delta) = 2\ln\left(\frac{det(G_N+\lambda I)^{1/2}}{\delta\lambda^{d/2)}}\right)$ *and* $C_N(\delta) = \ln(e/\delta)\big[1+\ln(\pi^2\ln(N)/6)/\ln(1/\delta)\big]$.

Proof of Theorem 3.9:

We start with the following decomposition

$$\begin{aligned}&\sum_{n=1}^{N}(y_n - \langle\theta_{N,\lambda},\varphi(x_n)\rangle)^2 \\ &\quad= (\theta^\star-\theta_{N,\lambda})^\top G_N(\theta^\star-\theta_{N,\lambda}) + ||E_N||^2 + 2(\theta^\star-\theta_{N,\lambda})^\top\Phi_N^\top E_N\,. \qquad (10)\end{aligned}$$

On the one hand, we can control the first term in (10) via

$$\begin{aligned}&(\theta^\star-\theta_{N,\lambda})^\top G_N(\theta^\star-\theta_{N,\lambda}) \\ &= [(I-G_{N,\lambda}^{-1}G_N)\theta^\star - G_{N,\lambda}^{-1}\Phi_N^\top E_N]^\top G_N[(I-G_{N,\lambda}^{-1}G_N)\theta^\star - G_{N,\lambda}^{-1}\Phi_N^\top E_N] \\ &= [\lambda\theta^\star - \Phi_N^\top E_N]^\top G_{N,\lambda}^{-1}G_N G_{N,\lambda}^{-1}[\lambda\theta^\star - \Phi_N^\top E_N] \\ &= [\lambda\theta^\star - \Phi_N^\top E_N]^\top [G_{N,\lambda}^{-1} - \lambda G_{N,\lambda}^{-2}][\lambda\theta^\star - \Phi_N^\top E_N] \\ &= ||\Phi_N^\top E_N||^2_{G_{N,\lambda}^{-1}} - \lambda||\Phi_N^\top E_N||^2_{G_{N,\lambda}^{-2}} + \lambda^2||\theta^\star||^2_{G_{N,\lambda}^{-1}} - \lambda^3||\theta^\star||^2_{G_{N,\lambda}^{-2}} \\ &\quad -2\lambda\theta^{\star\top}[G_{N,\lambda}^{-1} - \lambda G_{N,\lambda}^{-2}]\Phi_N^\top E_N\end{aligned}$$

where we used the fact that $I - G_{N,\lambda}^{-1} G_N = \lambda G_{N,\lambda}^{-1}$ and then that $G_{N,\lambda}^{-1} G_N G_{N,\lambda}^{-1} = G_{N,\lambda}^{-1} - \lambda G_{N,\lambda}^{-2}$. Likewise, we control the third term in (10) via

$$\begin{aligned} 2(\theta^\star - \theta_{N,\lambda})^\top \Phi_N^\top E_N &= 2[(I - G_{N,\lambda}^{-1} G_N)\theta^\star - G_{N,\lambda}^{-1} \Phi_N^\top E_N]^\top \Phi_N^\top E_N \\ &= 2[\lambda\theta^\star - \Phi_N^\top E_N]^\top G_{N,\lambda}^{-1} \Phi_N^\top E_N \\ &= 2\lambda {\theta^\star}^\top G_{N,\lambda}^{-1} \Phi_N^\top E_N - 2||\Phi_N^\top E_N||^2_{G_{N,\lambda}^{-1}} . \end{aligned}$$

