Structured Prediction in Online Learning

Pierre Boudart, Alessandro Rudi, Pierre Gaillard

STRUCTURED PREDICTION IN ONLINE LEARNING

A PREPRINT

Pierre Boudart
INRIA, École Normale Supérieure
CNRS, PSL Research University
Paris, France
pierre.boudart@inria.fr

Alessandro Rudi
INRIA, École Normale Supérieure
CNRS, PSL Research University
Paris, France
alessandro.rudi@inria.fr

Pierre Gaillard
Univ. Grenoble Alpes, Inria,
CNRS, Grenoble INP, LJK
Grenoble, France
pierre.gaillard@inria.fr

June 17, 2024

ABSTRACT

We study a theoretical and algorithmic framework for structured prediction in the online learning setting. The problem of structured prediction, i.e. estimating function where the output space lacks a vectorial structure, is well studied in the literature of supervised statistical learning. We show that our algorithm is a generalisation of optimal algorithms from the supervised learning setting, and achieves the same excess risk upper bound also when data are not i.i.d. Moreover, we consider a second algorithm designed especially for non-stationary data distributions, including adversarial data. We bound its stochastic regret in function of the variation of the data distributions.

1 Introduction

Online learning is a subfield of statistical learning in which a learner receives a flow of data generated by an environment (Cesa-Bianchi and Lugosi, 2006; Orabona, 2023; Hazan, 2023). The learner has to learn from the flow of data, and adapt to the data which could be non-stationary or adversarial. More formally, at each time step t, the learner receives a context $x_t \in \mathcal{X}$ from which he makes a prediction $\hat{z}_t = f_t(x_t) \in \mathcal{Z}$. His prediction is then compared to the true label $y_t \in \mathcal{Y}$, which is observed. The learner than pays an error $\Delta(\hat{z}_t, y_t)$ measured by a known loss function $\Delta : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$. The goal of the learner is to minimise his regret

$$R_T = \sum_{t=1}^{T} \Delta(\hat{z}_t, y_t) - \Delta(f_t^*(x_t), y_t)\,, \tag{1}$$

where $f_t^*(x_t) \in \arg\min_{z \in \mathcal{Z}} \Delta(z, y_t)$. The inputs x_t and labels y_t are generated sequentially by the environment and could be adversarial. This could model the change of behaviour of a customer or an evolution of the environment such as climate change. Note that in our framework, unlike the standard regret definition in online learning, the learner's performance is compared to the best function f_t at each round, similar to the approach used in dynamic regret (Herbster and Warmuth, 1998).

When the output space contains a vectorial structure, statistical learning provides many algorithms with statistical guarantees. However more and more applications involve an output space which lacks a linear structure, such as translation (Lacoste-Julien et al., 2006), image segmentation (Forsyth and Ponce, 2002), protein folding (Joachims et al., 2009), ranking (Duchi et al., 2010). These problems are often referred as structured prediction problems, because the output space may be represented for instance as a sequence, a graph, or an ordered set. In practice, an ad hoc method is designed to solved each of these problems and is most of the time based on surrogate methods and empirical risk

minimisation. If they achieve good results in practice, they however lack generalisation and are not built in order to have good theoretical guarantees.

We consider the structured prediction framework of *Implicit Loss Embedding (ILE)* (Ciliberto et al., 2020), in which the loss is of the form $\Delta(z, y) = \langle \psi(z), \varphi(y) \rangle$ for some unknown and infinite dimensional feature maps $\psi : \mathcal{Z} \to \mathcal{H}$ and $\varphi : \mathcal{Y} \to \mathcal{H}$ into an unknown RKHS $\mathcal{H}$ (see Definition 1). Such an assumption is satisfied by most losses for rich enough feature maps and used in the practical applications detailed above. (Ciliberto et al., 2020) study this framework in a statistical supervised learning setting and provide a general algorithm for general problems including discrete outputs and manifold regression. Their algorithm comes with statistical guarantees on the excess risk when data are i.i.d. only.

In the context of prediction of arbitrary sequences, the closer works to ours are (McMahan and Orabona, 2014) and (Pacchiano et al., 2018). On the one hand, (McMahan and Orabona, 2014) analyses a loss written as an inner product in a Hilbert space $\Delta(z, y) = \langle z, \varphi(y) \rangle$. However, they assume that the action space $\mathcal{Z}$ is itself a Hilbert space $\mathcal{H}$ which thus has a vectorial structure, contrary to the setting we consider. On the other hand, Pacchiano et al. (2018) also considers a loss expressed by a kernel with full information and partial feedback, but they do not consider contextual information x_t and require prior knowledge of the kernel feature maps ψ and φ, which we do not need.

Contributions Our work is the first to study structured prediction in the framework of prediction of arbitrary sequences.

We first introduce a new algorithm, called *OSKAAR* (Algorithm 1) and inspired by the work of (Ciliberto et al., 2020) in the statitical framework. Given a RKHS $\mathcal{G}$ from $\mathcal{X}$ to $\mathcal{H}$ associated to a kernel of the feature space $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a regularization parameter λ, *OSKAAR* achieves a regret upper-bound (Theorem 2) of order[1]:

$$R_T \lesssim \sqrt{T\big(d_{\text{eff}}(\lambda) + \min_{g \in \mathcal{G}} L_T(g)\big)}\,, \qquad \text{where} \quad L_T(g) := \sum_{t=1}^{T} ||g(x_t) - \varphi(y_t)||^2_{\mathcal{H}} + \lambda ||g||^2_{\mathcal{G}}\,,$$

where $d_{\text{eff}}(\lambda)$ is the effective dimension (13) that measures the size of the RKHS and $L_T(g)$ measures how the RKHS g is able to interpolate features of the data. In particular, if there is a function $g^* \in \mathcal{G}$ that perfectly models the features $(\varphi(y_t))_t$, i.e. $g^*(x_t) = \varphi(y_t)$ for all t, noting that $d_{\text{eff}}(\lambda) \lesssim T/\lambda$, the above result yields a regret bound of the order of $O(T^{3/4})$. However, such an assumption is strong even for i.i.d. data, and the above bound might be linear in T in the worst-case. in the worst-case scenario. This is not surprising, as the learner's performance is compared to the best possible baseline $\arg\min_{z \in \mathcal{Z}} \Delta(z, y_t)$ at each time step, which is generally unattainable.

To weaken the above assumption, we also prove the following expected regret bound for *OSKAAR* (Theorem 3):

$$\mathbb{E}[R_T] \lesssim \sqrt{T\big(d_{\text{eff}}(\lambda) + \min_{g \in \mathcal{G}} \bar{L}_T(g)\big)}, \quad \text{where} \quad \bar{L}_T(g) = \mathbb{E}\bigg[\sum_{t=1}^{T} \|g(x_t) - \mathbb{E}[\varphi(y_t)|x_t]\|^2_{\mathcal{H}}\bigg] + \lambda \|g\|^2_{\mathcal{G}}\,,$$

where the expectation is taken with respect to the possible randomness of the data (x_t, y_t). In the context of arbitrary sequences, the two above results exactly match. Yet, the assumption that there exists some g^* such that $g^*(x_t) = \mathbb{E}[\varphi(y_t)|x_t]$ for all t, is much weaker in general than assuming $g^*(x_t) = \varphi(y_t)$ since random variation of $\varphi(y_t)$ are not considered. Such an assumption is weak in the i.i.d. statistical framework and standard in the analysis of Kernel Ridge Regression (Caponnetto and De Vito, 2007; Steinwart and Christmann, 2008). It corresponds to assuming that the data distribution lies in the RKHS. In particular, we show that our analysis allows to recover (up to a log factor) the optimal rate of Ciliberto et al. (2020) in the i.i.d. setting, by designing an estimator $\bar{f}_T$ that satisfies the excess risk upper-bound:

$$\mathbb{E}_{x,y}[\Delta(\bar{f}_T(x), y) - \Delta(f^*(x), y)] \lesssim T^{-1/4} + T^{-1/2}\sqrt{\log(\delta^{-1})} \qquad \text{w.p. } 1 - \delta.$$

Our estimator $\bar{f}_T$ is constructed via a careful online to batch conversion to face with two challenges: the loss $\Delta(z, y)$ being non-convex in z standard online to batch conversion techniques that use $\bar{f}_T = \sum_{t=1}^{T} f_t$ are not possible here; our result holds with high-probability which is challenging to obtain with such techniques (van der Hoeven et al., 2023).

The above result still hold under the assumption that $g^*(x_t) = \mathbb{E}[\varphi(y_t)|x_t]$ for all t for some g^*, which is weak for stationary data but strong in our framework of arbitrary sequences. Our third contribution aims at relaxing this assumption. We design a second algorithm, referred to as *SALAMI* (Algorithm 2), that achieves under the assumption that there exists $g_t^* \in \mathcal{G}$ such that $g_t^*(x_t) = \mathbb{E}[\varphi(y_t)|x_t]$ for all t:

$$\mathbb{E}[R_T] = \begin{cases} \tilde{O}(V_{\mathcal{G}}^{1/6} T^{5/6}) & \text{if } \lambda = V_{\mathcal{G}}^{-1/3} T^{1/3} \\ \tilde{O}(V_0^{1/4} T^{3/4}) & \text{if } \lambda = V_0^{-1/2} T^{1/2} \end{cases}, \tag{2}$$

[1]The symbol $\lesssim$ is a rough inequality that neglects contants and logarithmic factors.

where V_0 and $V_{\mathcal{G}}$ are two different measures of the non-stationarity of the sequence (g_t^*):

$$V_0 = 1 + \sum_{t=2}^{T} 1\{g_t^* \neq g_{t-1}^*\} \qquad \text{and} \qquad V_{\mathcal{G}} := \|g_1^*\|_{\mathcal{G}} + \sum_{t=2}^{T} \|g_t^* - g_{t-1}^*\|_{\mathcal{G}} \,.$$

Paper outline In the next section, we recall the setting of the problem and the background on the *ILE* definition. In Section 3.1, we introduce our first algorithm *Online Structured prediction with Kernel Aggregating Algorithm Regression (OSKAAR)* and the algorithm from the batch setting. In section 3.2, we bound the regret of our algorithm. In Section 4, we recover the convergence rate from the batch setting without stochastic assumption. And in Section 5, we introduce our second algorithm *Structured prediction ALgorithm with Aggregating MIxture (SALAMI)* for non-stationary data and bound its stochastic regret. The details of the proofs can be found in appendix. Moreover, in Appendix C and E we provide bounds in high probability for both the stationary and the non-stationary settings.

2 Problem Setting and Background

We recall the setting and introduce the main notations used throughout the paper. We then discuss the limitations of the previous works. We denote by $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ respectively the input, label and output spaces of the learning problem. We denote by $\Delta : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ the loss function, which measures the error between a prediction in $\mathcal{Z}$ and a true label in $\mathcal{Y}$. Having two different spaces $\mathcal{Y}$ and $\mathcal{Z}$ allows to consider applications where the outputs do not match the labels such as ranking (Duchi et al., 2010).

Online Learning Framework Our online framework is formalised as a game between a learner and an environment, see Framework 1. At each time step $t \geq 1$, the user receives a context $x_t \in \mathcal{X}$, computes a prediction $\hat{z}_t = f_t(x_t) \in \mathcal{Z}$ based on the current context x_t and the history $(x_1, y_1, \ldots, x_{t-1}, y_{t-1})$. The true label $y_t \in \mathcal{Y}$ is then revealed to the learner, which incurs a loss $\Delta(\hat{z}_t, y_t)$. In this framework we are in the full information setting. That is to say that observing the label y_t enables the learner to compute the loss $\Delta(z, y_t)$ for all $z \in \mathcal{Z}$.

Framework 1: Online learning framework with contextual information

for *Each time step* t *in* $1 \ldots T$ **do**
- Get information $x_t \in \mathcal{X}$
- Compute the prediction $\hat{z}_t = f_t(x_t) \in \mathcal{Z}$
- Observe the label $y_t \in \mathcal{Y}$
- Get loss $\Delta(\hat{z}_t, y_t) \in \mathbb{R}$
- Update predictor f_{t+1}

end

The online learning setting allows us to also work with adversarial or non-stationary data, i.e. data that are not i.i.d. This could model a change of the environment. Throughout the paper we consider a loss $\Delta : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ that admits an *Implicit Loss Embedding (ILE)*, see Definition 1, with feature maps ψ, φ, and a Hilbert space $\mathcal{H}$.

Definition 1 (ILE (Ciliberto et al., 2020))**.** *A continuous map $\Delta : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ is said to admit an Implicit Loss Embedding (ILE) if there exists a separable Hilbert space $\mathcal{H}$ and two measurable bounded maps $\psi : \mathcal{Z} \to \mathcal{H}$ and $\varphi : \mathcal{Y} \to \mathcal{H}$, such that for any $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$ we have*

$$\Delta(z, y) = \langle \psi(z), \varphi(y) \rangle_{\mathcal{H}} \tag{3}$$

and $||\varphi(y)||_{\mathcal{H}} \leq 1$. Additionally, we define $c_\Delta = \sup_{z \in \mathcal{Z}} ||\psi(z)||_{\mathcal{H}}$.

In particular we do not assume that the loss is convex or differentiable. The metric to evaluate the performance of a learning algorithm is the regret defined as

$$R_T = \sum_{t=1}^{T} \Delta(\hat{z}_t, y_t) - \Delta(f_t^*(x_t), y_t) \tag{4}$$

where $f_t^*(x_t) \in \arg\min_{z \in \mathcal{Z}} \Delta(z, y_t)$ is used as the baseline. Taking the optimum inside the sum as we do is stronger than taking the optimum of the sum as is usually done.

Structured Prediction This is the most general setting in supervised learning. We say that a learning problem is structured if we have one of the following conditions (Vila, 2022):

- The loss is different than the 0-1 loss : $\Delta(z, y) \neq 1[z \neq y]$.
- The size of the output space is exponentially larger than the natural dimension of the output elements.

The first condition implies that some pairs of outputs and labels are closer than others. For instance, two sets that differ by only one element should be closer to each other compared to sets with an empty intersection. The second condition

characterizes a space of sequences, where the cardinality is exponential in the size of the dictionary used to build the sequences. The following spaces and losses are structured:

- Subsets of $[\![k]\!] := \{1, \dots, k\}$ with the negative F1 score $\Delta(z, y) = -2|z \cap y|/(|z| + |y|)$
- Ordered elements: $\mathcal{Z} = \mathcal{Y} = ([\![k]\!], <)$ with $\Delta(z, y) = |z - y|$
- Sequences of k elements of a dictionary $\mathcal{D}$ with the Hamming distance $\Delta(z, y) = \|z - y\|_0$
- Ranking, Information Retrieval: the goal is to predict an ordered list of documents or web pages from $x \in \mathcal{X}$ a query in a search engine. The output space $\mathcal{Z}$ is the space of permutations and the label space $\mathcal{Y}$ contains scalar scores representing the relevance of each document for the query (Duchi et al., 2010).

Note that we do not assume to have a vectorial structure in the output or the label space.

Standard Approach The classical learning approach, in the supervised learning setting, is Empirical Risk Minimisation (ERM) (Devroye et al., 2013). The expected risk is estimated by the empirical risk, and f_n computed as its minimiser. The underlying idea is that f_n should approach f^* as size of the sample n grows. The estimator f_n is defined as follows

$$f_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \Delta(f(x_i), y_i) \tag{5}$$

where $\mathcal{F}$ is a class of function and an hyper-parameter of the method. When the loss Δ is convex and the output space $\mathcal{Z}$ has a vectorial structure ERM becomes an efficient strategy for a large family of spaces $\mathcal{F}$. However this strategy presents some limitations (Ciliberto et al., 2020):

- **Modeling.** If we do not assume to have a vectorial structure on the output space $\mathcal{Z}$, it is not clear how to design a suitable function space $\mathcal{F}$. For instance, given $f_1, f_2 : \mathcal{X} \to \mathcal{Z}$, there is no guarantee that $f_1 + f_2$ takes values in $\mathcal{Z}$ as well.
- **Computations.** If the function space $\mathcal{F}$ is non-linear or the loss in non-convex, solving ERM can be challenging. Most approaches, such as gradient descent, are based on the regularity of the loss or the optimisation domain.

Existing results in the batch statistical framework We briefly recall the main results from Ciliberto et al. (2020). The authors introduced the *ILE* assumption (see Def. 1) and studied learning problems that satisfy this definition in the supervised learning setting. The mathematical constructs introduced in this definition, such as the feature maps ψ, φ and the Hilbert space $\mathcal{H}$, are used solely for analysis purposes and algorithm design. Notably, they are not required for making predictions. An important feature of their work, which we also achieve, is that our online algorithms do not need prior knowledge of ψ, φ and $\mathcal{H}$.

Let $(x_i, y_i)_{i=1}^n$ be a sample of i.i.d. data. Ciliberto et al. (2020) consider the ERM estimator $g_n : \mathcal{X} \to \mathcal{H}$ that learns the features $\varphi(y)$ as

$$g_n := \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \|\varphi(y_i) - g(x_i)\|_{\mathcal{H}}^2 + \lambda \|g\|_{\mathcal{G}}^2 \tag{6}$$

over a known kernel space $\mathcal{G}$. Choosing a kernel space gives us a closed form solution and strong algebraic properties to analyse the algorithm. Moving the problem to the feature space $\mathcal{H}$, enables us to enjoy the vectorial structure of $\mathcal{H}$. The authors then define the predictor $f_n : \mathcal{X} \to \mathcal{Z}$ as an optimisation problem using g_n as follows

$$f_n(x) := \arg\min_{z \in \mathcal{Z}} \langle \psi(z), g_n(x) \rangle_{\mathcal{H}} = \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^{n} \alpha_i(x) \Delta(z, y_i)\,, \tag{7}$$

where α_i are coefficients obtained by resorting to the representer theorem. Let $\mathcal{E}$ and $\mathcal{R}$ be the expected risk of $f_n : \mathcal{X} \to \mathcal{Z}$ and $g_n : \mathcal{X} \to \mathcal{H}$ respectively and f^* and g^* be their respective minimizers. Ciliberto et al. (2020) show that the excess risk of f_n is controlled by the one of g_n enabling them to carry out their analysis. The following comparison inequality and convergence rate are derived:

$$\mathcal{E}(f) - \mathcal{E}(f^*) \lesssim \sqrt{\mathcal{R}(g) - \mathcal{R}(g^*)} \leq O\left(n^{-1/4} \log\left(\delta^{-1}\right)\right) \quad \text{w.p. } 1 - \delta$$

where $\lesssim$ does not take into account multiplicative constants independent of n and δ.

