2019 18th European Control Conference (ECC)
Napoli, Italy, June 25-28, 2019

Online Optimisation for Online Learning and Control – From No-Regret to Generalised Error Convergence

J. Calliess[1]

Abstract— **This paper presents early work aiming at the development of a new framework for the design and analysis of algorithms for online learning based prediction and control. Firstly, we consider the task of predicting values of a function or time series based on incrementally arriving sequences of inputs by utilising online programming. Introducing a generalisation of standard notions of convergence, we derive theoretical guarantees on the asymptotic behaviour of the prediction accuracies when prediction models are updated by a no-external-regret algorithm. We prove generalised learning guarantees for online regression and provide an example of how this can be applied to online learning-based control. We devise a model-reference adaptive controller with novel online performance guarantees on tracking success in the presence of a priori dynamic uncertainty. Our theoretical results are accompanied by illustrations on simple regression and control problems.**

I. Introduction

Learning is useful in so far it enhances decision making. Often it is necessary to make decisions repeatedly in an uncertain dynamical world. Here, learning can be employed to inform a prediction model that forecasts the consequences of actions. In the light of information that becomes incrementally available over time, one would hope that a learning algorithm is capable of updating this model online with sufficient rapidity to facilitate better decisions and to adapt to changing dynamics. And, when actions have real-world impact, it is typically desired to have sufficient theoretical guarantees on the (long-term) dynamics of the system affected by the learning-based decisions. Since decisions will be based on predictions, understanding such dynamics will have to rest on guarantees on the online prediction performance of the learner.

No-regret algorithms and more generally, online programming algorithms can be utilised for fast online learning and prediction of time series (e.g. [12], [2]). If a no-external regret bound is achieved then, provided the prediction loss is convex in the parameters, (sub-) optimality guarantees can be given to bound the average prediction errors and the degree of sub-optimality of the parametric predictor whose parameter is the average of all choices of the adapted online predictors' parameters. This has given rise to algorithms that are no longer purely online learning and that decompose learning and prediction into two phases: a learning phase where the online learning method is employed to adapt the learner's parameter online for some time and a subsequent prediction phase making use of the average parameter obtained from the learning phase [12], [10].

[1]OMI, Dept. of Engineering Science, University of Oxford, UK.

With the aim to avoid such decompositions, we ask a more general question: Without the need to presuppose convexity or having a separate learning phase, does the no-regret property alone allow us to give guarantees on the increased online prediction success of the pure online learning method over time? While we show that, without further assumptions, the regret bound alone is not sufficient to ensure vanishing prediction losses in a pure online learning setup in the classical sense of convergence, we do show increasing prediction success in a more general sense: That is, while we will not be able to guarantee that the prediction loss will eventually remain below any upper bound forever, we can guarantee that it will do so for increasingly long durations, provided learning has been allowed to take place sufficiently long. This gives rise to a new generalised notion of convergence (and thereby of online-learnability) which we will refer to as convergence with *increasing permanence (i.p.)*.

Applying online programming algorithms to (parametric) online regression (cf. [21], [19]), we can then establish i.p.-convergence guarantees on the online prediction loss sequence. Moreover, building on a parametric online regression model to learn and predict a priori uncertain nonlinear dynamics, we derive theory that allows us to devise controllers that are guaranteed to regulate the state of an a priori uncertain nonlinear system to a desired region with increasing permanence. While this property is weaker then traditional desiderata (such as global asymptotic stability or convergence), we argue that it can be easier to achieve by computationally efficient learning-based controllers. Note, in combination with switching control architectures, our results can lead to control designs guaranteed to be eventually stable: For example, we may be satisfied to know our learning-based controller will eventually succeed to move the state into a region of state-space in which another (e.g. linear) controller is capable to take over and achieve stability.

We will introduce our generalised convergence concept in Sec. II and provide some general theoretical results. In Sec. III, we apply no-regret learning to online regression and provide new learning guarantees based on the results of the previous section. Sec. IV shows how to combine all preceding results into the design and theoretical analysis of a model-reference adaptive controller with theoretical guarantees on control success. The article will conclude with a brief summary and an outlook to future work. Owing to the page-limit, most of the derivations had to be confined to a preprint version of the paper [9].

II. Convergence with Increasing Permanence

Definition 1 (Convergence with increasing permanence): Let $\mathbb{S}$ denote a space endowed with metric $\mathfrak{d} : \mathbb{S}^2 \to \mathbb{R}$ and consider the sequence $\left(s_t\right)_{t\in\mathbb{N}}$ in $\mathbb{S}$. We say the sequence $(s_t)_{t\in\mathbb{N}}$ *converges* to $s' \in \mathbb{S}$ *with increasing permanence (.i.p.)*, written $s_t \rightsquigarrow s'$, if and only if the sequence remains in any given ball around s' for increasingly long durations. That is, $s_t \rightsquigarrow s' :\Leftrightarrow \forall \epsilon > 0, D, N \in \mathbb{N} \exists n \geq N \in \mathbb{N} \forall i \in \{1, ..., D\} : \mathfrak{d}(s', s_{n+i}) \leq \epsilon$.