Combining these two bounds, it holds

$$\begin{aligned} &\sum_{n=1}^N (y_n - \langle \theta_{N,\lambda}, \varphi(x_n)\rangle)^2 \\ &= ||E_N||^2 - ||\Phi_N^\top E_N||^2_{G_{N,\lambda}^{-1}} - \lambda ||\Phi_N^\top E_N||^2_{G_{N,\lambda}^{-2}} \\ &\quad + \lambda^2 ||\theta^\star||^2_{G_{N,\lambda}^{-1}} - \lambda^3 ||\theta^\star||^2_{G_{N,\lambda}^{-2}} + 2\lambda^2 {\theta^\star}^\top G_{N,\lambda}^{-2} \Phi_N^\top E_N \\ &\leqslant ||E_N||^2 + \frac{\lambda^2}{\lambda_{\min(G_{N,\lambda})}} ||\theta^\star||_2^2 \Big(1 - \frac{\lambda}{\lambda_{\max}(G_{N,\lambda})}\Big) + 2\frac{\lambda^2}{\lambda_{\min^{3/2}(G_{N,\lambda})}} ||\theta^\star||_2 ||\Phi_N^\top E_N||_{G_{N,\lambda}^{-1}} \\ &\geqslant ||E_N||^2 + \frac{\lambda^2}{\lambda_{\max}(G_{N,\lambda})} ||\theta^\star||_2^2 \Big(1 - \frac{\lambda}{\lambda_{\min(G_{N,\lambda})}}\Big) - 2\frac{\lambda^2}{\lambda_{\min^{3/2}(G_{N,\lambda})}} ||\theta^\star||_2 ||\Phi_N^\top E_N||_{G_{N,\lambda}^{-1}} \\ &\quad - ||\Phi_N^\top E_N||^2_{G_{N,\lambda}^{-1}} \Big(1 + \frac{\lambda}{\lambda_{\min(G_{N,\lambda})}}\Big) . \end{aligned}$$

Now, from Theorem 5.4, it holds on an event Ω_1 of probability higher than $1-\delta$,

$$0 \leqslant ||\Phi_N^\top E_N||^2_{G_{N,\lambda}^{-1}} \leqslant R^2 \underbrace{2\ln\left(\frac{det(G_N + \lambda I)^{1/2}}{\delta \lambda^{d/2})}\right)}_{D_{N,\lambda}(\delta)} .$$

On the other hand, we control the second term $||E_N||^2$ by Lemma 5.10, and obtain that with probability higher than $1 - 2\delta$,

$$\begin{aligned} ||E_N||^2 &\leqslant NR^2 + 2R^2\sqrt{2NC_N(\delta)} + 2R^2 C_N(\delta) \\ ||E_N||^2 &\geqslant NR^2 - 2R^2\sqrt{NC_N(\delta)} , \end{aligned}$$

where $C_N(\delta) = \ln(e/\delta)(1 + c_N/\ln(1/\delta))$.

Thus, combining these two results with a union bound, we deduce that with proba-

bility higher than $1-3\delta$ it holds

$$\begin{aligned}
\widehat{\sigma}^2_{N,\lambda} &\leqslant R^2 + 2R^2\sqrt{\frac{2C_N(\delta)}{N}} + \frac{2R^2 C_N(\delta)}{N} \\
&\quad + \frac{\lambda^2}{N\lambda_{\min(G_{N,\lambda})}}||\theta^\star||_2^2\Big(1-\frac{\lambda}{\lambda_{\min}(G_{N,\lambda})}\Big) - 2\frac{R\lambda^2}{N\lambda_{\min^{3/2}(G_{N,\lambda})}}||\theta^\star||_2\sqrt{D_{N,\lambda}(\delta)} \\
\widehat{\sigma}^2_{N,\lambda} &\geqslant R^2 - 2R^2\sqrt{\frac{C_N(\delta)}{N}} + \frac{\lambda^2}{N\lambda_{\min}(G_{N,\lambda})}||\theta^\star||_2^2\Big(1-\frac{\lambda}{\lambda_{\min(G_{N,\lambda})}}\Big) \\
&\quad -2\frac{\lambda^2 R}{N\lambda_{\min^{3/2}(G_{N,\lambda})}}||\theta^\star||_2\sqrt{D_{N,\lambda}(\delta)} - \frac{R^2 D_{N,\lambda}(\delta)}{N}\Big(1+\frac{\lambda}{\lambda_{\min(G_{N,\lambda})}}\Big)\,.
\end{aligned}$$