Limitations of previous works This work is limited to the batch statistical framework with i.i.d. data. However some applications involve a flow of data; or data generated by non-stationary distributions including adversarial data. Our work is the first to study structured prediction in the setting of arbitrary sequences.

3 A General Algorithm for Online Structured Prediction

In this section we introduce our algorithm *OSKAAR (Online Structured prediction with Kernel Aggregating Algorithm Regression)* and bound its regret.

3.1 Introducing our Algorithm: *OSKAAR*

To simplify notations, we may denote $\varphi(y_t)$ by $\varphi_t \in \mathcal{H}$. We recall that the feature map $\varphi : \mathcal{Y} \to \mathcal{H}$ is constant over time, the index t in this notation denotes only the variation of y_t over time.

Algorithm 1: *OSKAAR – Online Structured prediction with Kernel Aggregating Algorithm Regression*

Input: $\lambda > 0$, kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
for *Each time step* t *in* $1 \dots T$ **do**
  Get information $x_t \in \mathcal{X}$
  Update $\beta^t(x) = (K_t + \lambda I)^{-1} v_t(x)$ where K_t and v_t are defined after Eq. (10)
  $\hat{z}_t = \arg\min_{z \in \mathcal{Z}} \langle \psi(z), \hat{g}_t(x_t) \rangle_{\mathcal{H}} = \arg\min_{z \in \mathcal{Z}} \sum_{s=1}^{t-1} \beta_s^t(x_t) \Delta(z, y_s)$
  Observe ground truth $y_t \in \mathcal{Y}$
  Get loss $\Delta(\hat{z}_t, y_t) \in \mathbb{R}$
end

We introduce our first algorithm, see Algorithm 1, which is inspired by the learning procedure of Ciliberto et al. (2020). However, we use a variant of Kernel Ridge Regression that has a different regularisation which is crucial in the context of arbitrary data, *Kernel Aggregating Algorithm Regression (KAAR)*, see Gammerman et al. (2012); Jézéquel et al. (2019). At each time step $t \in [\![T]\!]$, we compute $\hat{g}_t : \mathcal{X} \to \mathcal{H}$ as follows

$$\hat{g}_t := \arg\min_{g \in \mathcal{G}} \sum_{s=1}^{t-1} \|g(x_s) - \varphi_s\|_{\mathcal{H}}^2 + \lambda \|g\|_{\mathcal{G}}^2 + \|g(x_t)\|_{\mathcal{H}}^2 \,, \tag{8}$$

where $\mathcal{G}$ is a vRKHS with feature map ϕ such that $\sup_{x \in \mathcal{X}} \|\phi(x)\| \leq \kappa < \infty$, see Appendix A for more details. And f_t is defined as an optimisation problem with respect to $\hat{g}_t$ as in the batch setting

$$f_t(x) := \arg\min_{z \in \mathcal{Z}} \langle \psi(z), \hat{g}_t(x) \rangle_{\mathcal{H}} = \arg\min_{z \in \mathcal{Z}} \sum_{s=1}^{t-1} \beta_s^t(x) \Delta(z, y_s) \,, \tag{9}$$

where the coefficients β_s^t come from the representer theorem, and are defined as follows

$$\beta^t(x) = (K_t + \lambda I)^{-1} v_t(x) \tag{10}$$

with $K_t \in \mathbb{R}^{t \times t}$ the Gram matrix defined by $(K_t)_{i,j} = k(x_i, x_j)$, and $v_t(x) \in \mathbb{R}^t$ defined by $(v_t(x))_s = k(x, x_s)$. Thus, at each time step t, the prediction is computed by

$$\hat{z}_t := f_t(x_t) = \arg\min_{z \in \mathcal{Z}} \sum_{s=1}^{t-1} \beta_s^t(x_t) \Delta(z, y_s). \tag{11}$$

Hence, we note that, as in the supervised learning setting, the mathematical objects $\psi, \varphi, \mathcal{H}$ introduced in the definition of ILE are not needed to make a prediction. We only need the knowledge of the different labels y_s in order to compute $\Delta(., y_s)$.

3.2 Regret Bound of *OSKAAR*

We start our analysis by proving a comparison inequality, see Lemma 1. It extends any bound on the empirical risk of $(\hat{g}_t)_t$ to a bound on the regret of $(f_t)_t$. We can therefore carry out the analysis on $(\hat{g}_t)_t$ for which we have a closed form solution and lies in a space with algebraic assumptions.

Lemma 1 (Online Comparison Inequality)**.** *Let* $(f_t)_t$ *and* $(\hat{g}_t)_t$ *be defined as in* (9) *and* (8) *respectively. Then we have*

$$R_T \leq 2 c_\Delta \sqrt{T} \sqrt{\sum_{t=1}^{T} ||\varphi_t - \hat{g}_t(x_t)||_{\mathcal{H}}^2} \,. \tag{12}$$

Compared to Ciliberto et al. (2020), our online comparison inequality does not provide an upper bound with respect to a global minimiser g^*. The baseline is not reflected in the right hand side of this inequality. Finding a comparison inequality which controls the regret with respect to a baseline in the same class of functions than our estimators is left for future works. However, this result still shifts the problem into the feature space which possesses a lot more algebraic properties. It allows us to derive the regret bound of Theorem 2. The regret bound is expressed with respect to the effective dimension $d_{\text{eff}}(\lambda)$ (Rudi et al., 2016; Zadorozhnyi et al., 2021) defined by

$$d_{\text{eff}}(\lambda) := \text{Tr}(K(K+\lambda I)^{-1}) \quad \forall \lambda > 0 \tag{13}$$

where $K \in \mathbb{R}^{T\times T}$ is the Gram matrix at time T. The effective dimension measures the complexity of the underlying RKHS based on a given data sample. It is a decreasing function of the scale parameter λ and $d_{\text{eff}}(\lambda) \to 0$ when $\lambda \to \infty$. And when $\lambda \to 0$ it converges to the rank of K. Moreover it is always upper bounded by $d_{\text{eff}}(\lambda) \leq \kappa^2 T/\lambda$. We obtain the following regret bound.

Theorem 2 (Regret Bound of *OSKAAR*)**.** *Let $(f_t)_t$ be defined as in* (9)*. Then for all $\lambda > 0$ and $T \geq 1$ we have*

$$R_T \leq 2c_\Delta \sqrt{T} \sqrt{\log\left(e + \frac{e\kappa^2 T}{\lambda}\right) d_{\text{eff}}(\lambda) + \min_{g\in\mathcal{G}} L_T(g)} \tag{14}$$

where $L_t(g) := \sum_{s=1}^t ||\varphi_s - g(x_s)||_{\mathcal{H}}^2 + \lambda ||g||_{\mathcal{G}}^2$ for every $g \in \mathcal{G}$.

The proof of this statement is postponed to Appendix A. In the worst case scenario, if φ_t is a Rademacher variable, $\min_{g\in\mathcal{G}} L_T(g)$ is linear in T yields a linear regret bound. This is due to the fact that we are comparing our model to the best possible $z \in \mathcal{Z}$ at each time step, which is much too rich and linear regret is unavoidable in the worst case. On the other hand, if there is a function $g^* \in \mathcal{G}$ that perfectly models the features $(\varphi_t)_t$, i.e. $g^*(x_t) = \varphi_t$ for all t, by taking $\lambda = \sqrt{T}/\|g^*\|_{\mathcal{G}}$ and bounding the effective dimension by $d_{\text{eff}}(\lambda) \leq \kappa^2 T/\lambda$, we obtain $R_T \leq O\left(T^{3/4}\right)$. However, the assumption $\sum_{t=1}^T \|g^*(x_t) - \varphi_t\|_{\mathcal{H}}^2 = 0$ is too strong for adversarial data and even for i.i.d. data with white noise. These considerations motivate the study of the expected regret in the next section and the cumulative risk in Appendix C.

Computation time At each time step t, we need to compute the vector $\beta^t(x_t) \in \mathbb{R}^t$. Thus the per round complexity is of $O(t^2)$. If the kernel satisfies the capacity condition $d_{\text{eff}}(\lambda) \leq (T/\lambda)^\beta$ for $\beta \in [0,1]$ (see Appendix D.1), using a method based on Nyström approximation (Jézéquel et al., 2019), it is possible to recover the same regret with a computational complexity of $O(d_{\text{eff}}(\lambda)^{4/(1-\lambda)})$.

4 Stochastic Regret Bounds

In this section, we generalize the results from the supervised learning setting in Ciliberto et al. (2020). We achieve the same convergence rate as in the batch statistical framework, although our results hold without stochastic assumptions. We are now interested in bounding the expected regret $\mathbb{E}[R_T]$, where the expectation is taken over the possible randomness of the data $(x_1, y_1, \ldots, x_T, y_T)$. Note that the data are still generated sequentially and can be adapted to the player, in particular they can be adversarial and follow Dirac distributions. Note that the data are still generated sequentially and can adapt to the player, meaning they can be adversarial and follow Dirac distributions. Taking the expectation helps to avoid the noise inherent in the data and enables us to obtain results closer to Ciliberto et al. (2020) by replacing φ_t with to $\mathbb{E}[\varphi_t|x_t]$ in our result. We study the same algorithm (*OSKAAR*) as in the previous section, see Algorithm 1. We obtain the following regret bound and its corollary proved in Appendix B.

Theorem 3 (Expected Regret Bound)**.** *Let $(f_t)_t$ be defined as in* (9)*. Then, for any $g^* \in \mathcal{G}$, $\lambda > 0$ and $T \geq 1$, we have*

$$\mathbb{E}[R_T] \leq 2c_\Delta \sqrt{T} \sqrt{d_{\text{eff}}(\lambda) \log\left(e + \tfrac{e\kappa^2 T}{\lambda}\right) + \lambda \|g^*\|_{\mathcal{G}}^2 + \mathbb{E}\left[\sum_{t=1}^T \|g^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2\right]}. \tag{15}$$

Corollary 4 (Expected Regret Bound)**.** *With the same assumptions than Theorem 3, with $\lambda = \sqrt{T}$. Assume that there exists $g^* \in \mathcal{G}$ such that $\mathbb{E}\left[\sum_{t=1}^T \|g^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2\right] = 0$. Then, we have*

$$\mathbb{E}[R_T] \leq 2c_\Delta T^{3/4} \sqrt{\kappa^2 \log\left(e + e\kappa^2\sqrt{T}\right) + \|g^*\|_{\mathcal{G}}^2} = O\left(T^{3/4}\sqrt{\log T}\right). \tag{16}$$

Our assumption on g^* is similar to the one done to obtain the convergence rate in Ciliberto et al. (2020). It is a common assumption in Kernel Ridge Regression theory (Caponnetto and De Vito, 2007; Steinwart and Christmann, 2008). We

are assuming that there exists a function $g^* \in \mathcal{G}$, such that for all $t \in [\![T]\!]$ we have $g^*(x_t) = \mathbb{E}[\varphi_t|x_t]$. That it to say, g^* interpolates the expectations of the data. Up to the log factor, we retrieve the same bound in the online setting as in the supervised setting, and without assuming that the data are i.i.d. Specifically, we make no assumption on the x_t, generalizing existing results that assume i.i.d. inputs.

High probability regret bound In Theorem 3 and Corollary 4, we bound the expectation of the regret. The expectation is taken over the whole data, including the whole history at each time step. Moreover a bound in expectation does not necessarily imply a bound in high probability. It is however possible to obtain a bound in high probability on the cumulative risk, defined as

$$\sum_{t=1}^{T} \mathbb{E}_{y_t}[\Delta(f_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)]\,, \tag{17}$$

where at each round the expectation is taken with respect to the randomness of the next output y_t only and not with all past data $x_1, y_1, \ldots, x_{t-1}, y_{t-1}, x_t$. Computing such bounds requires more recent mathematical tools such as van der Hoeven et al. (2023). In Appendix C, we prove a regret bound in high probability using a slightly different estimator.

High probability excess risk bound In Appendix C.1, we demonstrate that, when data are i.i.d., our previous results enable the design of a batch estimator $\bar{f}_T$ from the online predictors, yielding the following bound on the excess risk. With probability $1-\delta$

$$\mathbb{E}_{x,y}[\Delta(\bar{f}_T(x), y) - \Delta(f^*(x), y)] \leq O\left(T^{-1/4}\sqrt{\log(T)} + T^{-1/2}\sqrt{\log(\delta^{-1})}\right).$$

A standard online to batch conversion would have aggregated the predictors $(f_t)_t$ by setting $\bar{f}_t = \sum_{t=1}^{T} f_t$. However, this is not possible here because the output space $\mathcal{Z}$ is not convex. To design $\bar{f}_T$, we thus aggregate the feature estimators $(\hat{g}_t)_t$ into a unique function $\bar{g}_T$, which is used to construct $\bar{f}_T$. This construction requires recent technical tools (van der Hoeven et al., 2023). To sum up, our algorithm generalizes the supervised learning setting, achieving the same convergence rate up to a log factor. Moreover, our algorithm can learn from a stream of data, allowing sequential updates as data arrive step by step, instead of relying on a batch of data available from the start.

5 Non-Stationary Online Structured Prediction

In this section, we introduce *SALAMI (Structured prediction ALgorithm with Aggregating MIxture)*, see Algorithm 2, an algorithm designed to handle non-stationary data distributions, including adversarial data. The non-stationarity we consider is on $x \mapsto \mathbb{E}[\varphi_t|x]$ rather than on the baseline $(f_t^*)_t$, which is already non-stationary throughout the paper. We compare the feature predictors $(\hat{g}_t)_t$ to a non-stationary baseline $(g_t^*)_t \in \mathcal{G}^T$. This approach allows us to address data with a changing distribution over time, including adversarial data. Note that we handle general data distributions, including Dirac distributions. As the data distributions change, earlier data may become outdated. Therefore, we need to modify our previous predictor, which considers all past data equally. We treat predictors with different starting times as experts and use an expert selection algorithm to create a mixture of them. See Algorithm 2 for details.

In the previous section, we assume the existence of some fixed function $g^* \in \mathcal{G}$ such that $g^*(x_t) = \mathbb{E}[\varphi_t|x_t]$ for all t, which is weak when $x \mapsto \mathbb{E}[\varphi_t|x]$ is stationary. However it is not satisfied when the data distribution is non-stationary or even arbitrary. In this section, we assume that for each time step $t \in [\![T]\!]$, there exists a function $g_t^* \in \mathcal{G}$ such that $g_t^*(x_t) = \mathbb{E}[\varphi_t|x_t]$. This is a very weak assumption, as we can choose a different function g_t^* for each time step.

5.1 Regret Bound

In order to bound the regret, we define two quantities that measure the non-stationarity of the sequence (g_t^*): the continuous variation $V_{\mathcal{G}}$ and the discrete variation V_0 defined as follows

$$V_{\mathcal{G}} := \|g_1^*\|_{\mathcal{G}} + \sum_{t=2}^{T} \|g_t^* - g_{t-1}^*\|_{\mathcal{G}} \qquad \text{and} \qquad V_0 := 1 + \sum_{t=2}^{T} 1[g_t^* \neq g_{t-1}^*]. \tag{18}$$

We obtain the following regret bound.

Theorem 5 (Expected Regret in a Non-Stationary Environment)**.** *Assume that there exists (g_t^*) a sequence in $\mathcal{G}$ such that $\mathbb{E}\big[\sum_{t=1}^{T} \|g_t^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2\big] = 0$. Then, Algorithm 2 run with $\lambda > 0$ and $\eta = 1/2(\kappa \sup\|g\|_{\mathcal{G}} + 1)^2$ satisfies*

$$\mathbb{E}[R_T] = \begin{cases} \tilde{O}(V_{\mathcal{G}}^{1/6} T^{5/6}) & \text{if } \lambda = V_{\mathcal{G}}^{-1/3} T^{1/3} \\ \tilde{O}(V_0^{1/4} T^{3/4}) & \text{if } \lambda = V_0^{-1/2} T^{1/2} \end{cases}. \tag{19}$$

In Appendix D, we prove this statement and precise the constants and log terms in the bounds. Therefore our method obtains a sublinear regret with variations $V_{\mathcal{G}}, V_0$ up to T. As expected, we obtain a loss of performance when facing non-stationary data distributions compared to the stationary case in Theorem 4. With discrete distribution changes, the rate $\tilde{O}(T^{3/4})$ is unchanged compared to the stationary setting as soon as the number of changes remains constant.

Calibration of λ A limitation of Theorem 5 is the required knowledge of V_0 or $V_{\mathcal{G}}$ to tune the learning rate $\lambda > 0$. First, note that setting $\lambda = \Omega(T^{1/3})$ always yield a regret of order $\tilde{O}(T^{5/6})$, but at the cost of a worse dependence on the variation $V_{\mathcal{G}}$. Second, our algorithm can be easily adapted to calibrate λ automatically, by combining experts $\hat{g}_{s:t}^{(\lambda)}$ (see Eq. 21 and the algorithm details in the next section), indexed by both the starting time s and an hyperparameter λ, with λ chosen from a logarithmic finite grid.

