Low Complexity Online Contextual Learning with Continuous Actions

Mohsen Najjar
Tel Aviv University
Tel Aviv, Israel
mohsennajjar@mail.tau.ac.il

Tavor Baharav
Broad Institute
Cambridge, MA, USA
baharav@broadinstitute.org

Ilai Bistritz
Tel Aviv University
Tel Aviv, Israel
ilaibistritz@tauex.tau.ac.il

ABSTRACT

We study an online contextual learning problem, where an agent repeatedly observes independent and identically distributed (IID) contexts $c_t \in \mathbb{R}^d$ and selects actions $x_t \in \mathbb{R}^k$ to maximize its cumulative reward $r(x_t, c_t)$ over T rounds. The reward function is Lipschitz continuous in contexts, so good actions for a given context are also reasonably good for similar contexts. Current algorithms that leverage this structure are practically infeasible due to large runtime or memory complexity. In this paper, we propose Congrad, a simple kernel-based projected gradient ascent algorithm, which maintains $O(n)$ memory and $O(n(k+d))$ computational complexity per iteration by projecting policies onto a fixed n-dimensional function space. Congrad utilizes a kernel that at each turn updates the actions for contexts c near the observed c_t. The kernel initially has a large bandwidth to enable fast global learning, and progressively narrows for local refinement. We prove an expected regret bound of $O(T^{\frac{d+1}{d+2}})$, independent of the action space dimension k.

KEYWORDS

Contextual learning, Stochastic optimization

ACM Reference Format:
Mohsen Najjar, Tavor Baharav, and Ilai Bistritz. 2026. Low Complexity Online Contextual Learning with Continuous Actions. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/

1 INTRODUCTION

For many applications, the Markov Decision Process (MDP) framework considered by reinforcement learning is unnecessarily general, as the problem is more structured. One common structured case is contextual learning [23]. In contextual learning, there is a state in every round, called a context, that the agent can observe before taking its action, which affects its reward. However, the agent's actions do not affect this context. This models many applications, such as recommendation systems [3, 8, 14, 16] where the context is the features of the customer under consideration, cyber-physical systems [5, 9, 10, 13, 21] where the context can be a physical state of the environment (e.g. weather, demand, or traffic), or healthcare [6, 7, 24, 26] where the context represents the patient features.

This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/

Most works on contextual learning have focused on discrete action spaces [1, 2, 11, 12, 15, 22]. While discrete action spaces are closer to the classical MDP model, they cannot capture scenarios where the action is a location, velocity, probability vector, or continuous model parameter. Discrete actions also preclude the use of classical and efficient optimization tools, such as gradient descent.

With a discrete context space, a naive contextual learning approach would be to run an independent algorithm for every context when it appears. With a continuous context space, the natural extension of this scheme is to divide the context space into small enough balls to approximate the optimal policy [17]. However, such an algorithm is inefficient, as it does not exploit the similarity of the rewards when the contexts are similar; the optimal action must be similar for adjacent balls. Following the literature [4, 18, 19], we encode the similarity between contexts by assuming that the reward function and its gradients are Lipschitz continuous in c.

State-of-the-art contextual learning algorithms [18, 25] are based on splitting the context–action space adaptively (or an underlying metric space), selecting the best action within each refined region. These algorithms do not assume concave reward functions, and thus must search directly over the high-dimensional action space, yielding regret that inherently depends on the metric (or zooming) dimension of the action space. The computational complexity depends on T as the number of regions increases with T. Instead, we consider concave reward functions, which are natural in control and optimization and allow for gradient-based algorithms that inherently have lower complexity than exhaustive search in high dimensions. As with stochastic gradient descent, we only assume that a noisy, unbiased gradient estimate is available at each turn.

In contextual learning [4, 19], gradient-based methods in concave settings achieve regret bounds where the exponent of T is independent of the action space dimension k. However, the multiplicative factor of this T-dependent term still scales no better than linearly with k, as the number of regions ("bins") grows with k, rendering such methods impractical for even moderate dimensionality. Our algorithm achieves a regret bound independent of k, making it appealing for high-dimensional action spaces common in practice.