It is easy to see that any sequence that is convergent in the classical sense also is i.p.-convergent. However, convergence with .i.p. is a more general concept than standard convergence. To see this consider the following example:

Example 1: Define the index set $Z_T := \{t \in \mathbb{N} | t \leq T, \exists n \in \mathbb{N} : t = 2^n\}$. Define the sequence (q_t) with $q_t := \begin{cases} 1, t \in Z_\infty \\ 1/t^2, \text{ otherwise} \end{cases}$. It is easy to show that we have $s_t \rightsquigarrow 0 \wedge s_t \rightsquigarrow 0$ but $s_t \nrightarrow 0$.

Just as with standard convergence, it will be convenient to consider convergence to sets:

Definition 2: A sequence $\left(s_t\right)_{t\in\mathbb{N}}$ converges to a set S with increasing permanence, written $s_t \overset{t\to\infty}{\rightsquigarrow} S$, iff $\inf_{s\in S} \mathfrak{d}(s, s_t) \overset{t\to\infty}{\rightsquigarrow} 0$.

In what is to follow we will consider real-valued sequences and convergence with respect to the canonical metric $\mathfrak{d}(s, s') = |s - s'|$.

Lemma 1: Assume we are given a non-negative real-valued sequence $(s_t)_{t\in\mathbb{N}}$ with $S_T := \frac{1}{T}\sum_{t=1}^{T} s_t \overset{T\to\infty}{\to} 0$. Then we have: $s_t \overset{t\to\infty}{\rightsquigarrow} 0$.

Remark 1: Note, that generally, convergence $\frac{1}{T}\sum_{t=1}^{T} s_t \overset{T\to\infty}{\to} 0$ does *not* imply classical convergence $s_t \to 0$. For a counterexample, consider the sequence (q_t) from Ex. 1. It is easy to check that indeed $\frac{1}{T}\sum_{t=1}^{T} q_t \overset{T\to\infty}{\to} 0$. But, as discussed above, (q_t) does not converge to 0 in the classical sense.

A. Contractive dynamical systems with increasingly permanently bounded disturbances

In set-point or tracking control, controllers often generate actions with the aim to turn the closed-loop error dynamics of a plant into a stable system with equilibrium $x^* = 0$. This means that the closed-loop dynamics can be represented by a contraction with that fixed-point. However, when the dynamics are not know a priori but are learned online, the actual dynamics deviate from a contraction by some time-varying disturbance. If the online learning method succeeds, this disturbance will eventually become increasingly small for increasing durations of time. Motivated by the analysis of such situations, we will next give i.p. convergence guarantees for disturbed contractive systems.

Theorem 1: Let $(\mathcal{X}, \|\cdot\|)$ be a normed vector space and $\phi : \mathcal{X} \to \mathcal{X}$ be a contraction with fixed point $x_* \in \mathcal{X}$ and Lipschitz constant $\lambda < 1$ relative to the metric canonically induced by norm $\|\cdot\|$. Let $(y_t)_{t\in\mathbb{N}}, (d_t)_{t\in\mathbb{N}}$ be sequences in $\mathcal{X}$ satisfying

$$y_{t+1} = \phi(y_t) + d_t \tag{1}$$

for all time steps $t \in \mathbb{N}_0$. We assume the *disturbances* d_t to be bounded. Let $r \geq 0$. If $\|d_t\| \overset{n\to\infty}{\rightsquigarrow} [0, r]$ then we have:

$$\|y_t - x_*\| \overset{t\to\infty}{\rightsquigarrow} \left[0, \frac{r}{1-\lambda}\right].$$

For the special case that the disturbances d_t vanish with i.p., the theorem guarantees that the perturbed sequence (y_t) also converges with i.p. to the fixed point x_*. Another important special case, which we consider below, is when $\mathcal{X}$ is finite-dimensional and $\phi(x) = Mx$ for some Schur (i.e. stable) matrix M with $\rho(M) < 1$. It is a special case since then ϕ is an eventually contracting map and hence, a contraction relative to some metric $\tilde{d}$ that is uniformly equivalent to the metric $\mathfrak{d} : (x, x') \mapsto \|x - y\|$ [17]. In this case we have:

Theorem 2: Let $\|\cdot\|, |||\cdot|||$ denote the Euclidean and spectral norms, respectively. Consider the recurrence $x_{t+1} = Mx_t + d_t$ $(t \in \mathbb{N})$. Let $\sigma = \sum_{i=0}^{\infty} |||M^i||| < \infty$. If the sequence of disturbances (d_t) is bounded and vanishes with increasing permanence up to error $r > 0$, i.e. $\|d_t\| \rightsquigarrow [0, r] \wedge \exists b \forall t : \|d_t\| \leq b$ then we have:

$$\|x_t\| \overset{t\to\infty}{\rightsquigarrow} [0, \sigma r].$$

III. Programming for Online Regression

A. Background: Online (Convex) Programming and No-Regret Algorithms

In an *online programming (OP)* problem [16], [23], an arbitrary sequence of (stage) *cost*, or *loss*, functions, $(\ell_t)_{t\in\mathbb{N}}$, on a domain F, is revealed step by step. At each *stage* or *time step* t, one is asked to choose an *action* a_t from a *feasible set* F. The choice is to be made on the basis of information $\mathbb{I}_t$ about past actions and cost functions up to stage $t-1$. After the choice is made, information $\mathbb{F}_t$ about the current cost function ℓ_t is revealed, and the algorithm suffers a loss amounting to $\ell_t(a_t)$. The information that the OP algorithm can use to choose the action a_t is summarised in the information set $\mathbb{I}_t = \left\{\left(a_q\right)_{q<t}, \left(\mathbb{F}_q\right)_{q<t}\right\}$. OP problems can arise in varying setups depending on the kind of information available to the algorithm at the time of decision making and the nature of the *loss feedback* it receives after having made the decision.

To measure the performance of an OP algorithm, we can compare its accumulated loss up to time step T to an estimate of the best cumulative cost attainable against the sequence $(\ell_t)_{t=1}^T$. In particular, we estimate the best attainable cost as the cost of the best constant action choice $a_T^* \in \operatorname{argmin}_{a\in F} \sum_{t=1}^{T} \ell_t(a)$ chosen with knowledge of the entire sequence $\ell_1, \ldots, \ell_T$. This choice leads to a measure called *external regret* or just *regret*: $\mathcal{R}(T) = \sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(a_T^*)$. An algorithm is said to be *no-(external)-regret* if it guarantees that $\max\big(0, \mathcal{R}(T)\big)$ is not contained in $\Omega(T)$. That is, if there exists a nonnegative sublinear function $\Delta(T) \in o(T)$ with $\mathcal{R}(T) \leq \Delta(T)$. The term *no-regret* is motivated by the fact that the limiting average nonnegative regret of a no-regret algorithm vanishes, i.e.,

$\limsup_{T\to\infty} \max(0, \mathcal{R}(T))/T = 0$. The sublinear function Δ is called a *regret bound*.

A prominent special case arises when both the feasible set and the stage loss funtions are convex. The pertaining problem is then called an *Online Convex Programming (OCP)* problem. Devising no-regret algorithms and bounds for OCP problems is an active area of research in theoretical computer science and machine learning (e.g. [7], [18], [1], [8], [20], [7]). A particular well-known no-regret algorithm to solve OCPs is the *Greedy Projection* (GP) algorithm [23], which requires feedback about the gradients of the stage loss functions. While here, we focus on the case, where noise-free losses are observable, there exist no-regret algorithms for noise corrupted observations [13], [5]. Recently, *Online Projected Stochastic Gradient Descent (OPSGD)* [7] has been proposed which is applicable in the case of pure bandit loss feedback but provides stochastic no-regret bounds that hold true with arbitrarily adjustable probability. While we focus on the case of deterministic no-external-regret bounds, all our results do extend to such stochastic settings, albeit our guarantees would then merely hold with the pertaining probabilistic confidence provided by the probabilistic no-regret bound.

No-regret bounds and algorithms have been studied and deployed in a great many online learning scenarios, including, among others, time-series prediction in ARMA models [2], multi-agent coordination [10], game-theory [15], [6], [22]. For a classic text book, the reader is referred to [12], whereas a recent survey can be found in [19]. In what is to follow, we will illustrate the application of our new convergence concept to the particular domain of online regression. In the context of kernel methods, online regression algorithms are briefly touched upon in [21]. However, their theoretical guarantees are limited to online classification. Applications to other online learning domains provide ample avenues to future work. A recent survey of existing approaches is provided in [19].

B. Application to Online Learning and Prediction

To keep the exposition concrete, we consider the following online learning and prediction problem: Let $(\mathcal{X}, \mathfrak{d}_\mathcal{X}), (\mathcal{Y}, \mathfrak{d}_\mathcal{Y})$ be two metric spaces.

An algorithm is given the task to predict a time series $\Big(y_t\Big)_{t\in\mathbb{N}} \in \mathcal{Y}^\mathbb{N}$ online on the basis of incrementally observing a related time series $\Big(x_t\Big)_{t\in\mathbb{N}} \in \mathcal{X}^\mathbb{N}$ and obtaining feedback after each prediction. We assume there exists a functional relationship

$$y_t = f(x_t) \tag{2}$$

where $f : \mathcal{X} \to \mathcal{Y}$ is some (a priori uncertain) *target function* residing in some class $\mathcal{F}$. To make predictions, the algorithm has access to a *hypothesis space* $\hat{\mathcal{F}}$ of *predictors* $\hat{\mathfrak{f}}(\cdot;\theta) : \mathcal{X} \to \mathcal{Y}$ $(\theta \in \Theta)$ parametrised by *parameter space* Θ. We assume prediction accuracy is measured by a non-negative loss function

