Proceedings of Machine Learning Research vol XX:1–34, 2023

Regret Guarantees for Online Deep Control

Xinyi Chen XINYIC@PRINCETON.EDU
Princeton University, Google AI Princeton

Edgar Minasyan MINASYAN@PRINCETON.EDU
Princeton University, Google AI Princeton

Jason D. Lee JASONLEE@PRINCETON.EDU
Princeton University, Google AI Princeton

Elad Hazan EHAZAN@PRINCETON.EDU
Princeton University, Google AI Princeton

Abstract

Despite the immense success of deep learning in reinforcement learning and control, few theoretical guarantees for neural networks exist for these problems. Deriving performance guarantees is challenging because control is an online problem with no distributional assumptions and an agnostic learning objective, while the theory of deep learning so far focuses on supervised learning with a fixed known training set.

In this work, we begin to resolve these challenges and derive the first regret guarantees in online control over a *neural network*-based policy class. In particular, we show sublinear *episodic* regret guarantees against a policy class parameterized by deep neural networks, a much richer class than previously considered linear policy parameterizations. Our results center on a reduction from online learning of neural networks to online convex optimization (OCO), and can use any OCO algorithm as a blackbox. Since online learning guarantees are inherently agnostic, we need to quantify the performance of the best policy in our policy class. To this end, we introduce the interpolation dimension, an expressivity metric, which we use to accompany our regret bounds. The results and findings in online deep learning are of independent interest and may have applications beyond online control.

1. Introduction

The use of deep neural networks has been highly successful in reinforcement learning (RL) and continuous control problems. However, a theory for deep control and RL remains challenging. The main difficulty in applying the theory developed for supervised learning to the RL domain is the distributional assumptions and realizability goal made in the literature thus far. In control and RL, the environment is inherently online and often nonstochastic, and the goal is usually agnostic learning with respect to a policy class.

In this work, we consider the problem of online episodic control with neural network-based policies. We begin to resolve the aforementioned challenges and derive the first regret bound guarantees in this setting. Provable regret bounds in this domain have thus far been limited to linear controllers. However, most dynamical systems in the physical world are

© 2023 X. Chen, E. Minasyan, J.D. Lee & E. Hazan.

nonlinear and/or require nonlinear controls. An important tool that allows us to go beyond linear controllers is the emerging paradigm of online nonstochastic control: a methodology for control that is robust to adversarial noise in the dynamics. The important aspect of this paradigm to our study is that it uses policy classes that admit a convex parameterization.

It is natural to consider the online *episodic* control setting: although it is less challenging from a technical perspective than single-trajectory control, the policy learning procedure in empirical deep control is done episodically as detailed in the related work section. The main technical challenge to this goal is formalizing online learning over deep neural networks and proving accompanying regret bounds. Given this result, an extension to the single-trajectory setting is possible but often does not reflect standard practice in empirical research.

As the major technical component of this work, we propose a black-box reduction from online deep learning to online convex optimization (OCO) that attains provable regret bounds. These bounds apply to the general online learning setting with vector output predictors and arbitrary convex loss functions. Moreover, the regret guarantees are naturally *agnostic*, i.e. they show performance competitive to the best neural network in our policy class in hindsight without assuming it achieves zero loss. To capture agnostic learning and derive meaningful guarantees for online learning and control, we also introduce a new metric of expressivity, namely the "interpolation dimension", that accompanies our regret bounds.

An interesting conclusion from this reduction is the unifying view that provable convergence and/or generalization bounds for training deep neural networks can be derived for any OCO method, beyond online and stochastic gradient descent. This includes mirror descent, adaptive gradient methods, follow-the-perturbed leader and other algorithms. Previously, convergence and generalization analyses for neural networks were done in isolation for different optimization algorithms as detailed in the related work section.

Our contributions in this work can be summarized as follows:

- **Online episodic deep control:** We derive the first provable regret guarantees in online episodic control with policies based on deep neural networks. Furthermore, we demonstrate the richness of the considered policy class by showing that it can output the optimal *open-loop* control sequence of any single episode.
- **Online learning over neural networks:** We give a general reduction from online learning of neural networks to OCO that can use any OCO algorithm as a blackbox.
- **Interpolation dimension:** To state meaningful guarantees in online agnostic learning, we introduce the interpolation dimension as an expressivity metric. It is a fundamental notion and applies to any hypothesis class.
- **Unifying analysis:** Our proposed method applies to any OCO algorithm, including mirror descent and adaptive gradient methods widely used in deep learning. This leads to a unifying framework for optimization in deep learning: the online learning framework implies both convergence and generalization bounds in the supervised learning setting.