We can now derive a bound on $\sqrt{\widehat{\sigma}^2_{N,\lambda}}$. Indeed,

$$\begin{aligned}
\widehat{\sigma}^2_{N,\lambda} &\leqslant \left(R+\sqrt{\frac{2R^2C_N(\delta)}{N}}\right)^2 + \frac{\lambda^2}{N\lambda_{\min(G_{N,\lambda})}}||\theta^\star||_2^2\Big(1-\frac{\lambda}{\lambda_{\min}(G_{N,\lambda})}\Big) \\
\widehat{\sigma}^2_{N,\lambda} &\geqslant \left(R-\sqrt{\frac{R^2C_N(\delta)}{N}}\right)^2 - \frac{R^2}{N}\left(C_N(\delta)+D_{N,\lambda}(\delta)\Big(1+\frac{\lambda}{\lambda_{\min(G_{N,\lambda})}}\Big)\right) \\
&\quad -\frac{2\lambda^2 R}{N\lambda_{\min^{3/2}(G_{N,\lambda})}}||\theta^\star||_2\sqrt{D_{N,\lambda}(\delta)}\,.
\end{aligned}$$

Thus, using the inequality $\sqrt{a+b}\leqslant\sqrt{a}+\sqrt{b}$, on both inequalities, we get

$$\begin{aligned}
\sqrt{\widehat{\sigma}^2_{N,\lambda}} &\leqslant R + R\sqrt{\frac{2C_N(\delta)}{N}} + \frac{\lambda||\theta^\star||_2}{\sqrt{N\lambda_{\min(G_{N,\lambda})}}}\sqrt{1-\frac{\lambda}{\lambda_{\min}(G_{N,\lambda})}} \\
\sqrt{\widehat{\sigma}^2_{N,\lambda}} &\geqslant R - R\sqrt{\frac{C_N(\delta)}{N}} - R\sqrt{\frac{C_N(\delta)+D_{N,\lambda}(\delta)\Big(1+\frac{\lambda}{\lambda_{\min(G_{N,\lambda})}}\Big)}{N}} \\
&\quad -\lambda\sqrt{\frac{2R||\theta^\star||_2\sqrt{D_{N,\lambda}(\delta)}}{N\lambda_{\min^{3/2}(G_{N,\lambda})}}}\,.
\end{aligned}$$

□

Corollary 5.11 *Under the conditionally strongly-sub-Gaussian assumption, for any random stopping time N for the filtration of the past, with probability higher than $1-3\delta$, it holds that either $\lambda_{\min}(G_N)<\lambda_0$ or*

$$R\leqslant\sqrt{\widehat{\sigma}^2_N}\left(1-\sqrt{\frac{C_N(\delta)}{N}}-\sqrt{\frac{C_N(\delta)}{N}+\frac{D_N(\delta)}{N}}\right)_+^{-1}.$$

where $(x)_+ = \max(x, 0)$*, and using the convention that* $(x)_+^{-1} = 0$ *if* $x \leqslant 0$*. Likewise, for any* $\lambda > 0$*, with probability higher than* $1 - 3\delta$*, it holds that*

$$R^{1/2} \leqslant \frac{\left[\sqrt{\widehat{\sigma}^2_{N,\lambda}}\alpha + \frac{\lambda^2||\theta^\star||_2\sqrt{D_{N,\lambda}(\delta)}}{2N\lambda_{\min^{3/2}(G_{N,\lambda})}}\right]^{1/2} + \left[\frac{\lambda^2||\theta^\star||_2\sqrt{D_{N,\lambda}(\delta)}}{2N\lambda_{\min^{3/2}(G_{N,\lambda})}}\right]^{1/2}}{\alpha},$$

where we introduced the notation

$$\alpha = \left(1 - \sqrt{\frac{C_N(\delta)}{N}} - \sqrt{\frac{C_N(\delta) + D_{N,\lambda}(\delta)\left(1 + \frac{\lambda}{\lambda_{\min(G_{N,\lambda})}}\right)}{N}}\right)_+ .$$

Proof of Corollary 5.11:

For the ordinary least-squares estimate, the result is immediate. We thus focus on the regularized estimate. Using the notations of Theorem 3.9,it holds

$$\underbrace{\sqrt{\widehat{\sigma}^2_{N,\lambda}}}_{A} \geqslant R\Big[\underbrace{1 - \sqrt{\frac{C_N(\delta)}{N}} - \sqrt{\frac{C_N(\delta) + D_{N,\lambda}(\delta)\left(1 + \frac{\lambda}{\lambda_{\min(G_{N,\lambda})}}\right)}{N}}}_{C}\Big] - \sqrt{R}\underbrace{\sqrt{\frac{2\lambda^2||\theta^\star||_2\sqrt{D_{N,\lambda}(\delta)}}{N\lambda_{\min^{3/2}(G_{N,\lambda})}}}}_{B}.$$

Provided that $C > 0$, the inequality rewrites in the form $A \geqslant RC - \sqrt{R}B$. Introducing the quantity $y^2 = R$, this inequality holds provided that $y \geqslant 0$ and $A + yB - Cy^2 \geqslant 0$, that is when $0 \leqslant y \leqslant \frac{B+\sqrt{B^2+4AC}}{2C}$.