Refined regret bounds under the capacity condition A standard assumption when learning on RKHS is the capacity condition that assumes the existence of some $\beta \in [0, 1]$ and $Q > 0$ for which $d_{\text{eff}}(\lambda) \leq Q(T/\lambda)^{\beta}$ for all $\lambda > 0$. This assumption is weak since it is always verified for $\beta = 1$ but smaller values of β yield improved computational complexity for the algorithm (see for instance Jézéquel et al. (2019)) and improved regret guarantees. We further discuss this assumption and provide a refined regret bound in Appendix D.1. In the particular case of the Gaussian kernel, the effective dimension statisfies $d_{\text{eff}}(\lambda) \leq (\log(T/\lambda))^d$ (Altschuler et al., 2019), where d is the dimension of the input space $\mathcal{X}$. In this case, our result leads to the regret bound

$$\mathbb{E}[R_T] = \tilde{O}\big(T^{3/4} V_{\mathcal{G}}^{1/4} (\lambda + 1)^{1/4}\big)\,, \tag{20}$$

which improves the generic rate of Theorem 5 from $\tilde{O}(T^{5/6})$ to $\tilde{O}(T^{3/4})$. In this case, the extension to non-stationarity comes at no cost in the regret rate as soon as $V_{\mathcal{G}}$ does not grow with time. More details are provided in Appendix D.1.

5.2 Algorithm Design

We detail below our algorithm *SALAMI*. Let (g_t^*) be the unknown sequence satisfying the assumption of Theorem 5. We start from the observation that if one could identify breaking times $(t_i)_{1 \leq i \leq T}$ at which the sequence g_t^* changes, one could restart *OSKAAR* (Algorithm 1) at each t_i, considering that g_t^* is fixed from t_i to $t_{i+1} - 1$. The high-level idea of *SALAMI* is to learn these restart times through a meta-aggregation procedure that combines estimators $\hat{g}_{s:t}$ of the sequence g_t^*, indexed by $s = 1, \dots, t$, each assuming that (g_t^*) is fixed from s to t; defined by following the *KAAR* estimator starting in s

$$\hat{g}_{s:t} := \arg\min_{g \in \mathcal{G}} \sum_{\tau=s}^{t-1} \|\varphi_\tau - g(x_\tau)\|_{\mathcal{H}}^2 + \lambda \|g\|_{\mathcal{G}}^2 + \|g(x_t)\|_{\mathcal{H}}^2. \tag{21}$$

For each $t \in [\![T]\!]$, the feature predictor $\hat{g}_t$ of *SALAMI* is then defined as a convex combination of $\hat{g}_{s:t}$ for $1 \leq s \leq t$. Formally, *SALAMI* learns a probability vector $p_t \in \Delta_T$ and defines

$$\hat{g}_t = \sum_{s=1}^{T} p_t(s) \hat{g}_{s:t}. \tag{22}$$

The next part of the algorithm is how to choose the weights $p_t(s)$. To do so, this is done by using the exponentially weighted average forecaster (EWA) $w_t \in \Delta_T$, which needs a small adaptation to deal with the fact that $\hat{g}_{s:t}$ only produces predictions for $t \geq s$. Following the idea of Gaillard et al. (2014) for sleeping experts, this can be done by defining the auxiliary losses

$$\tilde{\ell}_t(s) = \begin{cases} \ell_t(\hat{g}_{s:t}) \text{ if } s \leq t \\ \ell_t(\hat{g}_t) \text{ if } s > t \end{cases} \qquad \text{where} \qquad \ell_t(g) := \|\varphi_t - g(x_t)\|_{\mathcal{H}}^2.$$

That is by assigning the loss of the algorithm itself $\ell_t(\hat{g}_t)$ to all expert that are inactive. The weights $p_t(s)$ are then defined as: $p_1(1) = 1$ and for $t > 1$:

$$p_t(s) = \frac{w_t(s)}{\sum_{k=1}^{t} w_t(k)} \qquad \text{where} \qquad w_t(k) \propto \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_s(k)\right) \tag{23}$$

for some learning rate $\eta > 0$. Finally, *SALAMI* defines the predictor $\hat{f}_t : \mathcal{X} \to \mathcal{Z}$ as:

$$\hat{f}_t(x) := \arg\min_{z \in \mathcal{Z}} \langle \psi(z), \hat{g}_t(x) \rangle_{\mathcal{H}}. \tag{24}$$

Computational complexity Note that in its current form, *SALAMI* needs to consider an increasing number of experts $\hat{g}_{s:t}$ over time, which increases the per-round space and time complexities of *OSKAAR* by a factor of $O(t)$. However, this problem can be addressed by using more sophisticated intervals than $[s, t]$ as done in Zhang et al. (2017); György et al. (2012); Daniely et al. (2015), which reduces the overhead in complexities to a factor of $O(\log t)$. Extending our work to such intervals is straightforward, but we have chosen to restrict ourselves to intervals $[s, t]$ to simplify the understanding of the algorithm.

Algorithm 2: *SALAMI – Structured prediction ALgorithm with Aggregating MIxture*

Input: $\lambda > 0$, exp-concavity constant η of $(\ell_t)_t$, kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
for *Each time step* t *in* $1 \dots T$ **do**
  Get information $x_t \in \mathcal{X}$
  for *Each expert* s *in* $1 \dots t$ **do**
    Compute $\hat{g}_{s:t} := \arg\min_{g \in \mathcal{G}} \sum_{\tau=s}^{t-1} \|\varphi_\tau - g(x_\tau)\|_{\mathcal{H}}^2 + \lambda \|g\|_{\mathcal{G}}^2 + \|g(x_t)\|_{\mathcal{H}}^2$
  end
  for *Each expert* s *in* $1 \dots T$ **do**
    Compute EWA $w_t(s) \propto w_{t-1}(s) \exp(-\eta \tilde{\ell}_t(s))$ where $\tilde{\ell}_t(s) = \begin{cases} \ell_t(\hat{g}_{s:t}) \text{ if } s \le t \\ \ell_t(\hat{g}_t) \text{ if } s > t \end{cases}$
    Compute $p_t(s) \propto \begin{cases} w_t(s) \text{ if } s \le t \\ 0 \text{ if } s > t \end{cases}$
  end
  Compute the aggregate predictor $\hat{g}_t = \sum_{s=1}^{T} p_t(s) \hat{g}_{s:t}$
  Compute the prediction $\hat{z}_t = \hat{f}_t(x_t) = \arg\min_{z \in \mathcal{Z}} \langle \psi(z), \hat{g}_t(x_t) \rangle_{\mathcal{H}}$
  Observe ground truth $y_t \in \mathcal{Y}$
  Get loss $\Delta(\hat{z}_t, y_t) \in \mathbb{R}$
end

Acknowledgements. A.R. acknowledges the support of the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and the support of the European Research Council (grant REAL 947908).

References

J. Altschuler, F. Bach, A. Rudi, and J. Niles-Weed. Massively scalable sinkhorn distances via the nyström method, 2019.

M. A. Alvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: a review, 2012.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

C. Ciliberto, L. Rosasco, and A. Rudi. A General Framework for Consistent Structured Prediction with Implicit Loss Embeddings, 2020.

A. Daniely, A. Gonen, and S. Shalev-Shwartz. Strongly adaptive online learning. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1405–1411, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/daniely15.html`.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

J. C. Duchi, L. W. Mackey, and M. I. Jordan. On the consistency of ranking algorithms. In *ICML*, pages 327–334, 2010.

D. A. Forsyth and J. Ponce. *Computer vision: a modern approach*. prentice hall professional technical reference, 2002.

P. Gaillard, G. Stoltz, and T. van Erven. A second-order bound with excess losses. In *Proceedings of COLT'14*, volume 35, pages 176–196. JMLR: Workshop and Conference Proceedings, 2014.

A. Gammerman, Y. Kalnishkan, and V. Vovk. On-line prediction with kernels and the complexity approximation principle, 2012.

A. György, T. Linder, and G. Lugosi. Efficient tracking of large classes of experts. In *2012 IEEE International Symposium on Information Theory Proceedings*, pages 885–889, 2012. doi: 10.1109/ISIT.2012.6284689.

E. Hazan. Introduction to online convex optimization, 2023.

M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine learning*, 32(2):151–178, 1998.

T. Joachims, T. Hofmann, Y. Yue, and C.-N. Yu. Predicting structured objects with support vector machines. *Communications of the ACM*, 52(11):97–104, 2009.

R. Jézéquel, P. Gaillard, and A. Rudi. Efficient online learning with kernels for adversarial large scale problems, May 2019. URL `http://arxiv.org/abs/1902.09917`. arXiv:1902.09917 [cs, math, stat].

S. Lacoste-Julien, B. Taskar, D. Klein, and M. I. Jordan. Word alignment via quadratic assignment. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 112–119, 2006.

H. B. McMahan and F. Orabona. Unconstrained online linear learning in hilbert spaces: Minimax algorithms and normal approximations, 2014.

C. Micchelli and M. Pontil. Kernels for multi–task learning. *Advances in neural information processing systems*, 17, 2004.

F. Orabona. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2023. URL `http://arxiv.org/abs/1912.13213`.

A. Pacchiano, N. S. Chatterji, and P. L. Bartlett. Online learning with kernel losses, 2018.

A. Raj, P. Gaillard, and C. Saad. Non-stationary online regression. *arXiv preprint arXiv:2011.06957*, 2020.

A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization, 2016.

I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

D. van der Hoeven, N. Zhivotovskiy, and N. Cesa-Bianchi. High-probability risk bounds via sequential predictors, 2023.

A. N. Vila. Structured supervised learning with theoretical guarantees, 2022.

O. Wintenberger. Stochastic online convex optimization. application to probabilistic time series forecasting. *Electronic Journal of Statistics*, 18(1):429–464, 2024.

O. Zadorozhnyi, P. Gaillard, S. Gerschinovitz, and A. Rudi. Online nonparametric regression with sobolev kernels, 2021.

L. Zhang, T. Yang, R. Jin, and Z.-H. Zhou. Dynamic regret of strongly adaptive methods. In *International Conference on Machine Learning*, 2017. URL `https://api.semanticscholar.org/CorpusID:49314405`.

APPENDIX

A Proof of Theorem 2: Regret Bound of *OSKAAR*

We first introduce some additional notations on kernels. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive semidefinite kernel, and $\mathcal{F} = \overline{\text{span}\{k(x,.)|x \in \mathcal{X}\}}$ its associated RKHS. We denote by $\phi : \mathcal{X} \to \mathcal{F}$ the feature map $\phi(x) = k(x,.)$. We assume ϕ to be bounded by $\|\phi(x)\|_{\mathcal{F}} \leq \kappa < \infty$. For more details on RKHS see Aronszajn (1950); Berlinet and Thomas-Agnan (2011). We may now introduce the following operators:

- $S_t : \mathcal{F} \to \mathbb{R}^t$, s.t. $f \in \mathcal{F} \mapsto (\langle \phi(x_s), f \rangle)_{s=1}^t$
- $S_t^* : \mathbb{R}^t \to \mathcal{F}$, s.t. $v = (v_i)_{i=1}^t \mapsto \sum_{s=1}^t v_i \phi(x_i)$
- $C_t = S_t^* S_t : \mathcal{F} \to \mathcal{F}$
- We have that $C_t = \sum_{s=1}^t \phi(x_s) \otimes \phi(x_s)$
- $K_t = S_t S_t^*$ is the empirical kernel matrix
- $A_\lambda = A + \lambda I$ for any symmetric linear operator A, where I is the identity and $\lambda \in \mathbb{R}$

For the space of function $\mathcal{G} : \mathcal{X} \to \mathcal{H}$ we choose an vector-valued RKHS $\mathcal{G} = \mathcal{H} \otimes \mathcal{F}$ (Micchelli and Pontil, 2004; Alvarez et al., 2012), which is a direct generalisation of scalar-valued RKHS.

We define L_t for $t \in [\![T]\!]$ in a more general setting and rewrite it using Hilbertian operators

$$L_t(g, g_t^*) = \sum_{s=1}^t \|g_s^* - g(x_s)\|_{\mathcal{H}}^2 + \lambda \|g\|_{\mathcal{G}}^2 = \|H_t\|^2 - 2g^* S_t^* H_t + \langle g, C_{t,\lambda} g \rangle$$

where H_t is the vector $(g_1^*, \ldots, g_t^*) \in \mathcal{H}^t$. And we denote by $L_t(g)$ the application $L_t(g, \varphi(Y_t))$, where $\varphi(Y_t)$ is the vector $(\varphi_1, \ldots, \varphi_t) \in \mathcal{H}^t$. We define the following functions

$$g_{t+1} = \arg\min_{g \in \mathcal{G}} L_t(g) = C_{t,\lambda}^{-1} S_t^* \varphi(Y_t)\,, \tag{25}$$

$$\hat{g}_{t+1} = \arg\min_{g \in \mathcal{G}} L_t(g) + \|g(x_{t+1})\|_{\mathcal{H}}^2 = C_{t+1,\lambda}^{-1} S_t^* \varphi(Y_t)\,. \tag{26}$$

$\hat{g}_{t+1}$ is directly used in the definition of the algorithm, while g_{t+1} is only used in the proof of its regret.

We recall and prove the results from in Section 3.2.

Lemma 1 (Online Comparison Inequality)**.** *Let $(f_t)_t$ and $(\hat{g}_t)_t$ be defined as in* (9) *and* (8) *respectively. Then we have*

$$R_T \leq 2c_\Delta \sqrt{T} \sqrt{\sum_{t=1}^T ||\varphi_t - \hat{g}_t(x_t)||_{\mathcal{H}}^2}\,. \tag{12}$$

Proof. We add and subtract two terms.

$$\begin{aligned}
R_T &= \sum_{t=1}^T \Delta(f_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t) \\
&= \sum_{t=1}^T \langle \psi(f_t(x_t)), \varphi_t \rangle - \langle \psi(f_t(x_t)), \hat{g}_t(x_t) \rangle \\
&\quad + \langle \psi(f_t(x_t)), \hat{g}_t(x_t) \rangle - \langle \psi(f_t^*(x_t)), \hat{g}_t(x_t) \rangle \\
&\quad + \langle \psi(f_t^*(x_t)), \hat{g}_t(x_t) \rangle - \langle \psi(f_t^*(x_t)), \varphi_t \rangle \\
&\leq \sum_{t=1}^T \langle \psi(f_t(x_t)), \varphi_t \rangle - \langle \psi(f_t(x_t)), \hat{g}_t(x_t) \rangle \\
&\quad + \langle \psi(f_t^*(x_t)), \hat{g}_t(x_t) \rangle - \langle \psi(f_t^*(x_t)), \varphi_t \rangle \\
&= \sum_{t=1}^T \langle \psi(f_t(x_t)) - \psi(f_t^*(x_t)), \varphi_t - \hat{g}_t(x_t) \rangle
\end{aligned}$$

where the inequality comes from the definition of f_t. We now apply successively Cauchy-Schwartz and Jensen's inequalities to conclude the proof.

$$\begin{aligned} R_T &\leq \sum_{t=1}^{T} ||\psi(f_t(x_t)) - \psi(f_t^*(x_t))||_{\mathcal{H}} \cdot ||\varphi_t - \hat{g}_t(x_t)||_{\mathcal{H}} \\ &\leq 2c_\Delta \sum_{t=1}^{T} ||\varphi_t - \hat{g}_t(x_t)||_{\mathcal{H}} \\ &\leq 2c_\Delta \sqrt{T} \sqrt{\sum_{t=1}^{T} ||\varphi_t - \hat{g}_t(x_t)||_{\mathcal{H}}^2} \end{aligned}$$

□

We bound the regret of the *KAAR* estimator. We generalise the proof of Jézéquel et al. (2019) to vRKHS.

Lemma 6 (General Regret *KAAR* Estimator)**.** *Let $(h_t)_{t=1}^T$ be bounded vectors in $\mathcal{H}$ such that $\|h_t\|_{\mathcal{H}} \leq B < \infty$ for all $t \in [\![T]\!]$. Let $H_T \in \mathcal{H}^T$ be the vector $(h_1, \ldots, h_T)$. Let $\lambda > 0$, and let us define the KAAR predictors as follow*

$$\hat{g}_t = \arg\min_{g \in \mathcal{G}} \sum_{s=1}^{t-1} \|h_s - g(x_s)\|_{\mathcal{H}}^2 + \lambda \|g\|_{\mathcal{G}}^2 + \|g(x_t)\|_{\mathcal{H}}^2.$$

Then we have

$$\sum_{t=1}^{T} \|h_t - \hat{g}_t(x_t)\|_{\mathcal{H}}^2 \leq B^2 \log\left(e + \frac{e\kappa^2 T}{\lambda}\right) d_{\mathit{eff}}(\lambda) + \min_{g \in \mathcal{G}} L_T(g, H_T). \tag{27}$$

Proof. We follow and adapt the proof of Theorem 9 from Jézéquel et al. (2019) to vector RKHS without Nyström approximation. We start by adding telescopic terms.

$$\begin{aligned} &\sum_{t=1}^{T} \|h_t - \hat{g}_t(x_t)\|_{\mathcal{H}}^2 \\ &\quad = \sum_{t=1}^{T} \|h_t - \hat{g}_t(x_t)\|_{\mathcal{H}}^2 - L_T(g_{T+1}, H_T) + L_T(g_{T+1}, H_T) \\ &\quad = \sum_{t=1}^{T} \left[\|h_t - \hat{g}_t(x_t)\|_{\mathcal{H}}^2 + L_{t-1}(g_t, H_{t-1}) - L_t(g_{t+1}, H_t)\right] + L_T(g_{T+1}, H_T)\,. \end{aligned}$$

Let $Z(t) = \|h_t - \hat{g}_t(x_t)\|_{\mathcal{H}}^2 + L_{t-1}(g_t, H_{t-1}) - L_t(g_{t+1}, H_t)$, and let us study its terms separately.