1.1. Related work

Online and nonstochastic control. Our study focuses on algorithms which enjoy sublinear regret for online control of dynamical systems; that is, whose performance tracks

a given benchmark of policies up to a term which is vanishing relative to the problem horizon. Abbasi-Yadkori and Szepesvári (2011) initiated the study of online control under the regret benchmark for linear time-invariant (LTI) dynamical systems. Bounds for this setting have since been improved and refined in Dean et al. (2018); Mania et al. (2019); Cohen et al. (2019); Simchowitz and Foster (2020). Our work instead adopts the online *nonstochastic* control setting (Agarwal et al., 2019), that allows for adversarially chosen (e.g. non-Gaussian) noise and general convex costs that may vary with time. This model has been studied for many extended settings, see Hazan and Singh (2021) for a comprehensive survey. Similar to our control framework, online episodic control is also studied in Kakade et al. (2020), but the regret definition differs from ours, the results are only information-theoretic and the system is linear in a kernel space. In terms of nonlinear systems, one common approach in control is iterative linearization which takes the local linear approximation via the gradient of the nonlinear dynamics. One can apply techniques from optimal control to solve the resulting changing linear system. Iterative planning methods such as iLQR (Tassa et al., 2012), iLC (Moore, 2012) and iLQG (Todorov and Li, 2005) fall into this category. Recent works (Roulet et al., 2022; Westenbroek et al., 2021) provide theoretical results and insights to this approach but many theoretical questions about the approach remain open.

The emerging theory of deep learning. For detailed background on the developing theory for deep learning, see the book draft (Arora et al., 2021). Among the various studies on the theory of deep learning, the neural tangent kernel (NTK or linearization) approach has emerged as the most complete and pervasive: it is not currently believed to fully explain the practical success but there is no alternative substantial theory yet. This technique shows that neural networks behave similar to their local linearization and proves that gradient descent converges to a global minimizer of the training loss (Soltanolkotabi et al., 2018; Du et al., 2018a,b; Jacot et al., 2018; Bai and Lee, 2019; Lee et al., 2019). The NTK approach/regime has been expanded to provide various generalization error bounds (Arora et al., 2019; Wei et al., 2019; Cao and Gu, 2019; Ji and Telgarsky, 2020), and adversarial training guarantees (Gao et al., 2019; Zhang et al., 2020). As opposed to our generic approach, a number of different optimization algorithms have been considered in isolation for analyzing deep learning theory in the NTK regime including (Wu et al., 2019; Cai et al., 2019; Wu et al., 2021; Zhang et al., 2019).

The results in this work extend upon the described deep learning theory literature; in particular, we use the same deep learning setup, and follow techniques and results from Gao et al. (2019); Allen-Zhu et al. (2019). Furthermore, several works in the literature (Cao and Gu, 2019; Gao et al., 2019; Zhang et al., 2020) have observed and used online components in their derivations of generalization and adversarial training guarantees. We note that all these works, unlike our contributions, operate in the supervised learning setting.

Online convex optimization and dimensionality notions in learning. The framework of learning in games has been extensively studied as a model for learning in adversarial and nonstochastic environments (Cesa-Bianchi and Lugosi, 2006). Online learning was infused with algorithmic techniques from mathematical optimization into the setting of online convex optimization, see (Hazan, 2019) for a comprehensive introduction. Learnability in the statistical and online learning settings was characterized using various notions of di-

mensionality, starting from the VC-dimension, fat-shattering dimension, Rademacher complexity, Littlestone dimension and more. For an extensive treatment see (Mohri et al., 2018; Shalev-Shwartz and Ben-David, 2014; Vapnik, 1999). Regarding interpolation, Bubeck and Sellke (2021) establish an inverse relationship between the interpolation ability and robustness of a function class. The notion of interpolation dimension that we introduce here has found applications in the theory of boosting (Alon et al., 2021).

Deep control. Deep neural networks have advanced the state of the art for continuous control, not only in simulated environments Tassa et al. (2018); Zhang et al. (2016); Duan et al. (2016), but also in real-world tasks such as robotic manipulation OpenAI et al. (2018, 2019) and temperature control in office buildings and datacenters Wang et al. (2017); Lazic et al. (2018). In many of these applications, the policy learning procedure is episodic, where the environment resets at the beginning of an episode. For example, OpenAI et al. (2018, 2019) train an LSTM policy for manipulating a rubik's cube with a robotic hand in the following manner: an environment is generated at the beginning of the episode, which interacts with the current policy for a fixed number of time steps; then, after collecting the episodic trajectory, the policy is updated according to a chosen optimization scheme. This setting is closely related to online episodic control, which we formally describe in Section 2, and motivates our theoretical analysis of neural network-based policies in this framework.

2. Problem Setting and Preliminaries

Notation. Let $\|\cdot\|$ denote the Euclidean norm and $\langle\cdot,\cdot\rangle$ the corresponding inner product between two vectors, matrices, or tensors of the same dimension: $\langle x, y\rangle = \mathrm{vec}(x)^\top \mathrm{vec}(y)$. Let $\mathbb{S}_p = \{x \in \mathbb{R}^p : \|x\| = 1\}$ denote the unit p-dimensional sphere, and for a convex set $\mathcal{K}$, let $\prod_{\mathcal{K}}$ denote projection onto $\mathcal{K}$.