□

6. Discussion

In this paper, we studied the construction of *streaming* confidence intervals in a modeled non-stationary setting with dependent data, when the number of observation may be a random stopping time. We analyzed, in a sub-Gaussian regression setting, both the ordinary and ridge estimates, and refined existing results. We focused also on deriving empirical Bernstein bounds in this challenging scenario, thus enriching previous work with novel results.

We paid special attention to providing further intuition regarding the loss function and the estimation of the future loss, which are important from a machine learning perspective. We also hope that the proof techniques are simple and detailed enough in order to be useful for pedagogical purpose.

Finally, although we refrained from applying our results to popular problems in machine learning and statistics such as linear multi-armed bandits, piece-wise linear regression, change point detection, or online aggregation of experts, we believe these domains can greatly benefit from this simple study.

Acknowledgments.

This research was supported by Inria Saclay and the Laboratoire de Recherche en Informatique (LRI), UMR 8623.

Appendix A: Regularized least-squares from a Bayesian standpoint

When the function space is of large dimension, d the matrix G_N is singular for $N < d$, and thus it becomes difficult to provide a confidence distributions on the parameter and next values. A natural way to fix this is to introduce a regularization, which forces to obtain a unique representative solution of the minimization problem. We can then control the error introduced by the regularization in a second step. Alternatively, one can also see regularization as an effective way to pick one solution in the space of all possible solutions, which becomes crucial when $d \to \infty$.

Using the previously introduced notations, the regularized estimate is given by

$$\theta_{N,\lambda} = (\mathbf{\Phi}_{[N]}^\top \mathbf{\Phi}_{[N]} + \lambda I)^{-1} \mathbf{\Phi}_{[N]}^\top \mathbf{Y}_{[N]} \,.$$

One way to better understand the regularization λ is to use the Bayesian point of view. To this end, the noise is not only assumed to be sub-Gaussian conditionally on the past, but is assumed to be exactly Gaussian, and independent on the past observations ($\xi(x_n) \sim \mathcal{N}(0, \sigma^2)$ for each n). We further assume that the observation points x_n are independent from the observations x_n, thus moving back to a classical Gaussian regression setting. If we pick a random function using $\theta \sim \mathcal{N}(0, \Sigma)$ (a prior which models how hard to get one function), and assume that the noise satisfies $\xi \sim \mathcal{N}(0, \sigma^2)$, ($\xi(x_n) \sim \mathcal{N}(0, \sigma^2)$ for each n) then the posterior mean function is in this case $\widehat{f}_N(x) | x, x_1, \dots, x_N, y_1, \dots, y_n \sim \widehat{\pi}_N(x) = \mathcal{N}(\mu_N(x), \sigma_N^2(x))$ where

$$\begin{aligned} \mu_N(x) &= \varphi(x)^\top (\mathbf{\Phi}_{[N]}^\top \mathbf{\Phi}_{[N]} + \sigma^2 \Sigma^{-1})^{-1} \mathbf{\Phi}_{[N]}^\top \mathbf{Y}_{[N]} \\ \sigma_N^2(x) &= \sigma^2 \varphi(x)^\top (\mathbf{\Phi}_{[N]}^\top \mathbf{\Phi}_{[N]} + \sigma^2 \Sigma^{-1})^{-1} \varphi(x) \,. \end{aligned}$$

This classical result (see (Rasmussen and Williams, 2006)) enables to directly interpret λ as being a variance term.