Note that $\langle g_{t+1}, C_{t,\lambda} g_{t+1} \rangle = H_t^* S_t C_{t,\lambda}^{-1} C_{t,\lambda} g_{t+1} = H_t^* S_t g_{t+1}$.

Therefore $L_t(g_{t+1}, H_t) = \|H_t\|^2 - \langle g_{t+1}, C_{t,\lambda} g_{t+1} \rangle$.

Let us focus now on $\|h_t - \hat{g}_t(x_t)\|^2 = \|h_t\|^2 - 2\langle h_t, \hat{g}_t(x_t) \rangle + \langle \hat{g}_t(x_t), \hat{g}_t(x_t) \rangle$. Note that

$$\begin{aligned} \langle h_t, \hat{g}_t(x_t) \rangle &= \hat{g}_t^* \phi(x_t) h_t \\ &= \hat{g}_t^* (S_t^* H_t - S_{t-1}^* H_{t-1}) \\ &= \hat{g}_t^* (C_{t,\lambda} g_{t+1} - C_{t-1,\lambda} g_t) \\ &= \langle \hat{g}_t, C_{t,\lambda} g_{t+1} - C_{t-1,\lambda} g_t \rangle \end{aligned}$$

and

$$\langle \hat{g}_t(x_t), \hat{g}_t(x_t) \rangle = \langle \hat{g}_t, \phi(x_t) \langle \phi(x_t), \hat{g}_t \rangle \rangle = \langle \hat{g}_t, [\phi(x_t) \otimes \phi(x_t)] \hat{g}_t \rangle = \langle \hat{g}_t, (C_{t,\lambda} - C_{t-1,\lambda}) \hat{g}_t \rangle\,.$$

Therefore we obtain

$$\|h_t - \hat{g}_t(x_t)\|^2 = \|h_t\|^2 - 2\langle \hat{g}_t, C_{t,\lambda} g_{t+1} - C_{t-1,\lambda} g_t \rangle + \langle \hat{g}_t, (C_{t,\lambda} - C_{t-1,\lambda}) \hat{g}_t \rangle\,.$$

Thus by putting everything together, we get

$$\begin{aligned} Z(t) &= -2\langle \hat{g}_t, C_{t,\lambda} g_{t+1} - C_{t-1,\lambda} g_t \rangle + \langle \hat{g}_t, (C_{t,\lambda} - C_{t-1,\lambda})\hat{g}_t \rangle - \langle g_t, C_{t-1,\lambda} g_t \rangle + \langle g_{t+1}, C_{t,\lambda} g_{t+1} \rangle \\ &= \langle \hat{g}_t - g_{t+1}, C_{t,\lambda}(\hat{g}_t - g_{t+1}) \rangle - \langle \hat{g}_t - g_t, C_{t-1,\lambda}(\hat{g}_t - g_t) \rangle \\ &\leq \langle \hat{g}_t - g_{t+1}, C_{t,\lambda}(\hat{g}_t - g_{t+1}) \rangle \,. \end{aligned}$$

Now note that we can factorise

$$\hat{g}_t - g_{t+1} = C_{t,\lambda}^{-1} S_{t-1}^* H_{t-1} - C_{t,\lambda}^{-1} S_t^* H_t = -C_{t,\lambda}^{-1} \phi(x_t) h_t \,.$$

We thus bound $Z(t)$ by

$$Z(t) \leq \langle \phi(x_t) h_t, C_{t,\lambda}^{-1} \phi(x_t) h_t \rangle = \|h_t\|^2 \langle \phi(x_t), C_{t,\lambda}^{-1} \phi(x_t) \rangle \leq B^2 \langle \phi(x_t), C_{t,\lambda}^{-1} \phi(x_t) \rangle \,.$$

Finally, we proved that

$$\sum_{t=1}^{T} \|h_t - g(x_t)\|_{\mathcal{H}}^2 \leq B^2 \sum_{t=1}^{T} \langle \phi(x_t), C_{t,\lambda}^{-1} \phi(x_t) \rangle + L_T(g_{T+1}, H_T) \,.$$

We conclude the proof by using Propositions 1 and 2 of Jézéquel et al. (2019). □

We now prove our main result from Section 3.2.

Theorem 2 (Regret Bound of *OSKAAR*)**.** *Let $(f_t)_t$ be defined as in* (9)*. Then for all $\lambda > 0$ and $T \geq 1$ we have*

$$R_T \leq 2 c_\Delta \sqrt{T} \sqrt{\log\left(e + \frac{e\kappa^2 T}{\lambda}\right) d_{\mathit{eff}}(\lambda) + \min_{g \in \mathcal{G}} L_T(g)} \tag{14}$$

where $L_t(g) := \sum_{s=1}^{t} ||\varphi_s - g(x_s)||_{\mathcal{H}}^2 + \lambda ||g||_{\mathcal{G}}^2$ for every $g \in \mathcal{G}$.

Proof. We apply the two previous lemmas and obtain

$$\begin{aligned} R_T \quad &= \quad \sum_{t=1}^{T} \Delta(f_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t) \\ &\overset{\text{(Lem. 1)}}{\leq} \quad 2 c_\Delta \sqrt{T} \sqrt{\sum_{t=1}^{T} ||\varphi(y_t) - \hat{g}_t(x_t)||_{\mathcal{H}}^2} \\ &\overset{\text{(Lem. 6)}}{\leq} \quad 2 c_\Delta \sqrt{T} \sqrt{\log\left(e + \frac{e\kappa^2 T}{\lambda}\right) d_{\text{eff}}(\lambda) + \min_{g \in \mathcal{G}} L_T(g)} \end{aligned}$$

where $B = 1 \geq \sup_{y \in \mathcal{Y}} \|\varphi(y)\|_{\mathcal{H}}$. □

B Proofs of Theorem 3 and Corollary 4: Stochastic Regret Bounds in Expectation

In this section, we prove the results from Section 4. We will denote by $\mathcal{F}_{t-1}$ the filter $(x_1, y_1, \ldots, x_t)$. We start by introducing the following comparison inequality. It is an equivalent to Lemma 1 for the expected regret. It allows to control the expected regret with respect to $\mathbb{E}[\varphi_t | x_t]$.

Lemma 7 (Comparison Inequality in Expectation)**.** *For any sequence of measurable functions $(\hat{g}_t : \mathcal{X} \to \mathcal{H})_t$. For all $t \in [\![T]\!]$, let $f_t : \mathcal{X} \to \mathcal{Z}$ be defined by $f_t(x) = \arg\min_{z \in \mathcal{Z}} \langle \psi(z), \hat{g}_t(x) \rangle_{\mathcal{H}}$. Then we have*

$$\mathbb{E}[R_T] \leq 2 c_\Delta \sqrt{T} \sqrt{\sum_{t=1}^{T} \mathbb{E}[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t | x_t]\|_{\mathcal{H}}^2]} \,. \tag{28}$$

Proof. We follow the proof from Lemma 1 and get

$$\mathbb{E}[\Delta(f_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)] \leq \mathbb{E}[\langle \psi(f_t(x_t)) - \psi(f_t^*(x_t)), \hat{g}_t(x_t) - \varphi_t \rangle] \,.$$

Remember that $\varphi_t := \varphi(y_t)$ depends on y_t. Now note that

$$\begin{aligned}\mathbb{E}[\langle \psi(f_t(x_t)) - \psi(f_t^*(x_t)), \mathbb{E}[\varphi_t|x_t] - \varphi_t\rangle] &= \langle \psi(f_t(x_t)) - \psi(f_t^*(x_t)), \mathbb{E}[\mathbb{E}[\varphi_t|x_t] - \varphi_t]\rangle \\ &= \langle \psi(f_t(x_t)) - \psi(f_t^*(x_t)), \mathbb{E}[\mathbb{E}_{y_t}[\mathbb{E}[\varphi_t|x_t] - \varphi_t|\mathcal{F}_{t-1}]]\rangle = 0\end{aligned}$$

since we condition on x_t in the expectation. Thus

$$\mathbb{E}[\Delta(f_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)] \leq \mathbb{E}[\langle \psi(f_t(x_t)) - \psi(f_t^*(x_t)), \hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\rangle]\,.$$

We now apply successively Cauchy-Schwartz inequality and Jensen's inequality as in the proof of Lemma 1 and obtain

$$\mathbb{E}[\Delta(f_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)] \leq 2c_\Delta\sqrt{T}\sqrt{\sum_{t=1}^T \mathbb{E}[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2]}\,. \tag{29}$$

It concludes the proof. □

We now recall and prove the expectation of the regret bound.

Theorem 3 (Expected Regret Bound)**.** *Let $(f_t)_t$ be defined as in* (9)*. Then, for any $g^* \in \mathcal{G}$, $\lambda > 0$ and $T \geq 1$, we have*

$$\mathbb{E}[R_T] \leq 2c_\Delta\sqrt{T}\sqrt{d_{\mathit{eff}}(\lambda)\log\left(e + \tfrac{e\kappa^2 T}{\lambda}\right) + \lambda\|g^*\|_{\mathcal{G}}^2 + \mathbb{E}\left[\sum_{t=1}^T \|g^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2\right]}\,. \tag{15}$$

Proof. We apply Lemma 7, then add and subtract a term to get

$$\begin{aligned}&\mathbb{E}[\Delta(f_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)] \\ &\leq 2c_\Delta\sqrt{T}\sqrt{\sum_{t=1}^T \mathbb{E}[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2]} \\ &= 2c_\Delta\sqrt{T}\sqrt{\sum_{t=1}^T \mathbb{E}[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2 - \|g^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2 + \|g^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2]}\,.\end{aligned}$$

We bound the difference between the first two terms.

$$\begin{aligned}&\sum_{t=1}^T \mathbb{E}[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2 - \|g^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2] \\ &\qquad = \sum_{t=1}^T \mathbb{E}[\mathbb{E}_{y_t}[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2 - \|g^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2|\mathcal{F}_{t-1}]]\end{aligned}$$

Now note that

$$\begin{aligned}\mathbb{E}_{y_t}[\|\hat{g}_t(x_t) - \varphi_t\|^2|\mathcal{F}_{t-1}] &= \mathbb{E}_{y_t}[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|^2 + \|\varphi_t - \mathbb{E}[\varphi_t|x_t]\|^2|\mathcal{F}_{t-1}] \\ &\quad -2\mathbb{E}_{y_t}[\langle \mathbb{E}[\varphi_t|x_t] - \varphi_t, \mathbb{E}[\varphi_t|x_t] - \hat{g}_t(x_t)\rangle|\mathcal{F}_{t-1}] \\ &= \mathbb{E}_{y_t}[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|^2 + \|\varphi_t - \mathbb{E}[\varphi_t|x_t]\|^2|\mathcal{F}_{t-1}]\end{aligned}$$

since we condition on x_t. The same equality holds for g^*. Therefore we obtain

$$\begin{aligned}&\sum_{t=1}^T \mathbb{E}[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2 - \|g^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2] \\ &\quad = \sum_{t=1}^T \mathbb{E}[\|\hat{g}_t(x_t) - \varphi_t\|_{\mathcal{H}}^2 - \|g^*(x_t) - \varphi_t\|_{\mathcal{H}}^2] \\ &\quad = \mathbb{E}\left[\sum_{t=1}^T \|\hat{g}_t(x_t) - \varphi_t\|_{\mathcal{H}}^2 - \|g^*(x_t) - \varphi_t\|_{\mathcal{H}}^2\right] \\ &\quad \leq \mathbb{E}\left[d_{\text{eff}}(\lambda)\log\left(e + \tfrac{e\kappa^2 T}{\lambda}\right) + \lambda\|g^*\|_{\mathcal{G}}^2\right] \\ &\quad = d_{\text{eff}}(\lambda)\log\left(e + \tfrac{e\kappa^2 T}{\lambda}\right) + \lambda\|g^*\|_{\mathcal{G}}^2\end{aligned}$$

where the inequality comes from Theorem 6. It concludes the proof. □

Corollary 4 (Expected Regret Bound)**.** *With the same assumptions than Theorem 3, with $\lambda = \sqrt{T}$. Assume that there exists $g^* \in \mathcal{G}$ such that $\mathbb{E}\left[\sum_{t=1}^{T} \|g^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2\right] = 0$. Then, we have*

$$\mathbb{E}[R_T] \leq 2c_\Delta T^{3/4} \sqrt{\kappa^2 \log\left(e + e\kappa^2\sqrt{T}\right) + \|g^*\|_{\mathcal{G}}^2} = O\left(T^{3/4}\sqrt{\log T}\right). \tag{16}$$

Proof. We bound the effective dimension by $d_{\text{eff}}(\lambda) \leq \frac{\kappa^2 T}{\lambda}$ and obtain

$$\mathbb{E}\left[R_T\right] \leq 2c_\Delta \sqrt{T} \sqrt{\tfrac{\kappa^2 T}{\lambda} \log\left(e + \tfrac{e\kappa^2 T}{\lambda}\right) + \lambda \|g^*\|_{\mathcal{G}}^2}\,.$$

By choosing $\lambda = \sqrt{T}$, we get

$$\mathbb{E}\left[R_T\right] \leq 2c_\Delta T^{3/4} \sqrt{\kappa^2 \log\left(e + e\kappa^2\sqrt{T}\right) + \|g^*\|_{\mathcal{G}}^2}\,.$$

□

C Stochastic Regret Bounds in High Probability

In this section we aim to retrieve the results from the supervised learning framework (Ciliberto et al., 2020) in the online learning setting. For all time step t, we define the filter $\mathcal{F}_{t-1} = (x_1, y_1, \ldots, x_{t-1}, y_{t-1}, x_t)$, and we denote by $\mathbb{E}_t[.]$ the expectation $\mathbb{E}_{y_t}[.|\mathcal{F}_{t-1}]$. We are now interested in bounding the cumulative risk (Wintenberger, 2024)

$$\sum_{t=1}^{T} \mathbb{E}_t[\Delta(f_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)]\,. \tag{30}$$

Taking the expectation in y_t will avoid considering the noise of the random variables $(y_t)_t$ and allow us to obtain closer results from Ciliberto et al. (2020) by bounding with respect to $\mathbb{E}[\varphi_t|x_t]$ instead of φ_t.

We use the proof of Theorem 1 of van der Hoeven et al. (2023), which allows to bound the cumulative risk in the feature space with high probability using a regret bound for an exp-concave loss. Moreover, van der Hoeven et al. (2023) enables us to aggregate our predictors into a unique function $\bar{f}_T$ and bound its cumulative risk in high probability, which is a setting similar to the supervised learning study. In order to apply this theorem, we need to modify our feature predictor $\hat{g}_t : \mathcal{X} \to \mathcal{H}$ and define it using a shifted version of the losses

$$\hat{g}_t = \arg\min_{g \in \mathcal{G}} \sum_{s=1}^{t-1} \|\tfrac{1}{2} g(x_s) + \tfrac{1}{2}\hat{g}_s(x_s) - \varphi_s\|_{\mathcal{H}}^2 + \lambda\|g\|_{\mathcal{G}}^2 + \tfrac{1}{4}\|g(x_t)\|_{\mathcal{H}}^2\,. \tag{31}$$

The predictor $f_t : \mathcal{X} \to \mathcal{Z}$ is then defined as follows

$$f_t(x) = \arg\min_{z \in \mathcal{Z}} \langle \psi(z), \hat{g}_t(x)\rangle_{\mathcal{H}} = \arg\min_{z \in \mathcal{Z}} \sum_{s=1}^{t-1} \beta_s(x)\Delta(z, y_s)\,. \tag{32}$$

As previously, we do not require the knowledge of ψ, φ and $\mathcal{H}$ to make a prediction.

We first introduce a comparison inequality. It is an equivalent to Lemma 1 for the cumulative risk.