2.1. Deep neural networks and the interpolation dimension

Deep neural networks. Let $x \in \mathbb{R}^p$ be the p-dimensional input. We define the depth H network with ReLU activation and scalar output as follows:

$$x^0 = Ax, \quad x^h = \sigma_{\mathrm{relu}}(\theta^h x^{h-1}),\ h \in [H], \quad f(\theta, x) = a^\top x^H,$$

where $\sigma_{\mathrm{relu}}(\cdot)$ is the ReLU function $\sigma_{\mathrm{relu}}(z) = \max(0, z)$, $A \in \mathbb{R}^{m\times p}$, $\theta^h \in \mathbb{R}^{m\times m}$, and $a \in \mathbb{R}^m$. Let $\theta = (\theta^1, \ldots, \theta^H)^\top \in \mathbb{R}^{H\times m\times m}$ denote the trainable parameters of the network and the parameters A, a are fixed after initialization. The initialization scheme is as follows: each entry in A and θ^h is drawn i.i.d. from the Gaussian distribution $\mathcal{N}(0, \frac{2}{m})$, and each entry in a is drawn i.i.d. from $\mathcal{N}(0, 1)$. This setup is common in recent literature and follows that of Gao et al. (2019).

For vector-valued outputs, we consider a scalar output network for each coordinate. Suppose for $i \in [d]$, f_i is a deep neural network with a scalar output; with a slight abuse of notation, for input $x \in \mathbb{R}^p$, denote

$$f(\theta; x) = (f_1(\theta[1]; x), \ldots, f_d(\theta[d]; x))^\top \in \mathbb{R}^d, \tag{2.1}$$

where $\theta[i] \in \mathbb{R}^{H\times m\times m}$ denotes the trainable parameters for the network f_i for coordinate i. Let $\theta = (\theta[1], \theta[2], \ldots, \theta[d]) \in \mathbb{R}^{d\times H\times m\times m}$ denote all the parameters for f.

In the online setting, the neural net receives an input $x_t \in \mathbb{R}^p$ at each round $t \in [T]$, and with parameter θ suffers loss $\ell_t(f(\theta; x_t))$. Note that this framework generalizes the supervised learning paradigm. We make the following standard assumptions:

Assumption 1 *The input x has unit norm, i.e. $x \in \mathbb{S}_p$, $\|x\|_2 = 1$.*

Assumption 2 *The loss functions $\ell_t(f(\theta; x))$ are L-Lipschitz and convex in $f(\theta; x)$.*

Interpolation dimension. Since we aim to prove regret bounds for online learning with families of deep neural networks as the comparator class, we need to ensure that they have non-trivial representation power. To this end, we introduce interpolation dimension, an expressivity metric that can be naturally applied to our setting. In real-valued learning, we say that a hypothesis class has interpolation dimension of at least k if one can assign arbitrary real labels to *any* k different inputs using a hypothesis from that class.

Definition 1 *The* **interpolation dimension** *of a hypothesis class $\mathcal{H} = \{h : \mathcal{X} \subseteq \mathbb{R}^p \to \mathbb{R}^d\}$ over input domain $\mathcal{X}$ at non-degeneracy $\gamma > 0$, denoted $\mathcal{I}_{\mathcal{X},\gamma}(\mathcal{H})$, is the largest cardinality k such that for* **any** *set of data points $\{(x_j, y_j)\}_{j=1}^k$ satisfying $\min_{j \neq l} \|x_j - x_l\|_2 \geq \gamma$, $y_j \in [-1, 1]^d$, $\forall j \in [k]$, $\inf_{h \in \mathcal{H}} \left[\sum_{j=1}^k \|y_j - h(x_j)\|^2\right] = 0$.*

The label bound above is 1 for simplicity, but can be extended to any $B > 0$. Henceforth, we show that over input domain $\mathcal{X} = \mathbb{S}_p$, neural networks that have $\text{poly}(k, \frac{1}{\gamma})$ width have $\mathcal{I}_{\mathcal{X},\gamma}(\mathcal{H}) \geq k$. This enables us to derive regret bounds for online agnostic learning over a class of neural networks that has interpolation dimension at least k.

In the case of binary classification, interpolation dimension can be seen as the "dual" of the VC dimension. More details on the interpolation dimension in binary classification, connection to VC dimension, and additional examples can be found in Appendix A.1.

2.2. Online convex optimization

In Online Convex Optimization (OCO), a decision maker sequentially chooses a point in a convex set $\theta_t \in \mathcal{K} \subseteq \mathbb{R}^d$, and suffers loss $\ell_t(\theta_t)$ according to a convex loss function $\ell_t : \mathcal{K} \mapsto \mathbb{R}$. The goal of the learner is to minimize her regret, defined as

$$\text{Regret}_T = \sum_{t=1}^{T} \ell_t(\theta_t) - \min_{\theta^* \in \mathcal{K}} \sum_{t=1}^{T} \ell_t(\theta^*) \ .$$

A host of techniques from classical optimization are applicable to this setting and give rise to efficient low-regret algorithms. To name a few methods, mirror descent, Newton's method, Frank-Wolfe and follow-the-perturbed leader all have online analogues, see e.g. Hazan (2019) for a comprehensive treatment.