$$\ell(\cdot;\cdot) : \mathcal{X} \times \Theta \to \mathbb{R}_{\geq 0}$$

which is zero for inputs x and parameters θ for which $f(x) = \hat{\mathfrak{f}}(x;\theta)$. For example, the loss might quantify the squared distance, i.e. $\ell(x;\theta) = \mathfrak{d}_\mathcal{Y}\big(f(x), \hat{\mathfrak{f}}(x;\theta)\big)^2$. To connect to the OP setup, we can define the stage loss function

$$\ell_t : \theta \mapsto \ell(x_t;\theta)$$

At the beginning of each stage at time step $t \in \mathbb{N}$, the algorithm has to pick a parameter θ_t and use it to make a prediction utilising the chosen predictor:

$$\hat{y}_t = \hat{\mathfrak{f}}(x_t;\theta_t).$$

The parameter θ_t (and hence, the predictor) is chosen on the basis of information set $\mathbb{I}_t$. For now, we assume this set contains all previous inputs $(x_1, ..., x_{t-1})$ (but might exclude x_t) and information about the pertaining prediction losses. (For example, the latter might be given by revelation of the true $y_t = f(x_t)$ after each prediction at time t, from which the loss function ℓ_t can be computed.) The stage at time t concludes by revelation of the loss information after the prediction was made and the process enters the next stage at the next time step $t + 1$.

A special case of this setting is online regression. Here the task is learning f online and becoming better at predicting its output values based on an i.i.d. input samples that become incrementally available.

1) Example: Online regression with Radial Basis Function Neural Networks: As a concrete example, consider the case where the hypotheses class $\mathcal{F}$ is a set of radial-basis function neural networks (RBFNN) with known structure:

The component functions $f_i : \mathbb{R}^d \to \mathbb{R}$ $(i = 1, ..., d')$ of target function f can be represented by an RBFNN with some weight parameter $\theta_i^* \in \mathbb{R}^m$. These weights are assumed to be unknown a priori but contained in some known convex feasible set $F \subset \mathbb{R}^m$. That is $\forall i \exists \theta_i^* \in F : f_i(x) = \langle \theta_i^*, \phi_i(x) \rangle$. Here, $\phi : \mathbb{R}^d \to \mathbb{R}^m$ with each component function $\phi_i(\cdot) = \exp(-\frac{\|\cdot - c_i\|^2}{\sigma_i^2})$ being a radial basis function.

In this representation, the online learning task can be reduced to updating the weights $\theta_t := (\theta_{1,t}, ..., \theta_{d,t})$ at time t and insert it into the prediction hypothesis $\hat{\mathfrak{f}}(\cdot;\theta) := \langle \theta, \phi(\cdot) \rangle$ to predict the next observation as per $\hat{y}_t := \hat{\mathfrak{f}}(x_t;\theta_t)$.

To reduce the learning task to solving an OCP, once we have received the true value $y_t = f(x_t)$, we can compute and feed back information about the stage loss function $\ell_t : \theta \mapsto \Big\| \hat{\mathfrak{f}}_t(x_t;\theta) - y_t \Big\|_2^2$. It is easy to check that, by construction, each $\ell_t(\cdot)$ is a convex function. Therefore, by interpreting the weights as actions $(a_t := \theta_t)$ in the OCP paradigm, the weight hypotheses θ_t can be generated online by utilising any existing no-regret algorithm designed for OCPs.

In the following example simulation, we have done so employing Greedy Projection [23] (refer to Fig. III-B.1). Here, the task was to predict the real-valued time series $y_t = f(x_t) = \langle \theta^*, \phi(x_t) \rangle$ where $x_t, y_t \in \mathbb{R}, \forall t$ and $\theta^* \in \mathbb{R}^4$ was drawn at random and we chose RBF centres $c_1 = -1, c_2 = -1/3, c_3 = 1/3, c_4 = 1$ and length scale parameters $\sigma_1 = -1, \sigma_2 = -1/3, \sigma_3 = 1/3, \sigma_4 = 1$. At the start

of the online prediction task the initial hypothesis parameter θ_1 was drawn at random. Subsequently we generated the time series y_t by sampling the sequence x_t uniformly i.i.d. at random from the interval [-2,2] and computing $y_t = f(x_t)$. The updates of the θ_t were computed with Greedy Projection (GP). Requiring gradient information of the stage loss functions, we fed GP the stage prediction loss gradient $\nabla_{\theta_t}\ell_t(\theta_t) = 2(\hat{\mathfrak{f}}(x_t;\theta_t))\phi(x_t)$ after it had computed a new weight θ_t at each time step $t = 1, ..., 100$.

Some simulation results are depicted in Fig. III-B.1. Note, at the final recorded time step $T = 100$, the hypothesis $\hat{\mathfrak{f}}(\cdot;\theta_t)$ closely matched the ground truth $f(\cdot)$ (cf. Fig. III-B.1.a and Fig. III-B.1.c). Furthermore, we note that the prediction losses seemed to vanish with increasing learning experience (cf. Fig. III-B.1.d). In the next subsection, we will investigate to what extent such a behavior can be predicted on the basis of the no-regret properties of the learning algorithm we utilised.