Lemma 8 (Comparison Inequality for Cumulative Risk)**.** *For any sequence of measurable functions $(\hat{g}_t : \mathcal{X} \to \mathcal{H})_t$. For all $t \in [\![T]\!]$, let $f_t : \mathcal{X} \to \mathcal{Z}$ be defined by $f_t(x) = \arg\min_{z \in \mathcal{Z}} \langle \psi(z), \hat{g}_t(x)\rangle_{\mathcal{H}}$. Then we have*

$$\sum_{t=1}^{T} \mathbb{E}_t[\Delta(f_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)] \leq 2c_\Delta \sqrt{T} \sqrt{\sum_{t=1}^{T} \mathbb{E}_t[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2]}\,.$$

Proof. We recall the notation $\varphi_t = \varphi(y_t)$, which therefore depends on y_t in the expectation. We start by adding and subtracting two terms

$$\begin{aligned}
&\mathbb{E}_t[\Delta(f_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)] \\
&\quad = \mathbb{E}_t[\langle \psi(f_t(x_t)), \varphi_t\rangle - \langle \psi(f_t^*(x_t)), \varphi_t\rangle] \\
&\quad = \mathbb{E}_t[\langle \psi(f_t(x_t)), \varphi_t\rangle - \langle \psi(f_t(x_t)), \hat{g}_t(x_t)\rangle \\
&\qquad + \langle \psi(f_t(x_t)), \hat{g}_t(x_t)\rangle - \langle \psi(f_t^*(x_t)), \hat{g}_t(x_t)\rangle \\
&\qquad + \langle \psi(f_t^*(x_t)), \hat{g}_t(x_t)\rangle - \langle \psi(f_t^*(x_t)), \varphi_t\rangle] \\
&\quad \leq \mathbb{E}_t[\langle \psi(f_t(x_t)), \varphi_t\rangle - \langle \psi(f_t(x_t)), \hat{g}_t(x_t)\rangle \\
&\qquad + \langle \psi(f_t^*(x_t)), \hat{g}_t(x_t)\rangle - \langle \psi(f_t^*(x_t)), \varphi_t\rangle] \\
&\quad = \mathbb{E}_t[\langle \psi(f_t(x_t)) - \psi(f_t^*(x_t)), \hat{g}_t(x_t) - \varphi_t\rangle]
\end{aligned}$$

where the inequality comes by definition of f_t. Now note that

$$\mathbb{E}_t[\langle \psi(f_t(x_t)) - \psi(f_t^*(x_t)), \mathbb{E}[\varphi_t|x_t] - \varphi_t\rangle] = \langle \psi(f_t(x_t)) - \psi(f_t^*(x_t)), \mathbb{E}_t[\mathbb{E}[\varphi_t|x_t] - \varphi_t]\rangle = 0$$

since we condition on x_t in $\mathbb{E}_t$. Thus

$$\begin{aligned}
&\mathbb{E}_t[\langle \psi(f_t(x_t)) - \psi(f_t^*(x_t)), \hat{g}_t(x_t) - \varphi_t\rangle] \\
&\quad = \mathbb{E}_t[\langle \psi(f_t(x_t)) - \psi(f_t^*(x_t)), \hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\rangle] \\
&\quad + \mathbb{E}_t[\langle \psi(f_t(x_t)) - \psi(f_t^*(x_t)), \mathbb{E}[\varphi_t|x_t] - \varphi_t\rangle] \\
&\quad = \mathbb{E}_t[\langle \psi(f_t(x_t)) - \psi(f_t^*(x_t)), \hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\rangle]\,.
\end{aligned}$$

We now apply successively Cauchy-Schwartz inequality and Jensen's inequality.

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}_t[\langle \psi(f_t(x_t)) - \psi(f_t^*(x_t)), \hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\rangle] &\leq 2c_\Delta \sum_{t=1}^T \mathbb{E}_t[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}] \\
&\leq 2c_\Delta \sqrt{T} \sqrt{\sum_{t=1}^T \mathbb{E}_t[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2]}
\end{aligned}$$

□

Let $\eta = 1/2(\kappa \sup\|g\|_{\mathcal{G}}+1)^2$ be an exp-concavity constant of the loss $\ell_t(g) = \|g(x_t)-\varphi_t\|_{\mathcal{H}}^2$. Let $m = 2(\kappa \sup\|g\|_{\mathcal{G}}+1)^2$ be such that $\ell(g) - \ell(g') \leq m$ for all $g, g' \in \mathcal{G}$. We define

$$\gamma = 4\max(\tfrac{1}{\eta}, m) \tag{33}$$

as in the Theorem 1 of van der Hoeven et al. (2023).

We bound the cumulative risk of our algorithm, see Theorem 9. Compared to our previous result Theorem 2, we obtain a bound in high probability. We now use $g^* \in \mathcal{G}$ to model $(\mathbb{E}[\varphi_t|x_t])_t$ instead of $(\varphi_t)_t$. This difference allows us not to consider the noise of the random variables $x_t \mapsto \varphi_t$. It also allows us to be closer to the framework of Ciliberto et al. (2020) that compares a model $g \in \mathcal{G}$ with the optimum and conditional expectation $x \mapsto \int_{\mathcal{Y}} \varphi(y) d\rho(y|x)$.

Theorem 9 (Average Cumulative Risk)**.** *Let $(f_t)_t$ be defined as in* (32)*. Let $\delta \in (0,1]$ and $\gamma = 8(\kappa \sup\|g\|_{\mathcal{G}} + 1)^2$. With $\lambda = \sqrt{T}$ and assuming that there exists a function $h \in \mathcal{G}$ such that for all $t \in [\![T]\!], h(x_t) = \mathbb{E}[\varphi_t|x_t]$, we have with probability $1-\delta$*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \mathbb{E}_t[\Delta(f_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)] \\
&\qquad \leq 2c_\Delta T^{-1/4} \sqrt{\tfrac{B^2\kappa^2}{8} \log\left(e + \tfrac{e\kappa^2\sqrt{T}}{4}\right) + 2\|g^*\|^2} + 2c_\Delta T^{-1/2}\sqrt{2\gamma \log(\delta^{-1})} \\
&\qquad = O\left(T^{-1/4}\sqrt{\log(T)} + T^{-1/2}\sqrt{\log(\delta^{-1})}\right).
\end{aligned}$$

The assumption we do on h is the same that the one that is done on g^* to obtain the convergence rate in Ciliberto et al. (2020), and is a common assumption for Kernel Ridge Regression (Caponnetto and De Vito, 2007; Steinwart and Christmann, 2008). Up to the log factor, we retrieve the same bound in the online learning setting and without assuming that the data are i.i.d. Therefore our result is more general than the original result in the batch statistical framework, however we are using T different functions to predict the outputs. In Theorem 10, we provide a similar result with a single estimator obtained by aggregation.

Proof. **Step 1: Controlling the regret of (f_t) by the regret of $(\hat{g}_t)$.** Applying Lemma 8, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_t[\Delta(f_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)] \\
&\leq 2c_\Delta \sqrt{T} \sqrt{\sum_{t=1}^{T} \mathbb{E}_t[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2]} \\
&= 2c_\Delta \sqrt{T} \sqrt{\sum_{t=1}^{T} \mathbb{E}_t[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2 - \|g^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2 + \|g^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2]} \,.
\end{aligned}$$

Now note that

$$\begin{aligned}
\mathbb{E}_t[\|\hat{g}_t(x_t) - \varphi_t\|^2] &= \mathbb{E}_t[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|^2 + \|\varphi_t - \mathbb{E}[\varphi_t|x_t]\|^2] \\
&\quad -2\mathbb{E}_t[\langle \mathbb{E}[\varphi_t|x_t] - \varphi_t, \mathbb{E}[\varphi_t|x_t] - \hat{g}_t(x_t)\rangle] \\
&= \mathbb{E}_t[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|^2 + \|\varphi_t - \mathbb{E}[\varphi_t|x_t]\|^2]
\end{aligned}$$

since we condition on x_t in $\mathbb{E}_t$. The same equality holds for $g^*(x_t)$. Thus

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_t[\|\hat{g}_t(x_t) - \mathbb{E}[\varphi_t|x_t]\|^2 - \|g^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|^2] \\
&\qquad = \sum_{t=1}^{T} \mathbb{E}_t[\|\hat{g}_t(x_t) - \varphi_t\|^2 - \|g^*(x_t) - \varphi_t\|^2] \,.
\end{aligned}$$

Step 2: Bounding the regret of the shifted *KAAR* estimator. We note that

$$\begin{aligned}
\hat{g}_t &= \arg\min_{g \in \mathcal{G}} \sum_{s=1}^{t-1} \|\tfrac{1}{2} g(x_s) + \tfrac{1}{2}\hat{g}_s(x_s) - \varphi_s\|_{\mathcal{H}}^2 + \lambda \|g\|_{\mathcal{G}}^2 + \tfrac{1}{4}\|g(x_t)\|_{\mathcal{H}}^2 \\
&= \arg\min_{g \in \mathcal{G}} \tfrac{1}{4} \sum_{s=1}^{t-1} \|g(x_s) + \hat{g}_s(x_s) - 2\varphi_s\|_{\mathcal{H}}^2 + \lambda \|g\|_{\mathcal{G}}^2 + \tfrac{1}{4}\|g(x_t)\|_{\mathcal{H}}^2 \\
&= \arg\min_{g \in \mathcal{G}} \sum_{s=1}^{t-1} \|g(x_s) + \hat{g}_s(x_s) - 2\varphi_s\|_{\mathcal{H}}^2 + 4\lambda \|g\|_{\mathcal{G}}^2 + \|g(x_t)\|_{\mathcal{H}}^2 \,.
\end{aligned}$$

Thus the function $\hat{g}_t$ aims to estimate $2\varphi_s - \hat{g}_s(x_s)$ with a regularisation parameter 4λ. We apply Lemma 6 and bound $2\varphi_s - \hat{g}_s(x_s)$ for all $s \in [\![T]\!]$,

$$\|2\varphi_s - \hat{g}_s(x_s)\|_{\mathcal{H}} \leq 2\|\varphi_s\| + \|\hat{g}_s(x_s)\| \leq 2 + \|\phi(x_s)\| \sup_{g \in \mathcal{G}} \|g\|_{\mathcal{G}} \leq 2 + \kappa \sup_{g \in \mathcal{G}} \|g\|_{\mathcal{G}} =: B \,.$$

We get

$$\sum_{t=1}^{T} \|2\hat{g}_t(x_t) - 2\varphi_t\|_{\mathcal{H}}^2 \leq B^2 \log\left(e + \tfrac{e\kappa^2 T}{4\lambda}\right) d_{\text{eff}}(4\lambda) + \min_{g \in \mathcal{G}} \left[\sum_{t=1}^{T} \|g(x_t) + \hat{g}_t(x_t) - 2\varphi_t\|_{\mathcal{H}}^2 + 4\lambda\|g\|_{\mathcal{G}}^2\right] .$$

We divide by 4 on both sides and obtain

$$\sum_{t=1}^{T} \|\hat{g}_t(x_t) - \varphi_t\|_{\mathcal{H}}^2 \leq \tfrac{B^2}{4} \log\left(e + \tfrac{e\kappa^2 T}{4\lambda}\right) d_{\text{eff}}(4\lambda) + \min_{g \in \mathcal{G}} \tilde{L}_T(g) \,. \tag{34}$$

Step 3: Applying van der Hoeven et al. (2023) to bound in high probability. We apply Theorem 1 of van der Hoeven et al. (2023) to obtain with probability $1 - \delta$,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_t[\Delta(f_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)] \\
&\leq 2c_\Delta \sqrt{T} \left(\tfrac{B^2}{2} d_{\text{eff}}(4\lambda) \log\left(e + \tfrac{e\kappa^2 T}{4\lambda}\right) + 2\gamma \log(\delta^{-1}) + 2\lambda\|g^*\|^2 + \sum_{t=1}^{T} \|g^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|^2\right)^{1/2}
\end{aligned} \tag{35}$$

where γ is defined in Eq. (33) and $B = 2 + \kappa \sup_{g \in \mathcal{G}} \|g\|_{\mathcal{G}}$. We bound the effective dimension by $d_{\text{eff}}(\lambda) \leq \frac{\kappa^2 T}{\lambda}$ and obtain

$$\sum_{t=1}^{T} \mathbb{E}_t[\Delta(f_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)] \leq 2c_\Delta \sqrt{T} \left(\frac{B^2 \kappa^2 T}{8\lambda} \log\left(e + \frac{e\kappa^2 T}{4\lambda}\right) + 2\gamma \log(\delta^{-1}) + 2\lambda \|g^*\|^2 \right)^{1/2} .$$

By choosing $\lambda = \sqrt{T}$ we get

$$\sum_{t=1}^{T} \mathbb{E}_t[\Delta(f_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)] \leq 2c_\Delta \sqrt{T} \left(\frac{B^2 \kappa^2 \sqrt{T}}{8} \log\left(e + \frac{e\kappa^2 \sqrt{T}}{4}\right) + 2\gamma \log(\delta^{-1}) + 2\sqrt{T} \|g^*\|^2 \right)^{1/2} .$$

□

C.1 Aggregating into a Unique Predictor

In this section, we build a unique predictor and we consider that the data $(x_t, y_t)_t$ are i.i.d., in order to be as close as possible from the supervised learning framework. As the output space $\mathcal{Z}$ does not have a vectorial structure, we cannot aggregate the $(f_t)_t$. Indeed, let $f, f' : \mathcal{X} \to \mathcal{Z}$, there is no guarantee that $f + f'$ takes values in $\mathcal{Z}$ as well. Therefore we build a unique predictor $\bar{f}_T$ from the T already computed feature predictors $(\hat{g}_t)_t$. We build an aggregate $\bar{g}_T : \mathcal{X} \to \mathcal{H}$ defined as the average of the $\hat{g}_t$,

$$\bar{g}_T = \frac{1}{T} \sum_{t=1}^{T} \hat{g}_t \tag{36}$$

and define the predictor $\bar{f}_T : \mathcal{X} \to \mathcal{Z}$ with respect to $\bar{g}_T$ as follows

$$\bar{f}_T(x) = \arg\min_{z \in \mathcal{Z}} \langle \psi(z), \bar{g}_T(x) \rangle_{\mathcal{H}} . \tag{37}$$

We bound the excess risk of the predictor $\bar{f}_T$.

Theorem 10 (Excess Risk)**.** *Let $f^* : \mathcal{X} \to \mathcal{Z}$ be a measurable function and $\gamma = 8(\kappa \sup \|g\|_{\mathcal{G}} + 1)^2$, with $\lambda = \sqrt{T}$ and assuming there is a function $g^* \in \mathcal{G}$ such that $\mathbb{E}_x \mathbb{E}_y[\|g^*(x) - \mathbb{E}[\varphi(y)|x]\|^2 | x] = 0$. Let $\bar{f}_T$ be defined as in* (37) *and let $\delta \in (0, 1]$. With probability $1 - \delta$, we have*

$$\begin{aligned}
&\mathbb{E}_{x,y}[\Delta(\bar{f}_T(x), y) - \Delta(f^*(x), y)] \\
&\quad \leq 2c_\Delta T^{-1/4} \sqrt{\frac{B^2 \kappa^2}{8} \log\left(e + \frac{e\kappa^2 \sqrt{T}}{4}\right) + 2\|g^*\|^2} + 2c_\Delta T^{-1/2} \sqrt{2\gamma \log(\delta^{-1})} \\
&\quad = O\left(T^{-1/4} \sqrt{\log(T)} + T^{-1/2} \sqrt{\log(\delta^{-1})} \right) .
\end{aligned}$$

Proof. We follow the proof of Lemma 8, with the difference that at the end we apply Jensen's inequality with respect to the expectation to obtain a comparison inequality

$$\begin{aligned}
&\mathbb{E}_{x,y}[\Delta(\bar{f}_T(x), y) - \Delta(f^*(x), y)] \\
&\quad \leq \mathbb{E}_{x,y}[\langle \psi(\bar{f}_T(x)) - \psi(f^*(x)), \bar{g}_T(x) - \varphi(y) \rangle] \\
&\quad = \mathbb{E}_x \mathbb{E}_y[\langle \psi(\bar{f}_T(x)) - \psi(f^*(x)), \bar{g}_T(x) - \mathbb{E}[\varphi(y)|x] \rangle | x] \\
&\quad + \mathbb{E}_x \mathbb{E}_y[\langle \psi(\bar{f}_T(x)) - \psi(f^*(x)), \mathbb{E}[\varphi(y)|x] - \varphi(y) \rangle | x] \\
&\quad = \mathbb{E}_x \mathbb{E}_y[\langle \psi(\bar{f}_T(x)) - \psi(f^*(x)), \bar{g}_T(x) - \mathbb{E}[\varphi(y)|x] \rangle | x] \\
&\quad \leq 2c_\Delta \mathbb{E}_x \mathbb{E}_y[\|\bar{g}_T(x) - \mathbb{E}[\varphi(y)|x]\| | x] \\
&\quad \leq 2c_\Delta \sqrt{\mathbb{E}_x \mathbb{E}_y[\|\bar{g}_T(x) - \mathbb{E}[\varphi(y)|x]\|^2 | x]} .
\end{aligned}$$

We now apply Theorem 1 from van der Hoeven et al. (2023) using Step 2 of the proof of Theorem 9.

$$
\begin{aligned}
&\mathbb{E}_{x,y}[\Delta(\bar{f}_T(x), y) - \Delta(f^*(x), y)] \\
&\quad = 2c_\Delta \left(\mathbb{E}_x \mathbb{E}_y[\|\bar{g}_T(x) - \mathbb{E}[\varphi(y)|x]\|^2 - \|g^*(x) - \mathbb{E}[\varphi(y)|x]\|^2 + \|g^*(x) - \mathbb{E}[\varphi(y)|x]\|^2 |x]\right)^{1/2} \\
&\quad = 2c_\Delta \left(\mathbb{E}_x \mathbb{E}_y[\|\bar{g}_T(x) - \varphi(y)\|^2 - \|g^*(x) - \varphi(y)\|^2 + \|g^*(x) - \mathbb{E}[\varphi(y)|x]\|^2 |x]\right)^{1/2} \\
&\quad = 2c_\Delta \left(\mathbb{E}_{x,y}[\|\bar{g}_T(x) - \varphi(y)\|^2 - \|g^*(x) - \varphi(y)\|^2] + \mathbb{E}_x \mathbb{E}_y[\|g^*(x) - \mathbb{E}[\varphi(y)|x]\|^2 |x]\right)^{1/2} \\
&\quad \leq 2c_\Delta \left(\frac{\frac{B^2}{2} d_{\text{eff}}(4\lambda) \log\left(e + \frac{e\kappa^2 T}{4\lambda}\right) + 2\lambda \|g^*\|_{\mathcal{G}}^2 + 2\gamma \log(\delta^{-1})}{T} + \mathbb{E}_x \mathbb{E}_y[\|g^*(x) - \mathbb{E}[\varphi(y)|x]\|^2 |x] \right)^{1/2} \\
&\quad = 2c_\Delta \left(\frac{\frac{B^2}{2} d_{\text{eff}}(4\lambda) \log\left(e + \frac{e\kappa^2 T}{4\lambda}\right) + 2\lambda \|g^*\|_{\mathcal{G}}^2 + 2\gamma \log(\delta^{-1})}{T} \right)^{1/2}
\end{aligned}
$$

where $B = 2 + \kappa \sup_{g \in \mathcal{G}} \|g\|_{\mathcal{G}}$. Upper-bounding the effective dimension by $d_{\text{eff}}(4\lambda) \leq \frac{\kappa^2 T}{4\lambda}$ and choosing $\lambda = \sqrt{T}$ gives the desired result. □

D Proof of Theorem 5: Dealing with Non-Stationary Data in Expectation

We define $m \in \mathbb{N}$ such that $1 = t_1 \leq t_2 \leq \cdots \leq t_{m+1} = T + 1$ and such that

$$
\sum_{t=t_i+1}^{t_{i+1}-1} \|g_t^* - g_{t-1}^*\|_{\mathcal{G}} \leq \frac{V_{\mathcal{G}}}{m} \quad \text{for all } i \in [\![1, m]\!]. \tag{38}
$$

That is to say that the variation of $(g_t^*)_t$ is small between t_i and $t_{i+1} - 1$. Note that the sum $\sum_{i=1}^m \sum_{t=t_i+1}^{t_{i+1}-1} \|g_t^* - g_{t-1}^*\|_{\mathcal{G}}$ does not take into account the norms of $\|g_{t_i}^* - g_{t_i-1}^*\|_{\mathcal{G}}$ for $i \in [\![2, m]\!]$. As in Raj et al. (2020), we define an approximation $(g_{t_i:t_{i+1}}^*)_{i=1}^m \in \mathcal{G}^m$ of $(g_t^*)_t$ with only m changes through the T time steps that occur between t_i and t_{i+1}. It is an hypothetical forecaster with m restart times. Formally we define

$$
\bar{g}_{t_i:t_{i+1}} := \arg\min_{g \in \mathcal{G}} \sum_{t=t_i}^{t_{i+1}-1} \|g - g_t^*\|_{\mathcal{G}}^2 = \frac{1}{t_{i+1} - t_i} \sum_{t=t_i}^{t_{i+1}-1} g_t^* \tag{39}
$$

and by $\bar{g}_t$ we denote $\bar{g}_{t_i:t_{i+1}}$ for all $t \in [\![t_i, t_{i+1} - 1]\!]$.