As an extension to the OCO framework, we show that regret bounds hold analogously for the online optimization of *nearly* convex functions. As we show in later sections, these regret bounds naturally carry over to the setting of online learning over neural networks.

Definition 2 *A function $\ell : \mathbb{R}^n \to \mathbb{R}$ is ε-nearly convex over the convex, compact set $\mathcal{K} \subseteq \mathbb{R}^n$ iff $\forall x, y \in \mathcal{K}$, $\ell(x) \geq \ell(y) + \nabla \ell(y)^\top (x - y) - \varepsilon$.*

The analysis of any algorithm for OCO, including the most fundamental method of online gradient descent (OGD), extends to this case in a straightforward manner. Let $\mathcal{A}$ be any regret minimization algorithm for OCO with a regret bound given by $\text{Regret}_T(\mathcal{A})$. This algorithm $\mathcal{A}$ can be applied on the surrogate loss functions $h_t(\theta) = \ell_t(\theta_t) + \nabla\ell_t(\theta_t)^\top(\theta - \theta_t)$ to obtain regret bounds on the nearly convex losses ℓ_t as given below. The described method is presented in Algorithm 3 which along with more details can be found in Appendix A.2.

Lemma 3 *Suppose $\ell_1, \ldots, \ell_T$ are ε-nearly convex, then Algorithm 3 has regret bounded by*

$$\sum_{t=1}^{T} \ell_t(\theta_t) - \min_{\theta^* \in \mathcal{K}} \sum_{t=1}^{T} \ell_t(\theta^*) \leq \textit{Regret}_T(\mathcal{A}) + \varepsilon T \ .$$

2.3. Online episodic control

Consider the following online episodic learning problem for nonstochastic control over linear time-varying (LTV) dynamics: there is a sequence of T control problems each with a horizon K and an initial state $x_1 \in \mathbb{R}^{d_x}$. In each episode, the state transition is given by

$$\forall k \in [1, K], \quad x_{k+1} = A_k x_k + B_k u_k + w_k, \tag{2.2}$$

where $x_k \in \mathbb{R}^{d_x}, u_k \in \mathbb{R}^{d_u}$. The system matrices $A_k \in \mathbb{R}^{d_x \times d_x}, B_k \in \mathbb{R}^{d_x \times d_u}$ along with the next state x_{k+1} are revealed to the learner *after* taking the action u_k. The disturbances $w_k \in \mathbb{R}^{d_x}$ are unknown and adversarial but can be a posteriori computed by the learner $w_k = x_{k+1} - A_k x_k - B_k u_k$. An episode loss is defined cumulatively over the rounds $k \in [1, K]$ according to the convex cost functions $c_k : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \to \mathbb{R}$ of state and action: for a policy π, the loss is $J(\pi; x_1, c_{1:K}) = \sum_{k=1}^{K} c_k(x_k^\pi, u_k^\pi)$. Like the system matrices, the cost function c_k is also revealed after taking action u_k. The transition matrices $(A_k, B_k)_{1:K}$, initial state x_1, disturbances $w_{1:K}$ and costs $c_{1:K}$ can *change arbitrarily* over different episodes. The goal of the learner is to minimize *episodic* regret by adapting its output policies π_t for $t \in [1, T]$,

$$\text{Regret}_T(\Pi) = \sum_{t=1}^{T} J_t(\pi_t; x_1^t, c_{1:K}^t) - \min_{\pi \in \Pi} \sum_{t=1}^{T} J_t(\pi; x_1^t, c_{1:K}^t), \tag{2.3}$$

where Π denotes the class of policies the learner competes against.

The model above is presented in its utmost generality: the system in an episode is LTV and these LTVs are allowed to change arbitrarily throughout episodes. Results for this model can be applied to derive guarantees for: (1) a simpler setting, learning to control a single LTV episodically; (2) a more complex setting, first-order guarantees in control or planning over *nonlinear* dynamics by taking the Jacobian linearization of the dynamics (Ahn et al., 2007; Westenbroek et al., 2021; Roulet et al., 2022). We make the following basic assumptions about the dynamical system *in each episode* that are common in the nonstochastic control literature (Agarwal et al., 2019).