More importantly, state-of-the-art gradient-based contextual learning algorithms [4, 19] suffer from a linear memory complexity in T. This memory explosion results from storing information from all previous time steps, making them infeasible when T is large.

To avoid exploding memory, we consider learning policies from a function space. Instead of storing all past information, we project the learned policy onto a fixed set of n orthogonal basis functions at each iteration. Consequently, the memory complexity of our algorithm is $O(n)$, independent of T. Computing gradient steps and projecting our learned policy results in a computational complexity of $O(n(k+d))$ per iteration.

2 PROBLEM FORMULATION

Consider an agent that takes actions in discrete time, receiving a stochastic context in each round. Let $C \subset \mathbb{R}^d$ be the compact context set and $\mathcal{X} = \mathbb{R}^k$ the action set. The payoff function $r : \mathcal{X} \times C \to \mathbb{R}$ is assumed to be concave in x.

In each round t, a context $c_t \in C$ is drawn independently from a fixed distribution D_c on C, independent of past contexts and the agent's actions. The probability density function of D_c is denoted by p_c. The agent selects an action $y_t \in \mathcal{X}$ and observes an unbiased estimate of the gradient:

$$\nabla_x r(y_t, c_t) + \epsilon_t \tag{1}$$

where ϵ_t is the IID noise in the gradient at turn t, modeling data-dependent gradients and estimation or measurement errors.

Our goal is to design an algorithm that learns an optimal policy $\pi^* \in \operatorname{argmax}_{\pi \in \Pi} \mathbb{E}\big[r(\pi(c), c)\big]$ from a predesigned function space Π, where the expectation is with respect to p_c. The expected regret measures how efficiently an algorithm learns the optimal policy:

$$\sum_{t=1}^{T} \mathbb{E}\big[r(\pi^*(c_t), c_t) - r(\pi_t(c_t), c_t)\big], \tag{2}$$

where the expectation is over $(c_1, \ldots, c_T)$ and the noise ϵ_t.

Our function space is defined by the orthogonal basis functions $\phi_1, \ldots, \phi_n : C \to \mathbb{R}^k$ and a bound on the coefficients M:

$$\Pi = \left\{ \pi(\cdot) = \sum_{i=1}^{n} \gamma_i \phi_i(\cdot) \;\middle|\; \|\gamma\|_\infty \le M \right\} \subset L^2(C, \mu; \mathbb{R}^k), \tag{3}$$

where $L^2(C, \mu; \mathbb{R}^k)$ is the space of square-integrable $\mathbb{R}^k$-valued functions on C with respect to the uniform (normalized Lebesgue) probability measure μ on C.

The effective action set $\mathcal{X}_\Pi := \{\pi(c) \mid \pi \in \Pi,\ c \in C\}$ that the policy can implement is compact, since each $\pi \in \Pi$ is a finite linear combination of bounded continuous basis functions with bounded coefficients, and C is compact.

The following assumption is common for context distributions [14, 19, 20]. The context space can also be a finite union of compact components if this assumption holds for each component.

Assumption 1. *Each context $c_t \in C \subset \mathbb{R}^d$ is drawn i.i.d. from a distribution D_c with a density p_c that satisfies*

$$p_{min} \le p_c(c) \le p_{max}, \quad \forall c \in C, \quad 0 < p_{min} \le p_{max} < \infty.$$

The next assumptions are standard in stochastic optimization.

Assumption 2. *The noise satisfies $\mathbb{E}[\epsilon_t] = 0$ and $\mathbb{E}[\|\epsilon_t\|^2] < \infty$.*

Assumption 3. $\|\nabla_x r(x, c)\| \le G', \quad \forall x \in \mathcal{X}_\Pi,\ c \in C.$

Our assumptions about the reward function are given next.

Assumption 4. *For all $c \in C$, the function $r(\cdot, c)$ is concave.*

Assumption 5. *$r(x, \cdot)$, $\nabla_x r(x, \cdot)$ are Lipschitz continuous $\forall x \in \mathcal{X}_\Pi$.*

The following assumption on the basis $\{\phi_i\}$ of Π ensures that π_t remains Lipschitz continuous at every t. This assumption is satisfied, for example, if each ϕ_i is twice continuously differentiable. Classical basis families such as Fourier series and polynomial bases meet this condition, for the compact context set.