For convenience, we now introduce the notation $G_{N,\sigma} = (\mathbf{\Phi}_{[N]}^\top \mathbf{\Phi}_{[N]} + \sigma^2 \Sigma^{-1})$. Thus it holds $\mu_N(x) = \varphi(x)^\top G_{N,\sigma}^{-1} \mathbf{\Phi}_{[N]}^\top \mathbf{Y}_{[N]}$ and $\sigma_N^2(x) = \sigma^2 \varphi(x)^\top G_{N,\sigma}^{-1} \varphi(x)$. Alternatively, we can derive the "functional" form

$$\begin{aligned}
\mu_N(x) &= \varphi(x)^\top \Sigma \mathbf{\Phi}_{[N]}^\top (\mathbf{\Phi}_{[N]} \Sigma \mathbf{\Phi}_{[N]}^\top + \sigma^2 I_N)^{-1} \mathbf{Y}_{[N]} \\
\sigma_N^2(x) &= \varphi(x)^\top \Sigma \varphi(x) - \varphi(x)^\top \Sigma \mathbf{\Phi}_{[N]}^\top (\mathbf{\Phi}_{[N]} \Sigma \mathbf{\Phi}_{[N]}^\top + \sigma^2 I_N)^{-1} \mathbf{\Phi}_{[N]} \Sigma \varphi(x) \,.
\end{aligned}$$

For convenience, we now introduce the notation $K_{N,\sigma} = (\mathbf{\Phi}_{[N]} \Sigma \mathbf{\Phi}_{[N]}^\top + \sigma^2 I_N)$ and $\kappa_{N,\Sigma}(x) = \mathbf{\Phi}_{[N]} \Sigma \varphi(x)$. Thus it holds $\mu_N(x) = \kappa_{N,\sigma}(x)^\top K_{N,\sigma}^{-1} \mathbf{Y}_{[N]}$, and on the other hand $\sigma_N^2(x) = \varphi(x)^\top \Sigma \varphi(x) - \kappa_{N,\Sigma}(x)^\top K_{N,\sigma}^{-1} \kappa_{N,\Sigma}(x)$. This functional form is convenient as it generalizes to infinite dimensions Indeed let us introduce $k(x,x)$ to generalize $\varphi(x)^\top \Sigma \varphi(x)$, as well as $k_N(x) = (k(x_n, x))_{n \in [N]}$ and $\mathbf{K}_N = (k(x_i, x_j))_{i,j \in [N]}$. This leads to

$$\begin{aligned}
\mu_N(x) &= k_N(x)(K_N + \sigma^2 I_N)^{-1} \mathbf{Y}_{[N]} \\
\sigma_N^2(x) &= k(x,x) - k_N(x)^\top (\mathbf{K}_N + \sigma^2 I_N)^{-1} k_N(x) \,.
\end{aligned}$$

The following lemma holds in the setting when each x_n is independent on the past observations $\{y_m\}_{m<n}$, and comes from standard derivations (see (Rasmussen and Williams, 2006)).

Lemma A.1 (Mean-variance in the Bayesian model) *Under the Bayesian model, then $y(x) - \mu_N(x)$ is Gaussian, with mean and variance given in the parametric form by*

$$\begin{aligned}
\mathbb{E}[y(x) - \mu_N(x) | x, \{x_n\}_{n \in [N]}, \theta^\star] &= \sigma^2 \varphi(x)^\top G_{N,\sigma}^{-1} \Sigma^{-1} \theta^\star \\
\mathbb{V}[y(x) - \mu_N(x) | x, \{x_n\}_{n \in [N]}, \theta^\star] &= \sigma^2 \left(\varphi(x)^\top G_{N,\sigma}^{-1} (I - \sigma^2 \Sigma^{-1} G_{N,\sigma}^{-1}) \varphi(x) + 1 \right),
\end{aligned}$$

and in the functional form by

$$\begin{aligned}
\mathbb{E}[y(x) - \mu_N(x) | x, \{x_n\}_{n \in [N]}, \theta^\star] &= f^\star(x) - k_N(x)^\top (K_N + \sigma^2 I_N)^{-1} \mathbf{f}_N^\star \\
\mathbb{V}[y(x) - \mu_N(x) | x, \{x_n\}_{n \in [N]}, \theta^\star] &= \sigma^2 \left(k_N(x)^\top (K_N + \sigma^2 I_N)^{-2} k_N(x) + 1 \right).
\end{aligned}$$

Further, we can replace $y(x)$ with $f^\star(x)$ in the left-hand side with no effect on the expectations, and by subtracting σ^2 to the variance. Similarly we can replace in the left-hand side $\mu_N(x)$ with $f_\theta(x)$ where $\theta \sim \mathcal{N}(0, \Sigma)$, with no effect on the expectations, and by adding $\sigma_N^2(x)$ to the variance term.