Assumption 3 *The disturbances satisfy $\forall k \in [K]$, $\|w_k\|_2 \leq W$.*

Assumption 4 (Sequential stability) [1] *There exist $C_1, C_2 \geq 1$, $0 < \rho_1 < 1$ such that the system matrices satisfy:*

$$\forall k \in [K], \forall n \in [1, k), \quad \left\| \prod_{i=k}^{k-n+1} A_i \right\|_{op} \leq C_1 \cdot \rho_1^n, \quad \|B_k\|_{op} \leq C_2 \ .$$

Assumption 5 *Each cost function $c_k : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \to \mathbb{R}$ is jointly convex and satisfies a generalized Lipschitz condition $\|\nabla c_k(x, u)\| \leq L_c \max\{1, \|x\| + \|u\|\}$ for some $L_c > 0$.*

The performance of the learner given by (2.3) directly depends on the policy class Π. In this work, we focus on disturbance based policies, i.e. policies that take past perturbations as input $u_k = f(w_{1:k-1})$, which are parameterized w.r.t. *policy-independent* inputs. This is in contrast to the commonly used state feedback policy $u_k = f(x_k)$. For example, the Disturbance Action Control (DAC) policy class, shown to be more general than linear state feedback policies (Agarwal et al., 2019), outputs controls linear in past finite disturbances, resulting in a *convex* parameterization of the state/control and enabling the design of efficient provable online methods. Our work expands the comparator class by considering policies that are *nonlinear* in the past disturbances, represented by neural networks.

Definition 4 (Disturbance Neural Feedback Control) *Let π^{θ}_{dnn} denote the policy with control outputs u_k given by*

$$\forall k \in [K], \quad u_k = f_\theta(w_{k-1}, w_{k-2}, \ldots, w_1) \in \mathbb{R}^{d_u},$$

where $f_\theta(\cdot) = f(\theta; \cdot)$ is a neural network defined in (2.1). The policy class is defined as $\Pi_{dnn}(f; \Theta) = \{\pi^{\theta}_{dnn} : \theta \in \Theta\}$ with Θ being the set of permissible parameters.

3. Online learning of deep neural networks

In this section, we consider the general framework of online learning with deep neural networks and state the accompanying regret guarantees. Our framework can use any OCO algorithm as a black-box as in Algorithm 1, but for our main result, we use projected Online Gradient Descent (OGD). Projected OGD has explicit regret bounds and variants of GD are widely used in practice. Observe that, in this case, the parameter update is equivalent to OGD on the original losses.

Algorithm 1 Online Learning over Neural Networks

Input: OCO algorithm $\mathcal{A}$, neural network $f(\cdot;\cdot)$, initial θ_1, parameter set $\Theta = B(R; \theta_1)$.
for $t = 1 \ldots T$ **do**
  Play θ_t, receive loss $\ell_t(\theta) = \ell_t(f(\theta; x_t))$.
  Construct $h_t(\theta) = \ell_t(\theta_t) + \nabla \ell_t(\theta_t)^\top (\theta - \theta_t)$.
  Update $\theta_{t+1} = \mathcal{A}(h_1, \ldots, h_t) \in \Theta$.
end

1. This condition is relaxed to sequential stabilizability in Appendix E

The main technical result, provided in Theorem 5, gives a regret bound on the online agnostic learning of deep neural networks. The benchmark hypothesis class is a class of deep neural networks with interpolation dimension of at least k where k is decided a priori and used in the construction of the network.

Theorem 5 *Suppose Assumptions 1 and 2 hold, and let* $\mathcal{H}_{\mathrm{NN}}(R;\theta_1) = \{f(\theta;\cdot) : \theta \in \Theta\}$ *denote the class of neural networks* $f(\theta;\cdot)$ *as in (2.1) with parameter set* $\Theta = B(R;\theta_1) = \{\theta : \|\theta[i] - \theta_1[i]\|_F \le R, \forall\ i \in [d]\}$ *and* $\mathcal{X} = \mathbb{S}_p$. *Suppose* $\gamma \in (0, O\left(\frac{1}{H}\right)]$, *take* $R = O\left(\frac{k^3 \log m}{\gamma\sqrt{m}}\right)$, *then for* $m \ge O(\frac{p^{3/2}(k^{24}H^{12}\log^8 m + d)^{3/2}}{\gamma^8})$, *with probability* $1 - O(H+d)e^{-\Omega(\log^2 m)}$ *over the random initialization,*

- *The function class* $\mathcal{H}_{\mathrm{NN}}(R;\theta_1)$ *has interpolation dimension* $\mathcal{I}_{\mathcal{X},\gamma}(\mathcal{H}_{\mathrm{NN}}(R;\theta_1)) \ \ge\ k$.
- *Algorithm 1 using OGD with* $\eta_t = \frac{2R\sqrt{d}}{LH\sqrt{m}} \cdot t^{-1/2}$ *for* $\mathcal{A}$ *attains regret bound*

$$\sum_{t=1}^{T} \ell_t(f(\theta_t; x_t)) \le \min_{g \in \mathcal{H}_{\mathrm{NN}}(R;\theta_1)} \sum_{t=1}^{T} \ell_t(g(x_t)) + \tilde{O}\left(\frac{k^3 LH\sqrt{dT}}{\gamma} + \frac{k^4 LH^{5/2}\sqrt{d}T}{\gamma^{4/3} m^{1/6}}\right),$$

  where $\tilde{O}(\cdot)$ *hides terms polylogarithmic in* m.