We bound the dynamic regret of the *KAAR* estimator, see Proposition 11. It is expressed with respect to the time dependent effective dimension $d_{\text{eff}}(\lambda, s - r)$ defined as

$$
d_{\text{eff}}(\lambda, s - r) := Tr(K_{s-r,s-r}(K_{s-r,s-r} + \lambda I)^{-1}) \quad \forall \lambda > 0, \tag{40}
$$

where $K_{s-r,s-r} \in \mathbb{R}^{(s-r-1)\times(s-r-1)}$ is defined by $(K_{s-r,s-r})_{ij} = k(x_{r+i-1}, x_{r+j-1})$.

Proposition 11 (Dynamic Regret of *KAAR*)**.** *Let $(\hat{g}_t)_t$ be defined as in* (22) *and let $(g_t^*)_t \in \mathcal{G}^T$. Let $m \in \mathbb{N}$ be defined as in* (38)*. Let $\eta = 1/2(\kappa \sup\|g\|_{\mathcal{G}} + 1)^2$. Then we have*

$$
\begin{aligned}
&\sum_{t=1}^{T} \|\hat{g}_t(x_t) - \varphi_t\|_{\mathcal{H}}^2 - \|g_t^*(x_t) - \varphi_t\|_{\mathcal{H}}^2 \\
&\qquad \leq \frac{m \log T}{\eta} + \log\left(e + \frac{e\kappa^2 T}{\lambda}\right) \sum_{i=1}^{m} d_{\mathit{eff}}(\lambda, t_{i+1} - t_i) + \lambda m \max_{t \in [\![T]\!]} \|g_t^*\|_{\mathcal{G}}^2 + \frac{4\kappa V_{\mathcal{G}} T}{m} =: R_T(\lambda, m)\,.
\end{aligned}
$$

Proof. **Step 1: We add two intermediary terms.** We introduce two new terms in the sum and bound the differences separately.

$$\begin{aligned}\sum_{t=1}^{T}\|\hat{g}_t(x_t)-\varphi_t\|_{\mathcal{H}}^2-\|g_t^*(x_t)-\varphi_t\|_{\mathcal{H}}^2 &= \sum_{i=1}^{m}\sum_{t=t_i}^{t_{i+1}-1}\|\hat{g}_t(x_t)-\varphi_t\|_{\mathcal{H}}^2-\|g_t^*(x_t)-\varphi_t\|_{\mathcal{H}}^2\\ &= \sum_{i=1}^{m}\sum_{t=t_i}^{t_{i+1}-1}\|\hat{g}_t(x_t)-\varphi_t\|_{\mathcal{H}}^2-\|\hat{g}_{t_i:t}(x_t)-\varphi_t\|_{\mathcal{H}}^2 \\ &\qquad +\|\hat{g}_{t_i:t}(x_t)-\varphi_t\|_{\mathcal{H}}^2-\|\bar{g}_{t_i:t_{i+1}}(x_t)-\varphi_t\|_{\mathcal{H}}^2 \\ &\qquad +\|\bar{g}_{t_i:t_{i+1}}(x_t)-\varphi_t\|_{\mathcal{H}}^2-\|g_t^*(x_t)-\varphi_t\|_{\mathcal{H}}^2\end{aligned} \tag{41}$$

Step 2: Bounding the first difference. For all $k^*\in[\![T]\!]$ and all $s_2\in[\![T]\!]$, we prove that

$$\sum_{t=1}^{s_2}\left(\ell_t\left(\sum_{k=1}^{K}p_t(k)\hat{g}_{k:t}\right)-\ell_t(\hat{g}_{k^*:t})\right)1[k^*\le t]\le\frac{\log K}{\eta}\,. \tag{42}$$

The proof is based on the proof of EWA applied to $\tilde{\ell}_t$. Let W_t be the normalisation constant of w_t.

$$\begin{aligned}W_{s_2+1} &= \sum_{k=1}^{K}\exp\left(-\eta\sum_{s=1}^{s_2}\tilde{\ell}_s(k)\right)\\ &= \sum_{k=1}^{K}\exp\left(-\eta\sum_{s=1}^{s_2-1}\tilde{\ell}_s(k)\right)\exp(-\eta\tilde{\ell}_{s_2}(k))\\ &= W_{s_2}\sum_{k=1}^{K}w_{s_2}(k)\exp(-\eta\tilde{\ell}_{s_2}(k))\\ &\le W_{s_2}\exp\left(-\eta\ell_{s_2}\left(\sum_{k=1}^{K}w_{s_2}(k)\tilde{g}_{k:s_2}\right)\right)\end{aligned}$$

where the inequality comes from Jensen's inequality and η-exp-concavity of ℓ_{s_2}, and where we define

$$\tilde{g}_{k:t}=\begin{cases}\hat{g}_{k:t}\text{ if }k\le t\\ \hat{g}_t\text{ if }k>t\end{cases}\,.$$

Now note that

$$\begin{aligned}\sum_{k=1}^{K}w_{s_2}(k)\tilde{g}_{k:s_2} &= \sum_{k\le s_2}w_{s_2}(k)\hat{g}_{k:s_2}+\sum_{k>s_2}w_{s_2}(k)\hat{g}_{s_2}\\ &= \left(\sum_{k\le s_2}w_{s_2}(k)\right)\sum_{k\le s_2}p_{s_2}(k)\hat{g}_{k:s_2}+\sum_{k>s_2}w_{s_2}(k)\hat{g}_{s_2}\\ &= \hat{g}_{s_2}\left(\sum_{k\le s_2}w_{s_2}(k)+\sum_{k>s_2}w_{s_2}(k)\right)\\ &= \hat{g}_{s_2}.\end{aligned}$$

Thus by induction we obtain

$$W_{s_2+1}\le W_1\exp\left(-\eta\sum_{s=1}^{s_2}\ell_s(\hat{g}_s)\right)\le K\exp\left(-\eta\sum_{s=1}^{s_2}\ell_s(\hat{g}_s)\right)$$

where the right inequality is obtained by choosing to initialise w_1 as the uniform probability over the K experts. We now compute a lower bound of W_{s_2+1},

$$\begin{aligned}
W_{s_2+1} &= \sum_{k=1}^{K} \exp\left(-\eta \sum_{s=1}^{s_2} \tilde{\ell}_s(k)\right) \\
&\geq \exp\left(-\eta \sum_{s=1}^{s_2} \tilde{\ell}_s(k^*)\right) \\
&= \exp\left(-\eta \sum_{s=1}^{s_2} \ell_s(\hat{g}_{k^*:s}) 1[k^* \leq s] + \ell_s(\hat{g}_s) 1[k^* > s]\right) .
\end{aligned}$$

By taking the log we obtain the desired result.

Step 3: Bounding the second difference. We have that

$$\begin{aligned}
&\sum_{t=t_i}^{t_{i+1}-1} \|\hat{g}_{t_i:t}(x_t) - \varphi_t\|_{\mathcal{H}}^2 - \|\bar{g}_{t_i:t_{i+1}}(x_t) - \varphi_t\|_{\mathcal{H}}^2 \\
&\qquad \leq d_{\text{eff}}(\lambda, t_{i+1} - t_i) \log\left(e + \tfrac{e\kappa^2(t_{i+1}-t_i)}{\lambda}\right) + \lambda \|\bar{g}_{t_i:t_{i+1}}\|_{\mathcal{G}}^2 .
\end{aligned}$$

This is a generalisation of Theorem 6 with a late starting point.

Step 4: Bounding the third difference. We bound the third term of the sum using the following derivation.

$$\begin{aligned}
&\|\bar{g}_{t_i:t_{i+1}}(x_t) - \varphi_t\|_{\mathcal{H}}^2 - \|g_t^*(x_t) - \varphi_t\|_{\mathcal{H}}^2 \\
&= -\left\|g_t^*(x_t) - \bar{g}_{t_i:t_{i+1}}(x_t)\right\|_{\mathcal{H}}^2 + 2\left\langle \varphi_t - \bar{g}_{t_i:t_{i+1}}(x_t), g_t^*(x_t) - \bar{g}_{t_i:t_{i+1}}(x_t)\right\rangle_{\mathcal{H}} \\
&\leq 2\left\|\varphi_t - \bar{g}_{t_i:t_{i+1}}(x_t)\right\|_{\mathcal{H}} \left\|g_t^*(x_t) - \bar{g}_{t_i:t_{i+1}}(x_t)\right\|_{\mathcal{H}} \\
&\leq 4\kappa \left\|g_t^* - \bar{g}_{t_i:t_{i+1}}\right\|_{\mathcal{G}}
\end{aligned}$$

We bound the norm by

$$\begin{aligned}
\|g_t^* - \bar{g}_{t_i:t_{i+1}}\|_{\mathcal{G}} &= \left\| g_t^* - \tfrac{1}{t_{i+1}-t_i} \sum_{s=t_i}^{t_{i+1}-1} \bar{g}_s \right\|_{\mathcal{G}} \\
&= \left\| \tfrac{1}{t_{i+1}-t_i} \sum_{s=t_i}^{t_{i+1}-1} g_t^* - \bar{g}_s \right\|_{\mathcal{G}} \\
&\leq \tfrac{1}{t_{i+1}-t_i} \sum_{s=t_i}^{t_{i+1}-1} \|g_t^* - \bar{g}_s\|_{\mathcal{G}} \\
&\leq \max_{s \in [\![t_i, t_{i+1}-1]\!]} \|g_t^* - \bar{g}_s\|_{\mathcal{G}}
\end{aligned}$$

where the first inequality comes from Jensen's inequality. We now separate the max in two terms at time step t, and use a telescopic sum.

$$\begin{aligned}
\max_{s \in [\![t_i, t_{i+1}-1]\!]} \|g_t^* - \bar{g}_s\|_{\mathcal{G}} &\leq \max_{s \in [\![t_i, t-1]\!]} \|g_t^* - \bar{g}_s\|_{\mathcal{G}} + \max_{s \in [\![t+1, t_{i+1}-1]\!]} \|g_t^* - \bar{g}_s\|_{\mathcal{G}} \\
&= \max_{s \in [\![t_i, t-1]\!]} \left\| \sum_{r=s+1}^{t} g_r^* - \bar{g}_{r-1} \right\| + \max_{s \in [\![t+1, t_{i+1}-1]\!]} \left\| \sum_{r=t+1}^{s} g_r^* - \bar{g}_{r-1} \right\| \\
&\leq \max_{s \in [\![t_i, t-1]\!]} \sum_{r=s+1}^{t} \|g_r^* - \bar{g}_{r-1}\| + \max_{s \in [\![t+1, t_{i+1}-1]\!]} \sum_{r=t+1}^{s} \|g_r^* - \bar{g}_{r-1}\| \\
&= \sum_{r=t_i+1}^{t_{i+1}-1} \|g_r^* - \bar{g}_{r-1}\| \\
&\leq V_{\mathcal{G}}/m
\end{aligned}$$

We may now sum to obtain

$$\begin{aligned}4\kappa\sum_{t=1}^{T}\|g_t^*-\bar{g}_t\|_{\mathcal{G}} &= 4\kappa\sum_{i=1}^{m}\sum_{t=t_i}^{t_{i+1}-1}\|g_t^*-\bar{g}_{t_i:t_{i+1}}\|_{\mathcal{G}}\\ &\leq 4\kappa\sum_{i=1}^{m}\sum_{t=t_i}^{t_{i+1}-1}V_{\mathcal{G}}/m\\ &= 4\kappa V_{\mathcal{G}}/m\sum_{i=1}^{m}t_{i+1}-t_i\\ &= 4\kappa V_{\mathcal{G}}T/m.\end{aligned}$$

Step 5: Putting everything together. We obtain

$$\begin{aligned}&\sum_{t=1}^{T}\|\hat{g}_t(x_t)-\varphi_t\|_{\mathcal{H}}^2-\|g_t^*(x_t)-\varphi_t\|_{\mathcal{H}}^2\\ &\leq \tfrac{m\log T}{\eta}+\log\Big(e+\tfrac{e\kappa^2T}{\lambda}\Big)\sum_{i=1}^{m}d_{\text{eff}}(\lambda,t_{i+1}-t_i)+\lambda\sum_{i=1}^{m}\|\bar{g}_{t_i:t_{i+1}}\|^2+\tfrac{4\kappa V_{\mathcal{G}}T}{m}.\end{aligned}$$

We conclude by noting that

$$\big\|\bar{g}_{t_i:t_{i+1}}\big\|_{\mathcal{G}}^2=\left\|\tfrac{1}{t_{i+1}-t_i}\sum_{t=t_i}^{t_{i+1}-1}g_t^*\right\|_{\mathcal{G}}^2\leq\tfrac{1}{t_{i+1}-t_i}\sum_{t=t_i}^{t_{i+1}-1}\|g_t^*\|_{\mathcal{G}}^2\leq\max_{t\in[\![t_i,t_{i+1}-1]\!]}\|g_t^*\|_{\mathcal{G}}^2$$

where the first inequality is by Jensen's inequality. □

We recall and prove our main result from Section 5.