As for conditional means and variances, it holds

$$\begin{aligned}
\mathbb{E}[y(x) - \mu_N(x) | x, \{x_n, y_n\}_{n \in [N]}, \theta^\star] &= f^\star(x) - \mu_N(x) \\
\mathbb{V}[y(x) - \mu_N(x) | x, \{x_n, y_n\}_{n \in [N]}, \theta^\star] &= \sigma^2 \\
\mathbb{E}[y(x) - f_\theta(x) | x, \{x_n, y_n\}_{n \in [N]}, \theta^\star] &= f^\star(x) - \mu_N(x) \\
\mathbb{V}[y(x) - f_\theta(x) | x, \{x_n, y_n\}_{n \in [N]}, \theta^\star] &= \sigma_N^2(x) + \sigma^2 \,.
\end{aligned}$$

The bias in the parametric form depends on $\theta^\star$ that is unknown. However, if a bound C on $||\Sigma^{-1}\theta^\star||_2$ is known, we can derive the bound

$$0 \leqslant |\sigma^2 \varphi(x)^\top G_{N,\sigma}^{-1} \Sigma^{-1} \theta^\star| \leqslant \frac{\sigma^2}{\lambda_{\min}^{1/2}(G_{N,\sigma})} ||\varphi(x)||_{G_{N,\sigma}^{-1}} ||\Sigma^{-1}\theta^\star||_2 \leqslant \frac{\sigma^2}{\lambda_{\min}^{1/2}(G_{N,\sigma})} ||\varphi(x)||_{G_{N,\sigma}^{-1}} C \, .$$

In particular, using the fact that $y(x) - \mu_N(x)$ is Gaussian, we deduce that

Corollary A.2 *For all $\delta \in (0,1)$,*

$$\mathbb{P}\left(y(x) - \mu_N(x) \geqslant \frac{\sigma^2}{\lambda_{\min}^{1/2}(G_{N,\sigma})} ||\varphi(x)||_{G_{N,\sigma}^{-1}} C + \sqrt{2\mathbb{V}_N(x)\log(1/\delta)} \middle| x, \{x_n\}_{n\in[N]}, \theta^\star \right) \leqslant \delta \, .$$

$$\mathbb{P}\left(\mu_N(x) - y(x) \geqslant \sqrt{2\mathbb{V}_N(x)\log(1/\delta)} \middle| x, \{x_n\}_{n\in[N]}, \theta^\star \right) \leqslant \delta \, .$$

where we introduced for convenience $\mathbb{V}_N(x) = \mathbb{V}[y(x) - \mu_N(x) | x, \{x_n\}_{n\in[N]}, \theta^\star]$.

Proof of Lemma A.1:

Let us start with the parametric form. In this case, we control the bias by

$$\begin{aligned}
&\mu_N(x) - f^\star(x) \\
&= \varphi(x)^\top \left((\mathbf{\Phi}_{[N]}^\top \mathbf{\Phi}_{[N]} + \sigma^2 \Sigma^{-1})^{-1} \mathbf{\Phi}_{[N]}^\top (\mathbf{\Phi}_{[N]} \theta^\star + E_{[N]}) - \theta^\star \right) \\
&= \varphi(x)^\top \left((\mathbf{\Phi}_{[N]}^\top \mathbf{\Phi}_{[N]} + \sigma^2 \Sigma^{-1})^{-1} \left[(\mathbf{\Phi}_{[N]}^\top \mathbf{\Phi}_{[N]} + \sigma^2 \Sigma^{-1}) \theta^\star - \sigma^2 \Sigma^{-1} \theta^\star + \mathbf{\Phi}_{[N]}^\top E_{[N]} \right] - \theta^\star \right) \\
&= \varphi(x)^\top (\mathbf{\Phi}_{[N]}^\top \mathbf{\Phi}_{[N]} + \sigma^2 \Sigma^{-1})^{-1} \left(\mathbf{\Phi}_{[N]}^\top E_{[N]} - \sigma^2 \Sigma^{-1} \theta^\star \right) .
\end{aligned}$$