The above theorem indicates that the average regret can be minimized up to arbitrary precision: for any $\varepsilon > 0$, if one chooses sufficiently large network width $m = \Omega(\varepsilon^{-6})$ and sufficiently large number of iterations $T = \Omega(\varepsilon^{-2})$, the average regret is bounded by ε. The interpolation dimension bound is established due to the seminal work Allen-Zhu et al. (2019), spelled out in the following lemma and proven in Appendix A.1.

Lemma 6 *Let* $\mathcal{H}_{\mathrm{NN}}(R;\theta_1) = \{f(\theta;\cdot) : \theta \in \Theta\}$ *denote the class of neural networks as in (2.1) where* $\Theta = B(R;\theta_1)$ *and* $\mathcal{X} = \mathbb{S}_p$. *Suppose* $\gamma \in \left(0, O(\frac{1}{H})\right]$, $m \ \ge\ \Omega\left(\frac{k^{24}H^{12}\log^5 m}{\gamma^8}\right)$ *and* $R = O\left(\frac{k^3 \log m}{\gamma\sqrt{m}}\right)$, *then with probability* $1 - d \cdot e^{-\Omega(\log^2 m)}$ *over random initialization of* θ_1,

$$\mathcal{I}_{\mathcal{X},\gamma}\left(\mathcal{H}_{\mathrm{NN}}(R;\theta_1)\right) \ \ge\ k\ . \tag{3.1}$$

3.1. Proof Sketch

Due to space constraints, we give a proof sketch here; for a more detailed analysis outline, see Appendix C, and for the full proof see Appendix D. There are 3 steps to the proof of Theorem 5. First, we show that the considered loss functions $\ell_t : \Theta \to \mathbb{R}$, $\ell_t(\theta) = \ell_t(f(\theta; x_t))$ are *nearly convex* with respect to the parameter θ. This is due to the observation that in the overparameterized regime, neural networks behave similarly to their local linearization.

Second, we can use the near convexity of the loss functions $\ell_t(\theta)$ for all $\theta \in B(R;\theta_1)$, and Lemma 3 to show a regret bound over the parameter set $\Theta = B(R;\theta_1)$. The bound is comprised of the sublinear regret of the OCO algorithm used for parameter update, and the worst-case linear penalty of near convexity $\varepsilon_{\mathrm{nc}} \cdot T$, where $\varepsilon_{\mathrm{nc}}$ is in terms of R and m.

Finally, we use Lemma 6 to ensure that our choice of R and m give the desired interpolation dimension, and derive the final regret guarantee in terms of k, γ and m, among other parameters.

4. Online episodic control with neural network controllers

The online episodic control problem described in Section 2.3 with the policy class $\Pi = \Pi_{\text{dnn}}(f;\Theta)$ can be reduced to online learning for neural networks. This reduction is done by following the current policy each episode, constructing the episode loss, and updating the policy via an OCO algorithm. Algorithm 2 below uses projected OGD but as in the previous section, any OCO algorithm can be used instead.

Algorithm 2 Deep Neural Network Episodic Control with OGD

Input: stepsize $\eta_t > 0$, initial parameter θ_1, parameter set $\Theta = B(R;\theta_1)$.
for $t = 1 \dots T$ **do**
  for $k = 1 \dots K$ **do**
    Construct $z_k^t = \text{vec}([w_{k-1}^t, \dots, w_1^t, \mathbf{0}, \dots, \mathbf{0}, k]) \in \mathbb{R}^{K \cdot d_x + 1}$ and normalize $\bar{z}_k^t = \frac{z_k^t}{\|z_k^t\|}$.
    Observe x_k^t and play $u_k^t = f(\theta_t, \bar{z}_k^t)$.
  end
  Construct loss function $\mathcal{L}_t(\theta) = \sum_{k=1}^K c_k^t(x_k^{t,\theta}, f(\theta, \bar{z}_k^t))$.
  Perform gradient update $\theta_{t+1} = \prod_\Theta [\theta_t - \eta_t \nabla_\theta \mathcal{L}_t(\theta_t)]$.
end

Theorem 7 *Suppose Assumptions 3, 4, 5 hold and let $\Pi_{dnn}(f;\Theta)$ denote the policy class given by Definition 4 with $\Theta = B(R;\theta_1)$. Take $R = O\left(\frac{K^3(2KW+H)\log m}{\sqrt{m}}\right)$, then for $m \geq \Omega(K^{46}H^{20}W^8(d_x d_u)^{3/2}\log^{12} m)$ with probability at least $1 - O(H + d_u)e^{-\Omega(\log^2 m)}$ over the randomness of initialization θ_1, Algorithm 2 with $\eta_t = O(\frac{R\sqrt{d_u}}{LH\sqrt{m}}t^{-1/2})$ satisfies*

$$\textit{Regret}_T(\Pi_{dnn}(f;\Theta)) \leq \tilde{O}\left(K^{10}L_c H^4 W^2 d_u d_x^{1/2} \cdot \sqrt{T} + \frac{K^{12}L_c H^6 W^3 d_u d_x^{1/2}}{m^{1/6}} \cdot T\right),$$

where $\Pi_{dnn}(f;\Theta)$ can output the optimal open-loop control sequence $u_{1:K}^\star \in [-1,1]^{K\times d_u}$ of any episode and $\tilde{O}(\cdot)$ hides terms polylogarithmic in m.