C. Prediction loss convergence with increasing permanence

Refer to Lem. 1. Its implications for no-regret learning in OPs are clear: while, without further assumptions, from the fact that $\frac{1}{T}\sum_{t=1}^{T}\ell_t(\theta_t) \to 0$ alone, we cannot infer that the prediction loss will converge to zero in the traditional sense, we can guarantee convergence with increasing permanence:

Theorem 3: Consider the online learning and prediction problem in an OP setting (as for instance considered above) where $\ell_t(\theta_t) \geq 0$ is the stage prediction error of the prediction model with parameter θ_t and where the prediction model class is sufficiently expressive to guarantee that $\min_\theta \ell_t(\theta) = 0, \forall t$ (which is the case e.g. when $\hat{\mathcal{F}} \supseteq \mathcal{F}$).

If the θ_t are updated with a no-regret algorithm suitable for the given OP then the prediction errors vanish with increasing permanence. That is we have:

$$\ell_t(\theta_t) \rightsquigarrow 0 \quad (\text{as } t \to \infty).$$

Proof: Let $\theta^* \in \Theta$ such that $\hat{\mathfrak{f}}_{\mathfrak{n}}(\cdot;\theta^*) = f(\cdot)$. Thus, $\ell_t(\theta^*) = \min_{\theta\in\Theta}\ell_t(\theta) = 0$. Considering that $\ell_t(\theta_t) \geq 0 \forall t$, the no-regret guarantee ensures $\frac{R(T)}{T} = \frac{1}{T}\sum_{t=1}^{T}\ell_t(\theta_t) - \frac{1}{T}\sum_{t=1}^{T}\ell_t(\theta^*) = \frac{1}{T}\sum_{t=1}^{T}\ell_t(\theta_t) \overset{T\to\infty}{\longrightarrow} 0$. Appealing to Lem. 1 gives the desired result. ∎

Note, the theorem applies to the online prediction setup of the previous section, guaranteeing that GP learning of the RBFNNs results in prediction errors that converge to zero with increasing permanence.

The theorem above is valid irrespective of the nature of the time series. However, the assumption that the target f is contained in the hypothesis space is a limitation. In practice, we might have the situation where the hypothesis space and the target class $\mathcal{F}$ are distinct with a distance given by some *representational model class error* $r = \sup_{f\in\mathcal{F}}\inf_{\hat{\mathfrak{f}}\in\hat{\mathcal{F}}}\mathfrak{d}(f,\hat{\mathfrak{f}})$ for some suitable metric $\mathfrak{d}$ whose choice depends on the concrete online learning problem. In what is to follow we extend our convergence guarantees to two such learning problems in the presence of representational error.

1) Online regression with i.i.d. inputs and representational mean-square model error: Consider online regression. In this standard learning scenario, the inputs are assumed to be drawn i.i.d. from a distribution with density p with support $\mathcal{X}$. We assume at each stage t, the new predictor's parameter θ_t is to be picked based on the history $\mathbb{I}_t$ of past observations of input -output pairs (x_i, y_i) (or just past prediction losses $\ell_i(\theta_i))\,(i < t)$. After this, (x_t, y_t) and the prediction loss $\ell(x_t;\theta_t)$ can be computed (alternatively, this new loss is revealed) which concludes the stage. Typically, one considers mean-square regression with a loss $\ell(x;\theta) = \left\|\hat{\mathfrak{f}}(x;\theta) - f(x)\right\|^2$. Let $\langle\cdot\rangle$ denote the expectation operator. The expectation $\langle\ell(x;\theta)\rangle$ is the standard stochastic mean-square loss. Relative to this loss, we can define the model error $r_f = \inf_{\hat{\mathfrak{f}}\in\hat{\mathcal{F}}}\mathfrak{d}(f,\hat{\mathfrak{f}})$ for a given function $f \in \mathcal{F}$ where $\mathfrak{d}(f,g) := \langle\|f-g\|_2^2\rangle$. Furthermore, we can consider the worst-case model class error $r_2 := \sup_{f\in\mathcal{F}} r_f$. Our theory developed so far assures us that the online mean-square prediction losses i.p.-converge to a set that is below this representational model error:

Theorem 4: Assume the online regression task is performed by a prediction algorithm that suffers sub-linear external regret, i.e. where the θ_t are chosen such that $\exists\Delta \in o(T)\forall T \in \mathbb{N} : \sum_{t=1}^{T}\ell(x_t;\theta_t) \leq \inf_{\theta\in\Theta}\sum_{t=1}^{T}\ell(x_t;\theta) + \Delta(T)$. Then the sequence $(\langle\ell(x_t;\theta_t)\rangle)_{t\in\mathbb{N}}$ of expected prediction losses converges to at most the representational error with i.p., that is:

$$\langle\ell(x_t;\theta_t)\rangle \overset{t\to\infty}{\rightsquigarrow} r_f \in [0, r_2].$$

Applied to our example of RBFN-based online regression, the theorem states that the mean-square prediction error of the predictors that are found online i.p. -converges to the best mean-square representational error attainable by the presupposed RBFN structure.