Thus, we deduce that

$$\begin{aligned}
\mathbb{E}[f^\star(x) - \mu_N(x) | x, \{x_n\}_{n\in[N]}, \theta^\star] &= \sigma^2 \varphi(x)^\top G_{N,\sigma}^{-1} \Sigma^{-1} \theta^\star \\
\mathbb{V}[f^\star(x) - \mu_N(x) | x, \{x_n\}_{n\in[N]}, \theta^\star] &= \sigma^2 \varphi(x)^\top G_{N,\sigma}^{-1} \mathbf{\Phi}_{[N]}^\top \mathbf{\Phi}_{[N]} G_{N,\sigma}^{-1} \varphi(x) \\
&= \sigma^2 \varphi(x)^\top G_{N,\sigma}^{-1} \left(I - \sigma^2 \Sigma^{-1} G_{N,\sigma}^{-1} \right) \varphi(x) \, .
\end{aligned}$$

Note that here $f^\star$ is assumed to be fixed, not randomly sampled from a Gaussian process, which explains why the variance is smaller than $\sigma_N^2(x)$. More precisely the missing variance is $\sigma^4 \varphi(x)^\top G_{N,\sigma}^{-1} \Sigma^{-1} G_{N,\sigma}^{-1} \varphi(x)$. It further holds that

$$\begin{aligned}
\mathbb{E}[f^\star(x) - f_\theta(x) | x, \{x_n\}_{n\in[N]}, \theta^\star] &= \mathbb{E}[f^\star(x) - \mu_N(x) | x, \{x_n\}_{n\in[N]}, \theta^\star] \\
\mathbb{V}[f^\star(x) - f_\theta(x) | x, \{x_n\}_{n\in[N]}, \theta^\star] &= \mathbb{V}[f^\star(x) - \mu_N(x) | x, \{x_n\}_{n\in[N]}, \theta^\star] + \sigma_N^2(x) \, .
\end{aligned}$$

Now, let us turn to the function form. We obtain by the Shermann-Morrison formula

$$\begin{aligned}
&\mathbb{E}[f^\star(x) - \mu_N(x) | x, \{x_n\}_{n\in[N]}, \theta^\star] \\
&\quad = \sigma^2 \varphi(x)^\top \Big(\sigma^{-2}\Sigma - \sigma^{-2}\Sigma \mathbf{\Phi}_{[N]}^\top \big(I_N + \sigma^{-2} \mathbf{\Phi}_{[N]} \Sigma \mathbf{\Phi}_{[N]}^\top\big)^{-1} \mathbf{\Phi}_{[N]} \sigma^{-2} \Sigma \Big) \Sigma^{-1} \theta^\star \\
&\quad = \varphi(x)^\top \theta^\star - \varphi(x)^\top \Sigma \mathbf{\Phi}_{[N]}^\top \Big(\sigma^2 I_N + \mathbf{\Phi}_{[N]} \Sigma \mathbf{\Phi}_{[N]}^\top \Big)^{-1} \mathbf{\Phi}_{[N]} \theta^\star \\
&\quad = f^\star(x) - \kappa_{N,\Sigma}(x)^\top K_{N,\sigma}^{-1} \mathbf{f}_N^\star \\
&\mathbb{V}[f^\star(x) - \mu_N(x) | x, \{x_n\}_{n\in[N]}, \theta^\star] \\
&\quad = \varphi(x)^\top \Sigma \Big(I - \mathbf{\Phi}_{[N]}^\top K_{N,\sigma}^{-1} \mathbf{\Phi}_{[N]} \Sigma \Big) \mathbf{\Phi}_{[N]}^\top K_{N,\sigma}^{-1} \mathbf{\Phi}_{[N]} \Sigma \varphi(x) \\
&\quad = -\varphi(x)^\top \Sigma \mathbf{\Phi}_{[N]}^\top \big(\sigma^2 I_N + \mathbf{\Phi}_{[N]} \Sigma \mathbf{\Phi}_{[N]}^\top \big)^{-1} \mathbf{\Phi}_{[N]} \Sigma \mathbf{\Phi}_{[N]}^\top K_{N,\sigma}^{-1} \mathbf{\Phi}_{[N]} \Sigma \varphi(x) \\
&\qquad + \varphi(x)^\top \Sigma \mathbf{\Phi}_{[N]}^\top K_{N,\sigma}^{-1} \mathbf{\Phi}_{[N]} \Sigma \varphi(x) \\
&\quad = -\kappa_{N,\sigma}(x)^\top \Big(I - \sigma^2 K_{N,\sigma}^{-1} \Big) K_{N,\sigma}^{-1} \kappa_{N,\sigma}(x) \\
&\qquad + \kappa_{N,\sigma}(x)^\top K_{N,\sigma}^{-1} \kappa_{N,\sigma}(x) \\
&\quad = \sigma^2 \kappa_{N,\Sigma}(x)^\top K_{N,\sigma}^{-2} \kappa_{N,\Sigma}(x) \,.
\end{aligned}$$