This theorem statement, analogous to Theorem 5, implies that arbitrarily small $\varepsilon > 0$ average episodic regret is attained with a large network width $m = \Omega(\varepsilon^{-6})$ and large number of iterations $T = \Omega(\varepsilon^{-2})$. The regret bound is against the benchmark policy class $\Pi_{\text{dnn}}(f;\Theta)$ which is chosen such that the neural network class has interpolation dimension $k = K$. This implies open-loop control optimality over a single episode in the following way. For simplicity, drop the episode index $t \in [T]$ and define the optimal *open-loop* control sequence of an episode.

Definition 8 *Define the optimal open-loop control sequence $u_{1:K}^\star \in [-1,1]^{K\times d_u}$ to be*

$$u_{1:K}^\star = \underset{\forall k, u_k \in [-1,1]^{d_u}}{\arg\min} \left\{ J(u_{1:K}; x_1, c_{1:K}) = \sum_{k=1}^K c_k(x_k, u_k) \right\}.$$

To demonstrate the capacity of the benchmark policy class $\Pi_{\text{dnn}}(f;\Theta)$ with $\Theta = B(R;\theta_1)$ we show that it can output the *optimal* open-loop control sequence of any *single* episode as detailed below.

Lemma 9 *Take* $R = O\left(\frac{K^3 \log m(2KW+H)}{\sqrt{m}}\right)$, *suppose* $m \geq \Omega\left(K^{24}H^{12}\log^5 m(2KW+H)^8\right)$, *then with probability* $1 - d_u \cdot e^{-\Omega(\log^2 m)}$ *over the random initialization of* θ_1, $\Pi_{dnn}(f;\Theta)$ *can output any open-loop control sequence* $u^\star_{1:K} \in [-1,1]^{K\times d_u}$*:*

$$\inf_{\pi^\theta_{dnn}\in\Pi_{dnn}(f;\Theta)} \left[\sum_{k=1}^{K} \|u^\theta_k - u^\star_k\|^2\right] = 0 \ .$$

4.1. Proof Sketch

To extend the online learning results of Theorem 5 to the online episodic control setting, we ensure the control setting satisfies the corresponding assumptions. For each $k \in [K]$, denote the padded input $z_k = \text{vec}([w_{k-1},\ldots,w_1,\mathbf{0},\ldots,\mathbf{0},k]) \in \mathbb{R}^{K\cdot d_x+1}$ where the index is padded to ensure inputs are separable (Definition 1). To satisfy Assumption 1, normalize the network inputs $\bar{z}_k = \frac{z_k}{\|z_k\|_2} \in \mathbb{S}_{K\cdot d_x+1}$.

For a policy π^θ_{dnn} the episode loss $\mathcal{L}(\theta) = J(\pi^\theta_{\text{dnn}}; x_1, c_{1:K})$ depends on the parameter θ through all the K controls $u^\theta_k = f(\theta;\bar{z}_k)$. Denote $\bar{f}(\theta) = [u^\theta_1,\ldots,u^\theta_K]^\top \in \mathbb{R}^{K\times d_u}$ and let $\mathcal{L}(\theta) = \mathcal{L}(\bar{f}(\theta))$ by abuse of notation. We demonstrate that the reduction to the online learning setting is achieved by showing that $\mathcal{L}(\bar{f}(\theta))$ satisfies the convexity (Lemma 23) and Lipschitz (Lemma 26) conditions. Hence, for each episode $t \in [T]$, the episode loss $\mathcal{L}_t(\theta) = J_t(\pi^\theta_{\text{dnn}}; x^t_1, c^t_{1:K})$ satisfies Assumption 2 and the rest of the derivation is analogous to that of Theorem 5. Finally, Lemma 9 uses the interpolation dimension property of the neural network class to conclude the open-loop optimality stated in the theorem. See Appendix E for full details.

5. Conclusions and Future Work

In this work, we derive the first regret guarantees for neural network based controllers in online control. Our results are in the online episodic control setting, which is motivated by empirical research in control and deep reinforcement learning. We propose algorithms that obtain sublinear episodic regret against the optimal open-loop control sequence of any episode, which relies on a general reduction from online deep learning to regret minimization.

We also introduce a new metric for the expressive power of a hypothesis class and use it for characterizing the expressivity of the benchmark neural network class. The definition of interpolation dimension enables this characterization to be isolated to neural networks but is in no way specific to them. Many intriguing questions about this expressivity notion remain open, such as the existence of an analogue to statistical learning theory, given its close relationship to the VC dimension.

We use the NTK deep learning theory paradigm to derive the control and online learning results in this work. However, there are still remaining open questions to understand the empirical success of neural networks. With such theoretical developments, the question of extending them to reinforcement learning and control problems remains open too.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119, 2019.