IV. No-Regret Learning-Based Model-Reference Adaptive Control

As mentioned above, our results are meaningful in online-learning based model-reference adaptive control. Consider a dynamical system $\ddot{x} = f(x) + a(x) + b(x)u$ where x denotes the state and u denotes the control action. We assume that a and b are known a priori and that the inverse b^{-1} can be computed for all states x. By contrast, f is uncertain and capturing model discrepancies due to environmental conditions of the system the plant operates i, which are hard to model a priori. If f was perfectly known, a standard approach would be to feedback linearise the system, setting $u(x) := b^{-1}(x)(-a(x) - f(x) + \ddot{x}_{ref})$ where $\ddot{x}_{ref}$ is some reference behaviour we desire the closed-loop dynamics to exhibit. For example, in tracking where we desire x_t to follow a target trajectory ξ_t, one approach would be to set $\ddot{x}_{ref} = K(\xi - x)$ where K a stabilising feedback matrix ensuring that ξ_t will eventually be tracked by the reference state trajectory x_{ref} with sufficient accuracy. In the absence of perfect knowledge of f we can replace $f(x_t)$ by a predictor $\hat{\mathfrak{f}}(x_t;\theta_t)$ in the feedback-linearising law $u(x_t)$ and to learn the parameters of the predictor online. There are

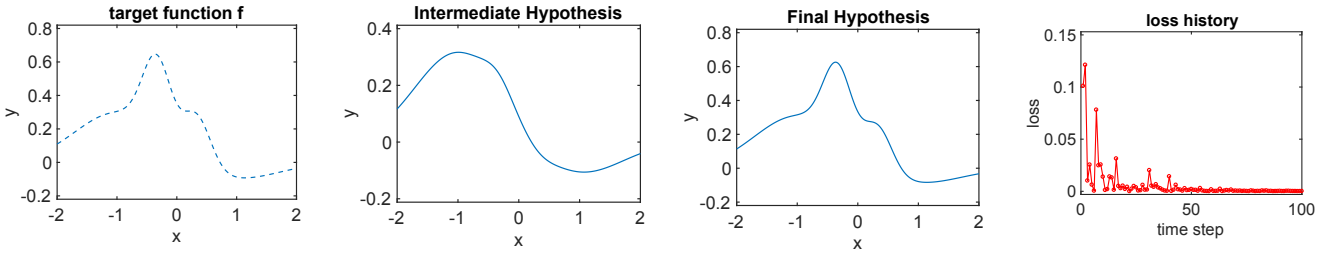


Fig. 1. (a): The ground truth f. (b): The hypothesis $\hat{\mathfrak{f}}(\cdot, \theta_5)$ used for prediction after 4 learning iterations. (c): The final hypothesis $\hat{\mathfrak{f}}(\cdot, \theta_{100})$. (d): The evolution of the prediction losses $\ell_t(\theta_t)$ for $t = 1, ..., 100$.

many learning methods we can employ to this end, including updates of neural network weights [3] or nonparametric learning methods such as Gaussian processes [14]. In this work, we propose to learn this predictor online with an OP-based online regression method as describedin in Sec. III.

It can be shown, that when defining the error by $e = x_{ref} - x$, in a first-order Euler-discretised version, the error dynamics become (see e.g. [11]):

$$e_{t+1} = Me_t + d_t.$$

Here, M is a stable matrix, disturbance $d_t = \tau f(x_t) - \tau \hat{\mathfrak{f}}(x_t; \theta_t)$ is the prediction error and τ is the time increment from the time-discretisation. Assume we employ online regression to online-learn $\hat{\mathfrak{f}}$ (by adapting the parameters θ_t) with a sublinear regret incurring algorithm OP algorithm. In that case, with some additional assumptions, we can guarantee that the error trajectory vanishes with increasing permanence.

Theorem 5: Suppose that (i) f resides in the model class $\hat{\mathcal{F}}$ and that (ii) $\sup_{\hat{\mathfrak{f}} \in \hat{\mathcal{F}}} \left\| f - \hat{\mathfrak{f}} \right\|_\infty$ is bounded. If the parameters θ_t are incrementally updated with an algorithm that incurs sublinear external regret and receives loss feedback $\ell_t(\theta_t) = \|d_t\|_2^2$ then the error trajectory vanishes with increasing permanence, i.e. $e_t \rightsquigarrow 0$. *Proof:* Owing to (i) and the no-regret assumption Thm. 3 is applicable, which entails that $d_t \rightsquigarrow 0$. In conjunction with (ii), this allows us to appeal to Thm. 2 which gives the desired statement. ■

A. Example – pendulum control

To illustrate the viability of no-regret learning based control on a simple example we, we consider the following pendulum control problem:

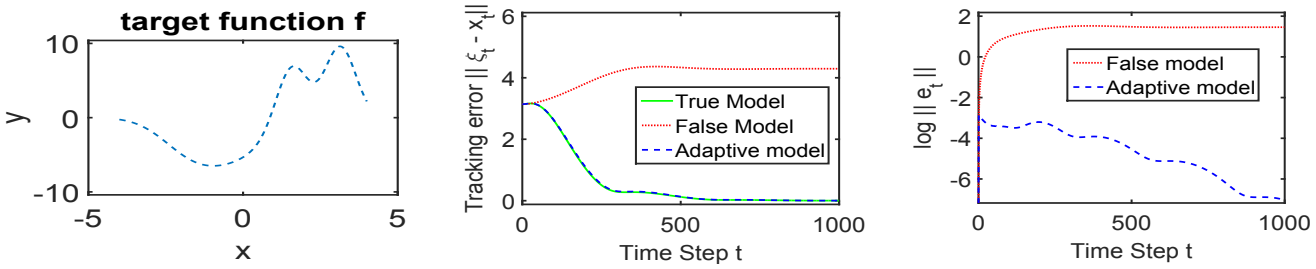
We explored our method's properties in simulations of a rigid pendulum with (a priori known) drift $a(x) := -\frac{g}{l}\sin(x_1) - \frac{r(x_1)}{ml^2}x_2$ and constant control input function $b(x) = \frac{1}{ml^2}$. Here, $x_1 = q, x_2 = \dot{q} \in \mathbb{R}$ are joint angle position and velocity, r denotes a friction coefficient, g is acceleration due to gravity l is the length and m the mass of the pendulum. The control input $u \in \mathbb{R}$ applied a torque to the joint that corresponds to joint-angle acceleration. With q denoting the joint angle, $q = 0$ pertains to a state where the pendulum is pointing downward and $q = \pi$ denotes a position in which the pendulum is upward. Given an initial configuration $x_0 = [0; 0]$, we desired to steer the state to a terminal configuration $\xi = [\pi, 0]$. We applied a feedback linearising control law corresponding to the continuous-time law $u(x_t; \theta_t) := b^{-1}(x_t)(-a(x_t) - \hat{\mathfrak{f}}(x_t; \theta_t) - K(\xi - x_t))$. Here K was (an underdamping) PD-controller feedback matrix ensuring global asymptotic stability of ξ in the perfectly feedback linearised reference system $\ddot{x}_{ref} = K(\xi - x_{ref})$. To connect to our theory, we discretised the system and control laws by a first-order Euler approximation. We documented the behavior of three different settings: (i) Where $f(\cdot) = \hat{\mathfrak{f}}(\cdot) = 0$, corresponding to a situation where there is no model error. (ii) Where $f(x) = \sum_{i=1}^4 w_i^* \exp(|x_1 - c_i| / \sigma_i)$ with $w^* = [-12, -10, 10, 12], c = [-\pi/2, 0, \pi/2, \pi], \sigma = [1, 1, .5, .5]$ being a mixture of Gaussians while the controller falsely modeled this function by the static predictor $\hat{\mathfrak{f}} \equiv 0$; (iii) Where f was as before but, at each time t, the predictor was chosen to be $\hat{\mathfrak{f}}(\cdot; \theta_t) = \sum_{i=1}^4 \theta_{t,i} \exp(|x_1 - c_i| / \sigma_i)$ and parameter θ_t updated with Greedy Projection[23]. The initial parameter was $\theta_0 = [0, 0, 0, 0]$.

As learning feedback prediction loss $\ell_t(\theta_t) = \left\| \hat{\mathfrak{f}}(x_t; \theta_t) - \left(\ddot{x}_t - a(x_t) - b(x_t)u(x_t; \theta_t)\right) \right\|_2^2$ was fed back to the no-regret algorithm after each discrete time step t.

According to our theory the reference error e_t should vanish (at least with increasing permanence). Simulation results are depicted in Fig. IV-A. The documented behaviour is consistent with our theoretical guarantees. Note, that the linearising controller without the adaptive element that operates on the basis of the static inaccurate model led to poor performance. By contrast, when the controller was set up with no-regret learning in place as described above (and matching the assumptions of Thm. 4) seamlessly managed to control the state to the target.

V. Conclusions

This paper discussed early work considering the application of online programming (in particular, of no-regret algorithms) in online learning-based control settings. As a first proposal of how to think about the implications of no-regret guarantees in control, we introduced the concept of convergence with *increasing permanence (i.p.)* as an objective for online learning and control. We discussed conditions under which such objectives are met when no-regret algorithms are employed in online regression and learning-based model-reference adaptive control. We have given illustrations of



(a) (b) (c)

Fig. 2. Pendulum control task. (a): The ground truth model error f. (b): Recorded tracking errors incurred by linearsing controllers with knowledge of the true f (green plot), under the false assumption that $f \equiv \hat{\mathrm{f}} \equiv 0$ (red plot) and when utilising our no-regret learning based prediction model $\hat{\mathrm{f}}(x_t; \theta_t)$ with the parameters updated with GP (blue plot). Note, how falsely fixing $\hat{\mathrm{f}} \equiv 0$ caused the second controller to completely fail to track the target. By contrast, our learning-based controller adapted quickly and learned to accurately track the target. (c): Logarithm of the norm of the error dynamics (deviation from the reference) for the linearising controller based on the false model (red plot) and for our learning based approach (blue plot).

such applications of no-regret algorithms to online regression and control tasks.