Theorem 5 (Expected Regret in a Non-Stationary Environment)**.** *Assume that there exists (g_t^*) a sequence in $\mathcal{G}$ such that $\mathbb{E}\big[\sum_{t=1}^{T}\|g_t^*(x_t)-\mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2\big]=0$. Then, Algorithm 2 run with $\lambda>0$ and $\eta=1/2(\kappa\sup\|g\|_{\mathcal{G}}+1)^2$ satisfies*

$$\mathbb{E}[R_T]=\begin{cases}\tilde{O}(V_{\mathcal{G}}^{1/6}T^{5/6}) & \text{if } \lambda=V_{\mathcal{G}}^{-1/3}T^{1/3}\\ \tilde{O}(V_0^{1/4}T^{3/4}) & \text{if } \lambda=V_0^{-1/2}T^{1/2}\end{cases}. \tag{19}$$

Precisely if $\lambda=V_{\mathcal{G}}^{-1/3}T^{1/3}$, we have

$$\mathbb{E}[R_T]\leq 2c_\Delta\sqrt{T}\left(\begin{array}{l}\frac{\left\lceil V_{\mathcal{G}}^{2/3}T^{1/3}\kappa^{-2/3}\right\rceil\log T}{\eta}+4\kappa^{5/3}V_{\mathcal{G}}^{1/3}T^{2/3}\\ +(V_{\mathcal{G}}^{2/3}T^{4/3}\kappa^{-2/3}+T)^{1/2}\Big(\kappa^2\log\Big(e+e\kappa^2(V_{\mathcal{G}}^{2/3}T^{4/3}\kappa^{-2/3}+T)^{1/2}\Big)+\max\limits_{t\in[\![T]\!]}\|g_t^*\|_{\mathcal{G}}^2\Big)\end{array}\right)^{1/2},$$

and if $\lambda=V_0^{-1/2}T^{1/2}$, we have

$$\mathbb{E}[R_T]\leq 2c_\Delta\sqrt{T}\sqrt{\frac{V_0\log T}{\eta}+\log\Big(e+e\kappa^2\sqrt{TV_0}\Big)\kappa^2\sqrt{TV_0}+\sqrt{TV_0}\max_{i\in[\![V_0]\!]}\|\bar{g}_{t_i:t_{i+1}}\|_{\mathcal{G}}^2}.$$

Proof. **Step 1: Controlling the regret of $(f_t)_t$ by the regret of $(\hat{g}_t)_t$.** From Lemma 7, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\Delta(f_t(x_t),y_t)-\Delta(f_t^*(x_t),y_t)\right]\leq 2c_\Delta\sqrt{T}\sqrt{\sum_{t=1}^{T}\mathbb{E}[\|\hat{g}_t(x_t)-\mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2]}.$$

We add and subtract a term and then follow the proof of Theorem 3,

$$\mathbb{E}\left[\sum_{t=1}^{T}\Delta(f_t(x_t),y_t)-\Delta(f_t^*(x_t),y_t)\right]$$
$$\leq 2c_\Delta\sqrt{T}\sqrt{\sum_{t=1}^{T}\mathbb{E}[\|\hat{g}_t(x_t)-\mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2-\|g_t^*(x_t)-\mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2+\|g_t^*(x_t)-\mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2]}\,.$$

We now apply Proposition 11 to obtain

$$\mathbb{E}[R_T]\leq 2c_\Delta\sqrt{T}\sqrt{R_T(\lambda,m)+\mathbb{E}\left[\sum_{t=1}^{T}\|g_t^*(x_t)-\mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2\right]} \tag{43}$$

where $R_T(\lambda,m)$ is as in Proposition 11 and is defined as

$$R_T(\lambda,m)=\tfrac{m\log T}{\eta}+\log\left(e+\tfrac{e\kappa^2T}{\lambda}\right)\sum_{i=1}^{m}d_{\text{eff}}(\lambda,t_{i+1}-t_i)+\lambda m\max_{t\in[\![T]\!]}\|g_t^*\|_{\mathcal{G}}^2+\tfrac{4\kappa V_{\mathcal{G}}T}{m}.$$

Case 1: Continuous variations. We bound the effective dimension by $d_{\text{eff}}(\lambda,t_{i+1}-t_i)\leq\frac{\kappa^2(t_{i+1}-t_i)}{\lambda}$. Thus we can bound the sum by $\sum_{i=1}^{m}d_{\text{eff}}(\lambda,t_{i+1}-t_i)\leq\kappa^2T/\lambda$ and obtain

$$R_T(\lambda,m)\leq\tfrac{m\log T}{\eta}+\log\left(e+\tfrac{e\kappa^2T}{\lambda}\right)\tfrac{\kappa^2T}{\lambda}+\lambda m\max_{t\in[\![T]\!]}\|g_t^*\|_{\mathcal{G}}^2+\tfrac{4\kappa V_{\mathcal{G}}T}{m}.$$

We choose $\lambda=\sqrt{T/m}$ and get

$$R_T(\lambda,m)\leq\tfrac{m\log T}{\eta}+\log(e+e\kappa^2\sqrt{Tm})\kappa^2\sqrt{Tm}+\sqrt{Tm}\max_{t\in[\![T]\!]}\|g_t^*\|_{\mathcal{G}}^2+\tfrac{4\kappa V_{\mathcal{G}}T}{m}.$$

We choose $m=\left\lceil V_{\mathcal{G}}^{2/3}T^{1/3}\kappa^{-2/3}\right\rceil$ and get

$$R_T(\lambda,m)\leq\frac{\left\lceil V_{\mathcal{G}}^{2/3}T^{1/3}\kappa^{-2/3}\right\rceil\log T}{\eta}+4\kappa^{5/3}V_{\mathcal{G}}^{1/3}T^{2/3}$$
$$+(V_{\mathcal{G}}^{2/3}T^{4/3}\kappa^{-2/3}+T)^{1/2}\left(\kappa^2\log(e+e\kappa^2(V_{\mathcal{G}}^{2/3}T^{4/3}\kappa^{-2/3}+T)^{1/2})+\max_{t\in[\![T]\!]}\|g_t^*\|_{\mathcal{G}}^2\right).$$

We conclude by noting that $V_{\mathcal{G}}=\|g_1^*\|_{\mathcal{G}}+\sum_{t=2}^{T}\|g_t^*-g_{t-1}^*\|_{\mathcal{G}}\leq(2T-1)\sup\|g\|_{\mathcal{G}}$.

Case 2: Discrete variations. In the case of discrete distributions data variations there is no more need to approximate the data distributions $(g_t^*)_t$ by the hypothetical forecasters $(\bar{g}_{t_i:t_{i+1}})_i$. Thus the third term of the sum in Eq. (41) is not necessary. And we can replace $R_T(\lambda,m)$ by $R_T^0(\lambda,V_0)$ defined as

$$R_T^0(\lambda,V_0):=\frac{V_0\log T}{\eta}+\log\left(e+\tfrac{e\kappa^2T}{\lambda}\right)\sum_{i=1}^{V_0}d_{\text{eff}}(\lambda,t_{i+1}-t_i)+\lambda\sum_{i=1}^{V_0}\|\bar{g}_{t_i:t_{i+1}}\|_{\mathcal{G}}^2.$$

We bound the effective dimension by $d_{\text{eff}}(\lambda,t_{i+1}-t_i)\leq\frac{\kappa^2(t_{i+1}-t_i)}{\lambda}$. Thus we obtain $\sum_{i=1}^{V_0}d_{\text{eff}}(\lambda,t_{i+1}-t_i)\leq\kappa^2T/\lambda$ and

$$R_T^0(\lambda,V_0)\leq\frac{V_0\log T}{\eta}+\log\left(e+\tfrac{e\kappa^2T}{\lambda}\right)\tfrac{\kappa^2T}{\lambda}+\lambda V_0\max_{i\in[\![V_0]\!]}\|\bar{g}_{t_i:t_{i+1}}\|_{\mathcal{G}}^2. \tag{44}$$

By choosing $\lambda=\sqrt{T/V_0}$, we get

$$R_T^0(\lambda,V_0)\leq\frac{V_0\log T}{\eta}+\log\left(e+e\kappa^2\sqrt{TV_0}\right)\kappa^2\sqrt{TV_0}+\sqrt{TV_0}\max_{i\in[\![V_0]\!]}\|\bar{g}_{t_i:t_{i+1}}\|_{\mathcal{G}}^2. \tag{45}$$

Finally we bound the expected regret by

$$\mathbb{E}[R_T]\leq 2c_\Delta\sqrt{T}\sqrt{\frac{V_0\log T}{\eta}+\log\left(e+e\kappa^2\sqrt{TV_0}\right)\kappa^2\sqrt{TV_0}+\sqrt{TV_0}\max_{i\in[\![V_0]\!]}\|\bar{g}_{t_i:t_{i+1}}\|_{\mathcal{G}}^2}\,.$$

We now note that $V_0\leq T$ to conclude the proof. □

D.1 Refined Regret Bounds

Capacity Condition There exists $\beta \in [0,1]$ and $Q > 0$ for which

$$d_{\text{eff}}(\lambda) \leq Q \left(\frac{T}{\lambda}\right)^{\beta} \quad , \forall \lambda > 0. \tag{46}$$

When the kernel is bounded the condition above is always satisfied for $\beta = 1$. Indeed we can always bound the effective dimension by $d_{\text{eff}}(\lambda) \leq \frac{\kappa^2 T}{\lambda}$. Moreover if the eigenvalues of the covariance operator C decay polynomially $\sigma_i(C) \leq cj^{-\mu}$, for $c > 0, \mu > 1$ and $j \in \mathbb{N}$, then the capacity condition is satisfied with $Q = c$ and $\beta = -1/\mu$. Using this bound we may derive a refined bound of the expectation of the dynamic regret.

Assuming that the capacity condition holds, and that there exists (g_t^*) a sequence in $\mathcal{G}$ such that $\mathbb{E}\left[\sum_{t=1}^T \|g_t^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2\right] = 0$. Then choosing $\lambda = T^{\frac{\beta}{\beta+1}} m^{\frac{-\beta}{\beta+1}}$ and $m = \left\lceil V_{\mathcal{G}}^{\frac{\beta+1}{\beta+2}} T^{\frac{1}{\beta+2}} \kappa^{\frac{-(\beta+1)}{\beta+2}} \right\rceil$ leads to the following bound on the expected regret

$$\mathbb{E}[R_T] = \tilde{O}\left(V_{\mathcal{G}}^{\frac{1}{2(\beta+2)}} T^{\frac{2\beta+3}{2(\beta+2)}}\right). \tag{47}$$

Indeed using the capacity condition, we bound the effective dimension by $d_{\text{eff}}(\lambda, t_{i+1} - t_i) \leq Q\left(\frac{t_{i+1}-t_i}{\lambda}\right)^{\beta}$. Using Jensen's inequality, we then derive

$$\begin{aligned}
\sum_{i=1}^m d_{\text{eff}}(\lambda, t_{i+1} - t_i) &\leq Q \sum_{i=1}^m \left(\frac{t_{i+1}-t_i}{\lambda}\right)^{\beta} \\
&= Qm \sum_{i=1}^m \frac{1}{m} \left(\frac{t_{i+1}-t_i}{\lambda}\right)^{\beta} \\
&\leq Qm \left(\sum_{i=1}^m \frac{t_{i+1}-t_i}{\lambda m}\right)^{\beta} \\
&= Qm^{1-\beta} \left(\frac{T}{\lambda}\right)^{\beta}.
\end{aligned}$$

We obtain

$$R_T(\lambda, m) \leq \frac{m \log T}{\eta} + \log\left(e + \frac{e\kappa^2 T}{\lambda}\right) Q m^{1-\beta} \left(\frac{T}{\lambda}\right)^{\beta} + \lambda m \max_{t \in [\![T]\!]} \|g_t^*\|_{\mathcal{G}}^2 + \frac{4\kappa V_{\mathcal{G}} T}{m}.$$

We choose $\lambda = T^{\frac{\beta}{\beta+1}} m^{\frac{-\beta}{\beta+1}}$, and obtain

$$R_T(\lambda, m) \leq \frac{m \log T}{\eta} + m^{\frac{1}{\beta+1}} T^{\frac{\beta}{\beta+1}} \left(Q \log\left(e + e\kappa^2 m^{\frac{\beta}{\beta+1}} T^{\frac{1}{\beta+1}}\right) + \max_{t \in [\![T]\!]} \|g_t^*\|_{\mathcal{G}}^2\right) + \frac{4\kappa V_{\mathcal{G}} T}{m}.$$

We choose $m = \left\lceil V_{\mathcal{G}}^{\frac{\beta+1}{\beta+2}} T^{\frac{1}{\beta+2}} \kappa^{\frac{-(\beta+1)}{\beta+2}} \right\rceil$ and obtain

$$\begin{aligned}
R_T(\lambda, m) \leq{}& \frac{\left\lceil V_{\mathcal{G}}^{\frac{\beta+1}{\beta+2}} T^{\frac{1}{\beta+2}} \kappa^{\frac{-(\beta+1)}{\beta+2}} \right\rceil \log T}{\eta} + 4\kappa^{\frac{2\beta+3}{\beta+2}} V_{\mathcal{G}}^{\frac{1}{\beta+2}} T^{\frac{\beta+1}{\beta+2}} \\
&+ \left(V_{\mathcal{G}}^{\frac{1}{\beta+2}} T^{\frac{\beta+1}{\beta+2}} \kappa^{\frac{-1}{\beta+2}} + T^{\frac{\beta}{\beta+1}}\right) \left(Q \log\left(e + e\kappa^{\frac{2\beta+3}{\beta+2}} V_{\mathcal{G}}^{\frac{\beta}{\beta+2}} T^{\frac{2}{\beta+2}} + eT^{\frac{1}{\beta+1}}\right) + \max_{t \in [\![T]\!]} \|g_t^*\|_{\mathcal{G}}^2\right).
\end{aligned}$$

We note that the variation $V_{\mathcal{G}} = \|g_1^*\|_{\mathcal{G}} + \sum_{t=2}^T \|g_t^* - g_{t-1}^*\|_{\mathcal{G}} = O(T)$ to conclude.

Gaussian kernel Moreover, in the case of the Gaussian kernel we can bound the effective dimension by $d_{\text{eff}}(\lambda) \leq \left(\log\left(\frac{T}{\lambda}\right)\right)^d$ (Altschuler et al., 2019) where d is the dimension of the input space $\mathcal{X}$, to obtain a smaller regret. By choosing $m = \sqrt{V_{\mathcal{G}} T/(\lambda+1)}$, we obtain

$$\mathbb{E}[R_T] = \tilde{O}\left(T^{3/4} V_{\mathcal{G}}^{1/4} (\lambda+1)^{1/4}\right). \tag{48}$$

In this particular case, if we choose a constant λ, we retrieve the power $T^{3/4}$ from the stationary case.

E Dealing with Non-Stationary Data in High Probability

In this section we deal with non-stationary data distributions as in Section 5, however we bound the cumulative risk in high probability instead of bounding the expected regret. As in Appendix C, we define the feature predictor on a shifted version of the losses in order to apply Theorem 1 of van der Hoeven et al. (2023), see Algorithm 3. We define the loss ℓ_t as

$$\ell_t(g) = \left\|\varphi_s - \tfrac{1}{2}\hat{g}_t(x_t) - \tfrac{1}{2}g(x_t)\right\|_{\mathcal{H}}^2. \tag{49}$$

We recall the definition of the filter $\mathcal{F}_{t-1} = (x_1, y_1, \ldots, x_{t-1}, y_{t-1}, x_t)$ and the notation $\mathbb{E}_t[.]$ that stands for $\mathbb{E}_{y_t}[.|\mathcal{F}_{t-1}]$.

Algorithm 3: *SALAMI – Structured prediction ALgorithm with Aggregating MIxture* – for the high probability setting

Input: $\lambda > 0$, exp-concavity constant η of $(\ell_t)_t$, kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
for *Each time step* t *in* $1 \ldots T$ **do**
  Get information $x_t \in \mathcal{X}$
  for *Each expert* s *in* $1 \ldots t$ **do**
    Compute $\hat{g}_{s:t} = \arg\min_{g\in\mathcal{G}} \sum_{s=s_1}^{t-1} \left\|\varphi_s - \frac{1}{2}\hat{g}_s(x_s) - \frac{1}{2}g(x_s)\right\|_{\mathcal{H}}^2 + \lambda\|g\|_{\mathcal{G}}^2 + \frac{1}{4}\|g(x_t)\|_{\mathcal{H}}^2$
  end
  for *Each expert* s *in* $1 \ldots T$ **do**
    Compute the auxiliary loss $\tilde{\ell}_t(s) = \begin{cases}\ell_t(\hat{g}_{s:t}) \text{ if } s \le t \\ \ell_t(\hat{g}_t) \text{ if } s > t\end{cases}$
    Compute the probability using EWA $w_t(s) \propto w_{t-1}(s)\exp(-\eta\tilde{\ell}_t(s))$
    Compute the probability $p_t(s) \propto \begin{cases}w_t(s) \text{ if } s \le t \\ 0 \text{ if } s > t\end{cases}$
  end
  Compute the aggregate predictor $\hat{g}_t = \sum_{s=1}^{T} p_t(s)\hat{g}_{s:t}$
  Compute the prediction $\hat{z}_t = \hat{f}_t(x_t) = \arg\min_{z\in\mathcal{Z}}\langle\psi(z), \hat{g}_t(x_t)\rangle_{\mathcal{H}}$
  Observe ground truth $y_t \in \mathcal{Y}$
  Get loss $\Delta(\hat{z}_t, y_t) \in \mathbb{R}$
end

The only difference with Appendix D is the definition of the experts. We shift the losses by $\frac{1}{2}\hat{g}_s(x_s)$,

$$\hat{g}_{s_1:t} := \arg\min_{g\in\mathcal{G}} \sum_{s=s_1}^{t-1} \left\|\varphi_s - \tfrac{1}{2}\hat{g}_s(x_s) - \tfrac{1}{2}g(x_s)\right\|_{\mathcal{H}}^2 + \lambda\|g\|_{\mathcal{G}}^2 + \tfrac{1}{4}\|g(x_t)\|_{\mathcal{H}}^2 \tag{50}$$

where $\hat{g}_s$ are the aggregate functions defined as in Appendix D, see Algorithm 3. The predictor $\hat{f}_t$ and the prediction $\hat{z}_t$ are computed as an optimisation problem in function of $\hat{g}_t(x_t)$

$$\hat{z}_t = \hat{f}_t(x_t) = \arg\min_{z\in\mathcal{Z}}\langle\psi(z), \hat{g}_t(x_t)\rangle_{\mathcal{H}}. \tag{51}$$

Analysis We now analyse the regret of this algorithm.

We introduce the following technical lemma, which bounds the cumulative risk of the feature predictors by the regret in high probability using Theorem 1 of van der Hoeven et al. (2023).