Hyo-Sung Ahn, YangQuan Chen, and Kevin L. Moore. Iterative learning control: Brief survey and categorization. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(6):1099–1121, 2007. doi: 10.1109/TSMCC.2007.905759.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization, 2019.

Noga Alon, Alon Gonen, Elad Hazan, and Shay Moran. Boosting simple learners. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 481–489, 2021.

Raman Arora, Sanjeev Arora, Joan Bruna, Nadav Cohen, Simon Du, Rong Ge, Suriya Gunasekar, Chi Jin, Jason Lee, Tengyu Ma, and Behnam Neyshabur. *Theory of Deep Learning*. 2021.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

Yu Bai and Jason D Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*, 2019.

Sebastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=z71OSKqTFh7.

Tianle Cai, Ruiqi Gao, Jikai Hou, Siyu Chen, Dong Wang, Di He, Zhihua Zhang, and Liwei Wang. Gram-gauss-newton method: Learning overparameterized neural networks for regression problems. *arXiv preprint arXiv:1905.11675*, 2019.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in Neural Information Processing Systems*, 32: 10836–10846, 2019.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1300–1309, 2019.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197, 2018.

Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018a.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018b.

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1329–1338. JMLR.org, 2016.

Ruiqi Gao, Tianle Cai, Haochuan Li, Cho-Jui Hsieh, Liwei Wang, and Jason D Lee. Convergence of adversarial training in overparametrized neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/file/348a38cd25abeab0e440f37510e9b1fa-Paper.pdf.

Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Elad Hazan and Karan Singh. Tutorial: online and non-stochastic control, July 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks, 2020.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15312–15325. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/aee5620fa0432e528275b8668581d9a8-Paper.pdf.

Nevena Lazic, Tyler Lu, Craig Boutilier, MK Ryu, Eehern Jay Wong, Binz Roy, and Greg Imwalle. Data center cooling using model-predictive control. In *Proceedings of the Thirty-second Conference on Neural Information Processing Systems (NeurIPS-18)*, pages 3818–3827, Montreal, QC, 2018. URL https://papers.nips.cc/paper/7638-data-center-cooling-using-model-predictive-control.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir. The connection between approximation, depth separation and learnability in neural networks. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3265–3295. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/malach21a.html.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, pages 10154–10164, 2019.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Kevin L Moore. *Iterative learning control for deterministic systems*. Springer Science & Business Media, 2012.

OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation, 2018. URL https://arxiv.org/abs/1808.00177.

OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving rubik's cube with a robot hand, 2019. URL https://arxiv.org/abs/1910.07113.

Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 555–561, 2008. doi: 10.1109/ALLERTON.2008.4797607.

Vincent Roulet, Siddhartha Srinivasa, Maryam Fazel, and Zaid Harchaoui. Complexity bounds of iterative linear quadratic optimization algorithms for discrete time nonlinear control. *arXiv preprint arXiv:2204.02322*, 2022.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 2018.

Y. Tassa, T. Erez, and E. Todorov. Synthesis and stabilization of complex behaviors through online trajectory optimization. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4906–4913, 2012.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. Deepmind control suite, 2018. URL https://arxiv.org/abs/1801.00690.

Emanuel Todorov and Weiwei Li. A generalized iterative lqg method for locally-optimal feedback control of constrained nonlinear stochastic systems. In *Proceedings of the 2005, American Control Conference, 2005.*, pages 300–306. IEEE, 2005.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

Yuan Wang, Kirubakaran Velswamy, and Biao Huang. A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems. *Processes*, 5(3), 2017. ISSN 2227-9717. doi: 10.3390/pr5030046. URL https://www.mdpi.com/2227-9717/5/3/46.

Colin Wei, Jason Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. 2019.

Tyler Westenbroek, Max Simchowitz, Michael I Jordan, and S Shankar Sastry. On the stability of nonlinear receding horizon control: a geometric perspective. *arXiv preprint arXiv:2103.15010*, 2021.

Xiaoxia Wu, Simon S Du, and Rachel Ward. Global convergence of adaptive gradient methods for an over-parameterized neural network. *arXiv preprint arXiv:1902.07111*, 2019.

Xiaoxia Wu, Yuege Xie, Simon Du, and Rachel Ward. Adaloss: A computationally-efficient and provably convergent adaptive gradient method. *arXiv preprint arXiv:2109.08282*, 2021.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/5481b2f34a74e427a2818014b8e103b0-Paper.pdf.

Guodong Zhang, James Martens, and Roger B Grosse. Fast convergence of natural gradient descent for over-parameterized neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/file/1da546f25222c1ee710cf7e2f7a3ff0c-Paper.pdf.

Tianhao Zhang, Gregory Kahn, Sergey Levine, and P. Abbeel. Learning deep control policies for autonomous aerial vehicles with mpc-guided policy search. *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 528–535, 2016.

Yi Zhang, Orestis Plevrakis, Simon S. Du, Xingguo Li, Zhao Song, and Sanjeev Arora. Over-parameterized adversarial training: An analysis overcoming the curse of dimensionality, 2020.