Being work at an early stage, many possible directions of further enquiry remain. Firstly, i.p.-convergence appears to be a relatively weak notion that may not be satisfactory to be assured of in many control settings. Future work could investigate how regret-bounds can lead to refined convergence rates and lead to stronger guarantees if additional properties are known such as bounded excitation (i.e. target function values), Lipschitz constants and convexity properties. Secondly, on the practical side, we will devise nonparametric extensions of no-regret learning based controllers that would be able to learn and control dynamics more flexibly. In addition, we will investigate of how to systematically deal with observational noise and representational model error both from a practical and a theoretical perspective. For simplicity, our exposition was limited to the case of noise-free observations (translating to noise-free loss observations). Fortunately however, there exist a variety of online convex programming algorithms that address noisy losses [13], [5], [4], [2] the offer expected regret bounds that can be converted into high-probability bounds. If we applied those, all our results presented in this paper would hold with arbitrarily high probability. Furthermore, we would find it interesting to investigate applications to model-predictive control.

In summary, this paper presented a first step towards drawing a bridge between control and the online optimisation community in theoretical machine learning. While many open theoretical challenges remain and details will have to be worked on in the future, we hope to have provided first evidence that utilisation of no-regret learning techniques might be a potentially fruitful direction in the design and analysis of new online-learning based controllers.

References

[1] Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Interior-point methods for full-information and bandit online learning. *IEEE Transactions on Information Theory*, 58(7):4164–4175, 2012.

[2] O. Anava, E. Hazan, S. Mannor, and O. Shamir. Online Learning for Time Series Prediction. *arXiv preprint arXiv:1302.6927*, 2013.

[3] Ryan T. Anderson, Girish Chowdhary, and Eric N. Johnson. Comparison of rbf and shl neural network based adaptive control. *Journal of Intelligent and Robotic Systems*, 54(1):183–199, Mar 2009.

[4] A. Argawal, D. P. Foster, D. Hsu, S. M. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. *Siam J. Optim.*, 2013.

[5] E. V. Belmega, P. Mertikopoulos, R. Negrel, and L. Sanguinetti. Online Convex Optimisation and No-Regret Learning: Algorithms, Guarantees and Applications. *eprint arXiv:1804.04529v1*, 2018.

[6] Avrim Blum, Eyal Even-Dar, and Katrina Ligett. Routing without regret: on convergence to nash equilibria of regret-minimizing algorithms in routing games. In *PODC '06: Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing*, pages 45–52, 2006.

[7] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

[8] Sébastien Bubeck, Ofer Dekel, Tomer Koren, and Yuval Peres. Bandit convex optimization: $\sqrt{T}$ regret in one dimension. In *In Proceedings of the 28st Annual Conference on Learning Theory (COLT)*, 2015.

[9] J. Calliess. Online Optimisation for Online Learning and Control – From No-Regret to Generalised Error Convergence. *eprint arXiv*, 2019.

[10] J. Calliess and Geoffrey J. Gordon. No-regret learning and a mechanism for distributed multiagent planning. In *AAMAS*, 2008.

[11] Jan-Peter Calliess. *Conservative decision-making and inference in uncertain dynamical systems*. PhD thesis, University of Oxford, 2014.

[12] Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, 2006.

[13] N. Cesa-Bianchi, Shai Shalev-Shwartz, and Ohad Shamir. Online Learning of Noisy Data. *IEEE Trans. Inf. Th.*, 2010.

[14] G. Cho, G. Chowdhary, A. Kingravi, J. P. . How, and A. Vela. A Bayesian nonparametric approach to adaptive control using Gaussian processes. In *CDC*, 2013.

[15] Yoav Freund and Robert E. Shapire. Game theory, on-line prediction and boosting. In *COLT*, 1996.

[16] Geoffrey J. Gordon. Regret bounds for prediction problems. In *COLT: Workshop on Computational Learning Theory*, 1999.

[17] B. Hasselblatt and A. Katok. *A First Course in Dynamics with a Panorama of Recent Developments*. Cambridge University Press, 2003.

[18] Elad Hazan and Yuanzhi Li. An optimal algorithm for bandit convex optimization. *arXiv preprint arXiv:1603.04350*, 2016.

[19] S. C. H. Hoi, D. Sahoo, J. Lu, and P. Zhao. Online Learning: A Comprehensive Survey. *CoRR, arXiv preprint*, arXiv:1802.02871v1, 2018.

[20] Xiaowei Hu, Prashanth LA, András György, and Csaba Szepesvari. (Bandit) Convex Optimization with Biased Noisy Gradient Oracles. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 819–828, 2016.

[21] J. Lu, S. C. Hoi, J. Wang, P. Zhao, and Z.-Y. Liu. Large scale online kernel learning. *Journal of Machine Learning Research*, 2016.

[22] T. Roughgarden, V. Syrgkanis, and E. Tardos. The Price of Anarchy in Auctions. *eprint arXiv:1607.07684*, 2016.

[23] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, 2003.