Lemma 12 (Dynamic Cumulative Risk of the Feature Predictors)**.** *Let $(\hat{g}_t)_t$ be defined as in Algorithm 3, $m \in \mathbb{N}$ be defined as in* (38)*, any sequence $(g_t^*)_t \in \mathcal{G}^T$ and $\delta \in (0, 1]$. Let $R_T(\lambda, m)$ be defined as follows*

$$\sum_{t=1}^{T}\|\hat{g}_t(x_t) - \varphi_t\|_{\mathcal{H}}^2 - \left\|\tfrac{1}{2}g_t^*(x_t) + \tfrac{1}{2}\hat{g}_t(x_t) - \varphi_t\right\|_{\mathcal{H}}^2 \le R_T(\lambda, m). \tag{52}$$

Let γ be defined as in Eq. (33)*. Then with probability $1 - \delta$*

$$\sum_{t=1}^{T}\mathbb{E}_t[\|\hat{g}_t(x_t) - \varphi_t\|_{\mathcal{H}}^2 - \|g_t^*(x_t) - \varphi_t\|_{\mathcal{H}}^2] \le 2R_T(\lambda, m) + 2\gamma\log(\delta^{-1}). \tag{53}$$

Proof. Let us recall the notations of van der Hoeven et al. (2023) in order to apply the Theorem 1 of their paper. Let $\tilde{\ell}_t$ denote the shifted loss as in van der Hoeven et al. (2023)

$$\tilde{\ell}_t(g) = \ell_t\left(\tfrac{1}{2}g + \tfrac{1}{2}\hat{g}_t\right) = \left\|\tfrac{1}{2}g(x_t) + \tfrac{1}{2}\hat{g}_t(x_t) - \varphi_t\right\|_{\mathcal{H}}^2. \tag{54}$$

We defined $R_T(\lambda, m)$ such that

$$\sum_{t=1}^{T} \tilde{\ell}_t(\hat{g}_t(x_t)) - \tilde{\ell}_t(g_t^*) \leq R_T(\lambda, m)$$

with the only assumption on $(g_t^*)_t$ that they are functions in the space $\mathcal{G}$. Using the convexity of $\tilde{\ell}$, we use Jensen inequality

$$\tilde{\ell}\left(\mathbb{E}_{g\sim Q_t}[g]\right) \leq \mathbb{E}_{g\sim Q_t}[\tilde{\ell}(g)],$$

where $(Q_t)_t$ are some distributions over $\mathcal{G}$, and by convexity of the space $\mathcal{G}$ we have that $\mathbb{E}_{g\sim Q_t}[g] \in \mathcal{G}$. We derive the following inequality

$$\sum_{t=1}^{T} \tilde{\ell}_t(\hat{g}_t) - \mathbb{E}_{g\sim Q_t}[\tilde{\ell}_t(g)] \leq \sum_{t=1}^{T} \tilde{\ell}_t(\hat{g}_t) - \tilde{\ell}\left(\mathbb{E}_{g\sim Q_t}[g]\right) \leq R_T(\lambda, m).$$

We now remark that the proof of the Theorem 1 of van der Hoeven et al. (2023) can be applied with a non-stationary baseline $(Q_t)_t$ instead of Q. This concludes the proof. □

We use this lemma and a comparison inequality to bound the cumulative risk of the predictors.

Theorem 13 (Expected Regret in a Non-Stationary Environment)**.** *Assume that there exists (g_t^*) a sequence in $\mathcal{G}$ such that $\mathbb{E}\left[\sum_{t=1}^{T}\|g_t^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2\right] = 0$. Let $\delta \in (0, 1]$. Then, Algorithm 3 run with $\lambda > 0$ and $\eta = 1/2(\kappa \sup\|g\|_{\mathcal{G}} + 1)^2$ satisfies with probability $1 - \delta$*

$$\mathbb{E}[R_T] = \begin{cases} \tilde{O}\left(V_{\mathcal{G}}^{1/6}T^{5/6} + T^{1/2}\sqrt{\log(\delta^{-1})}\right) & \text{if } \lambda = V_{\mathcal{G}}^{-1/3}T^{1/3} \\ \tilde{O}\left(V_0^{1/4}T^{3/4} + T^{1/2}\sqrt{\log(\delta^{-1})}\right) & \text{if } \lambda = V_0^{-1/2}T^{1/2} \end{cases}. \tag{55}$$

Precisely, let γ let be defined as in Eq. (33), if $\lambda = V_{\mathcal{G}}^{-1/3}T^{1/3}$ we have

$$\begin{aligned} &\sum_{t=1}^{T} \mathbb{E}_t[\Delta(\hat{f}_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)] \\ &\leq 2c_\Delta\sqrt{2T}\left(\begin{array}{l} \frac{\left\lceil V_{\mathcal{G}}^{2/3}T^{1/3}\kappa^{-2/3}\right\rceil \log T}{\eta} + (1 + \kappa\sup\|g\|_{\mathcal{G}})\,\kappa^{5/3}V_{\mathcal{G}}^{1/3}T^{2/3} \\ + (V_{\mathcal{G}}^{2/3}T^{4/3}\kappa^{-2/3} + T)^{1/2}\frac{\kappa^2B^2}{16}\log\left(e + \frac{e\kappa^2}{4}(V_{\mathcal{G}}^{2/3}T^{4/3}\kappa^{-2/3} + T)^{1/2}\right) \\ + (V_{\mathcal{G}}^{2/3}T^{4/3}\kappa^{-2/3} + T)^{1/2}\max_{t\in[\![T]\!]}\|g_t^*\|_{\mathcal{G}}^2 + \gamma\log(\delta^{-1}) \end{array}\right)^{1/2}, \end{aligned}$$

and if $\lambda = V_0^{-1/2}T^{1/2}$ we have

$$\begin{aligned} &\sum_{t=1}^{T} \mathbb{E}_t[\Delta(\hat{f}_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)] \\ &\qquad \leq 2c_\Delta\sqrt{2T}\left(\frac{V_0\log T}{\eta} + \frac{B^2\kappa^2\sqrt{TV_0}}{16}\log\left(e + \frac{e\kappa^2\sqrt{TV_0}}{4}\right) + \sqrt{TV_0}\max_i\|\bar{g}_{t_i:t_{i+1}}\|^2 + \gamma\log(\delta^{-1})\right)^{1/2}. \end{aligned}$$

Proof. **Step 1: Controlling the regret of $(f_t)_t$ by the regret of $(\hat{g}_t)_t$.** From Lemma 8, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}_t[\Delta(\hat{f}_t(x_t),y_t)-\Delta(f_t^*(x_t),y_t)]\\
&\quad\le 2c_\Delta\sqrt{T}\sqrt{\sum_{t=1}^{T}\mathbb{E}_t[\|\hat{g}_t(x_t)-\mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2]}\\
&\quad= 2c_\Delta\sqrt{T}\sqrt{\sum_{t=1}^{T}\mathbb{E}_t[\|\hat{g}_t(x_t)-\mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2-\|g_t^*(x_t)-\mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2+\|g_t^*(x_t)-\mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2]}\\
&\quad= 2c_\Delta\sqrt{T}\sqrt{\sum_{t=1}^{T}\mathbb{E}_t[\|\hat{g}_t(x_t)-\varphi_t\|_{\mathcal{H}}^2-\|g_t^*(x_t)-\varphi_t\|_{\mathcal{H}}^2]+\mathbb{E}_t[\|g_t^*(x_t)-\mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2]}.
\end{aligned}$$

We may now apply Lemma 12 to obtain

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}_t[\Delta(\hat{f}_t(x_t),y_t)-\Delta(f_t^*(x_t),y_t)]\\
&\quad\le 2c_\Delta\sqrt{T}\sqrt{2R_T(\lambda,m)+2\gamma\log(\delta^{-1})+\sum_{t=1}^{T}\|g_t^*(x_t)-\mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2}
\end{aligned}\tag{56}$$

where $R_T(\lambda,m)$ is defined as

$$R_T(\lambda,m)=\tfrac{m\log T}{\eta}+\tfrac{B^2}{4}\log\left(e+\tfrac{e\kappa^2T}{4\lambda}\right)\sum_{i=1}^{m}d_{\text{eff}}(4\lambda,t_{i+1}-t_i)+\lambda m\max_{t\in[\![T]\!]}\|g_t^*\|_{\mathcal{G}}^2+\tfrac{(1+\kappa\sup\|g\|_{\mathcal{G}})\kappa V_{\mathcal{G}}T}{m}$$

and comes from Proposition 11 for shifted losses.

Case 1: Continuous variations. We bound the effective dimension by $d_{\text{eff}}(4\lambda,t_{i+1}-t_i)\le\frac{\kappa^2(t_{i+1}-t_i)}{4\lambda}$. Thus we can bound the sum by $\sum_{i=1}^{m}d_{\text{eff}}(4\lambda,t_{i+1}-t_i)\le\kappa^2T/4\lambda$ and obtain

$$R_T(\lambda,m)\le\tfrac{m\log T}{\eta}+\tfrac{B^2}{4}\log\left(e+\tfrac{e\kappa^2T}{4\lambda}\right)\tfrac{\kappa^2T}{4\lambda}+\lambda m\max_{t\in[\![T]\!]}\|g_t^*\|_{\mathcal{G}}^2+\tfrac{(1+\kappa\sup\|g\|_{\mathcal{G}})\kappa V_{\mathcal{G}}T}{m}.$$

We choose $\lambda=\sqrt{T/m}$ and get

$$R_T(\lambda,m)\le\tfrac{m\log T}{\eta}+\tfrac{B^2}{4}\log\left(e+\tfrac{e\kappa^2\sqrt{Tm}}{4}\right)\tfrac{\kappa^2\sqrt{Tm}}{4}+\sqrt{Tm}\max_{t\in[\![T]\!]}\|g_t^*\|_{\mathcal{G}}^2+\tfrac{(1+\kappa\sup\|g\|_{\mathcal{G}})\kappa V_{\mathcal{G}}T}{m}.$$

We choose $m=\left\lceil V_{\mathcal{G}}^{2/3}T^{1/3}\kappa^{-2/3}\right\rceil$ and get

$$\begin{aligned}
R_T(\lambda,m)\le\ &\frac{\left\lceil V_{\mathcal{G}}^{2/3}T^{1/3}\kappa^{-2/3}\right\rceil\log T}{\eta}+(1+\kappa\sup\|g\|_{\mathcal{G}})\,\kappa^{5/3}V_{\mathcal{G}}^{1/3}T^{2/3}\\
&+(V_{\mathcal{G}}^{2/3}T^{4/3}\kappa^{-2/3}+T)^{1/2}\left(\tfrac{\kappa^2B^2}{16}\log\left(e+\tfrac{e\kappa^2}{4}(V_{\mathcal{G}}^{2/3}T^{4/3}\kappa^{-2/3}+T)^{1/2}\right)+\max_{t\in[\![T]\!]}\|g_t^*\|_{\mathcal{G}}^2\right).
\end{aligned}$$

We conclude by noting that $V_{\mathcal{G}}\le(2T-1)\sup\|g\|_{\mathcal{G}}$.

Case 2: Discrete variations. In the case of discrete distributions data variations there is no more need to approximate the data distributions $(g_t^*)_t$ by the hypothetical forecasters $(\bar{g}_{t_i:t_{i+1}})_i$. Thus the third term of the sum in Eq. (41) is not necessary. And we can replace $R_T(\lambda,m)$ by $R_T^0(\lambda,V_0)$ defined as

$$R_T^0(\lambda,V_0):=\frac{V_0\log T}{\eta}+\tfrac{B^2}{4}\log\left(e+\tfrac{e\kappa^2T}{4\lambda}\right)\sum_{i=1}^{V_0}d_{\text{eff}}(4\lambda,t_{i+1}-t_i)+\lambda\sum_{i=1}^{V_0}\|\bar{g}_{t_i:t_{i+1}}\|_{\mathcal{G}}^2.$$

We bound the effective dimension by $d_{\text{eff}}(4\lambda, t_{i+1}-t_i) \leq \frac{\kappa^2(t_{i+1}-t_i)}{4\lambda}$. Thus we bound the sum by $\sum_{i=1}^{V_0} d_{\text{eff}}(4\lambda, t_{i+1}-t_i) \leq \kappa^2 T/4\lambda$ and obtain

$$R_T^0(\lambda, V_0) \leq \frac{V_0 \log T}{\eta} + \frac{B^2\kappa^2 T}{16\lambda} \log\left(e + \frac{e\kappa^2 T}{4\lambda}\right) + \lambda V_0 \max_{i \in [\![V_0]\!]} \|\bar{g}_{t_i:t_{i+1}}\|_{\mathcal{G}}^2. \tag{57}$$

By choosing $\lambda = \sqrt{T/V_0}$, we get

$$R_T^0(\lambda, V_0) \leq \frac{V_0 \log T}{\eta} + \frac{B^2\kappa^2\sqrt{TV_0}}{16} \log\left(e + \frac{e\kappa^2\sqrt{TV_0}}{4}\right) + \sqrt{TV_0} \max_{i \in [\![V_0]\!]} \|\bar{g}_{t_i:t_{i+1}}\|_{\mathcal{G}}^2. \tag{58}$$

We then bound the cumulative risk using Step 1.

$$\sum_{t=1}^{T} \mathbb{E}_t[\Delta(\hat{f}_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)]$$
$$\leq 2c_\Delta\sqrt{2T}\sqrt{\frac{V_0 \log T}{\eta} + \frac{B^2\kappa^2\sqrt{TV_0}}{16} \log\left(e + \frac{e\kappa^2\sqrt{TV_0}}{4}\right) + \sqrt{TV_0} \max_{i \in [\![V_0]\!]} \|\bar{g}_{t_i:t_{i+1}}\|_{\mathcal{G}}^2 + \gamma \log(\delta^{-1})}$$

We now note that $V_0 \leq T$ to conclude the proof. □

E.1 Refined Regret Bounds

In this section we assume that the capacity condition holds and we derive refined bounds of the cumulative risk. For more details about the capacity condition see Appendix D.1. Assuming that there exists (g_t^*) a sequence in $\mathcal{G}$ such that $\mathbb{E}\left[\sum_{t=1}^T \|g_t^*(x_t) - \mathbb{E}[\varphi_t|x_t]\|_{\mathcal{H}}^2\right] = 0$. Let $\delta \in (0, 1]$. Then choosing $\lambda = T^{\frac{\beta}{\beta+1}} m^{\frac{-\beta}{\beta+1}}$ and $m = \left\lceil V_{\mathcal{G}}^{\frac{\beta+1}{\beta+2}} T^{\frac{1}{\beta+2}} \kappa^{\frac{-(\beta+1)}{\beta+2}} \right\rceil$ leads to the following bound on the expected regret with probability $1-\delta$

$$\sum_{t=1}^{T} \mathbb{E}_t[\Delta(\hat{f}_t(x_t), y_t) - \Delta(f_t^*(x_t), y_t)] = \tilde{O}\left(V_{\mathcal{G}}^{\frac{1}{2(\beta+2)}} T^{\frac{2\beta+3}{2(\beta+2)}} + T^{1/2}\sqrt{\log(\delta^{-1})}\right). \tag{59}$$

Indeed using the capacity condition, we bound the effective dimension by $d_{\text{eff}}(4\lambda, t_{i+1} - t_i) \leq Q\left(\frac{t_{i+1}-t_i}{4\lambda}\right)^\beta$. Using Jensen's inequality, we then derive

$$\begin{aligned}
\sum_{i=1}^{m} d_{\text{eff}}(4\lambda, t_{i+1} - t_i) &\leq Q \sum_{i=1}^{m} \left(\frac{t_{i+1} - t_i}{4\lambda}\right)^\beta \\
&= Qm \sum_{i=1}^{m} \frac{1}{m} \left(\frac{t_{i+1} - t_i}{4\lambda}\right)^\beta \\
&\leq Qm \left(\sum_{i=1}^{m} \frac{t_{i+1} - t_i}{4\lambda m}\right)^\beta \\
&= Qm^{1-\beta} \left(\frac{T}{4\lambda}\right)^\beta.
\end{aligned}$$

We obtain

$$R_T(\lambda, m) \leq \frac{m \log T}{\eta} + \frac{QB^2}{4} m^{1-\beta} \left(\tfrac{T}{4\lambda}\right)^\beta \log\left(e + \tfrac{e\kappa^2 T}{4\lambda}\right) + \lambda m \max_{t \in [\![T]\!]} \|g_t^*\|_{\mathcal{G}}^2 + \tfrac{(1+\kappa \sup\|g\|_{\mathcal{G}})\kappa V_{\mathcal{G}} T}{m}.$$

We choose $\lambda = T^{\frac{\beta}{\beta+1}} m^{\frac{-\beta}{\beta+1}}$, and obtain

$$R_T(\lambda, m) \leq \frac{m \log T}{\eta} + \frac{(1+\kappa \sup\|g\|_{\mathcal{G}})\kappa V_{\mathcal{G}} T}{m} + m^{\frac{1}{\beta+1}} T^{\frac{\beta}{\beta+1}} \left(\frac{QB^2 4^{-\beta}}{4} \log\left(e + \tfrac{e\kappa^2}{4} m^{\frac{\beta}{\beta+1}} T^{\frac{1}{\beta+1}}\right) + \max_{t \in [\![T]\!]} \|g_t^*\|_{\mathcal{G}}^2\right).$$

We choose $m = \left\lceil V_{\mathcal{G}}^{\frac{\beta+1}{\beta+2}} T^{\frac{1}{\beta+2}} \kappa^{\frac{-(\beta+1)}{\beta+2}} \right\rceil$ to obtain

$$R_T(\lambda, m) \leq \frac{\left\lceil V_{\mathcal{G}}^{\frac{\beta+1}{\beta+2}} T^{\frac{1}{\beta+2}} \kappa^{\frac{-(\beta+1)}{\beta+2}} \right\rceil \log T}{\eta} + (1 + \kappa \sup\|g\|_{\mathcal{G}})\, \kappa^{\frac{2\beta+3}{\beta+2}} V_{\mathcal{G}}^{\frac{1}{\beta+2}} T^{\frac{\beta+1}{\beta+2}}$$
$$+ \left(V_{\mathcal{G}}^{\frac{1}{\beta+2}} T^{\frac{\beta+1}{\beta+2}} \kappa^{\frac{-1}{\beta+2}} + T^{\frac{\beta}{\beta+1}} \right) \left(\frac{QB^2 4^{-\beta}}{4} \log \left(e + e^{\frac{\kappa^{\frac{2\beta+3}{\beta+2}} V_{\mathcal{G}}^{\frac{\beta}{\beta+2}} T^{\frac{2}{\beta+2}} + T^{\frac{1}{\beta+1}}}{4}} \right) + \max_{t \in [\![T]\!]} \|g_t^*\|_{\mathcal{G}}^2 \right).$$

We conclude by noting that $V_{\mathcal{G}} = \|g_1^*\|_{\mathcal{G}} + \sum_{t=2}^{T} \|g_t^* - g_{t-1}^*\|_{\mathcal{G}} \leq (2T-1) \sup\|g\|_{\mathcal{G}}$.