In function form, it then holds

$$\begin{aligned}
\mathbb{E}[f^\star(x) - \mu_N(x) | x, \{x_n\}_{n\in[N]}, \theta^\star] &= f^\star(x) - k_N(x)^\top (K_N + \sigma^2 I_N)^{-1} \mathbf{f}_N^\star \\
\mathbb{V}[f^\star(x) - \mu_N(x) | x, \{x_n\}_{n\in[N]}, \theta^\star] &= \sigma^2 k_N(x)^\top (K_N + \sigma^2 I_N)^{-2} k_N(x) \\
\mathbb{E}[y(x) - f_\theta(x) | x, \{x_n\}_{n\in[N]}, \theta^\star] &= f^\star(x) - k_N(x)^\top (K_N + \sigma^2 I_N)^{-1} \mathbf{f}_N^\star \\
\mathbb{V}[y(x) - f_\theta(x) | x, \{x_n\}_{n\in[N]}, \theta^\star] &= \sigma^2 k_N(x)^\top (K_N + \sigma^2 I_N)^{-2} k_N(x) + \sigma_N^2(x) + \sigma^2 \,.
\end{aligned}$$

□

References

Abbasi-Yadkori, Y., Pal, D. and Szepesvari, C. (2011). Improved Algorithms for Linear Stochastic Bandits. *Nips* 1–19.

Audiffren, J. and Ralaivola, L. (2014). Stationary Mixing Bandits. *arXiv preprint arXiv:1406.6020*.

Campbell, S. L. and Meyer, C. D. (2009). *Generalized inverses of linear transformations* **56**. SIAM.

Cappé, O., Garivier, A. and Maillard, O. (2013). Kullback-Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics* **41** 1–56.

Chu, W., Li, L., Reyzin, L. and Schapire, R. E. (2011). Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics* 208–214.

Csiszar, I. (1998). The method of types [information theory]. *IEEE Transactions on Information Theory* **44** 2505–2523.

Grünewälder, S., Audibert, J.-Y., Opper, M. and Shawe-Taylor, J. (2010). Regret bounds for gaussian process bandit problems. In *AISTATS 2010-Thirteenth International Conference on Artificial Intelligence and Statistics* **9** 273–280.

Krause, A. and Ong, C. S. (2011). Contextual gaussian process bandit optimization. In *Advances in Neural Information Processing Systems* 2447–2455.

Kuznetsov, V. and Mohri, M. (2015). Learning Theory and Algorithms for Forecasting Non-Stationary Time Series. In *Advances in Neural Information Processing Systems* 541–549.

Leadbetter, M. R., Lindgren, G. and Rootzén, H. (2012). *Extremes and related properties of random sequences and processes*. Springer Science & Business Media.

Maurer, A. and Pontil, M. (2009). Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*.

Merhav, N. and Feder, M. (1998). Universal prediction. *IEEE Transactions on Information Theory* **44** 2124–2147.

Peña, V. H., Lai, T. L. and Shao, Q.-M. (2008). *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media.

Rasmussen, C. and Williams, C. (2006). Gaussian Processes for Machine Learning. *Gaussian Processes for Machine Learning*.

Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research* **35** 395–411.

Valko, M., Korda, N., Munos, R., Flaounas, I. and Cristianini, N. (2013). Finite-Time Analysis of Kernelised Contextual Bandits. In *The 29th Conference on Uncertainty in Artificial Intelligence*.