Assumption 6. *Each basis function $\phi_i : C \to \mathbb{R}^k$ is L'-Lipschitz continuous for some $L' < \infty$.*

Algorithm 1 Congrad: Contextual Gradient-Based Learning

Input: Basis functions $\{\phi_i\}_{i=1}^n$, bound $M > 0$, learning rates $\{\eta_t\}$, bandwidths $\{\lambda_t\}$, kernels $\{K_t\}$
Initialize: $\gamma_1 \in [-M, M]^n$ and $\pi_1(\cdot) = \sum_{i=1}^n \gamma_{i,1} \phi_i(\cdot)$
for $t = 1, 2, \ldots, T$ **do**
  Observe context $c_t \in C$
  Compute action $y_t = \pi_t(c_t) = \sum_{i=1}^n \gamma_{i,t}\, \phi_i(c_t)$
  Play y_t and receive noisy gradient $G_t := \nabla_x r(y_t, c_t) + \epsilon_t$
  Compute $\overline{K}_t(c_t) = \int_C K_t(c, c_t)\, dc$
  For $i = 1, \ldots, n$, update the policy by updating the coefficients:

$$\tilde{\gamma}_{i,t+1} := \gamma_{i,t} + \eta_t \int_C \frac{K_t(c, c_t)}{\overline{K}_t(c_t)} \langle G_t, \phi_i(c) \rangle\, dc$$

$$\gamma_{i,t+1} := \mathrm{Proj}_{[-M,M]}\big(\tilde{\gamma}_{i,t+1}\big)$$

end for

3 CONGRAD: CONTEXTUAL PROJECTED GRADIENT ASCENT

Congrad is described in Algorithm 1. Since Π is a linear subspace (up to the box constraints), the kernel-smoothed gradient step can be projected onto Π by updating the coefficients directly, thereby avoiding an explicit projection of the full updated policy. The resulting d-dimensional integral can be approximated using a quadrature or Monte Carlo rule, yielding a per-iteration complexity of $O(mnk)$ for m evaluation points when d is small.

We assume that our sequence of kernels K_t have the form

$$K_t(c, c_0) = \kappa\Big(\lambda_t^{-1} \|c - c_0\|\Big), \tag{4}$$

where $\kappa : [0, \infty) \to [0, 1]$ is a non-increasing function with compact support, so that K_t has compact support in the d-dimensional ball of radius λ_t; i.e., $\kappa(u) = 0$ for all $u > 1$. Consequently, $K_t(c, c_0) = 0$ if $\|c - c_0\| > \lambda_t$. An example is $K_t(c, c_0) = \max(0, 1 - \lambda_t^{-1}\|c - c_0\|)$.

Assumption 7 (Kernel Mass Condition)**.** *For all $t \ge 1$ and all $c_0 \in C$ up to a set of measure zero:*

$$c_d \lambda_t^d \le \int_C K_t(c, c_0)\, dc \le C_d \lambda_t^d, \tag{5}$$

for constants $C_d, c_d > 0$ depending only on the context dimension d.

Assumption 7 is needed since a kernel that is too narrow (i.e., assigning near-zero weight to most contexts) would hinder generalization and slow down learning.

The kernel at round t is normalized by $\overline{K}_t(c_t) = \int_C K_t(c, c_t)\, dc$ so that updates have comparable magnitude even for contexts near the boundary of C, where the unnormalized kernel mass is smaller.

Finally, we present our main theorem: an expected regret bound for Congrad that is independent of the action space dimension.

Theorem 1 (Congrad Regret Bound)**.** *Suppose Assumptions 1-7 hold. Let the step size and kernel bandwidth sequences be defined by $\lambda_t = t^{-\frac{1}{d+2}}$ and $\eta_t = t^{-\frac{d+1}{d+2}}$. Then, for every T, the expected regret satisfies*

$$\sum_{t=1}^{T} \mathbb{E}\left[r\big(\pi^*(c_t), c_t\big) - r\big(\pi_t(c_t), c_t\big)\right] \le B T^{\frac{d+1}{d+2}} \tag{6}$$

where B is a constant independent of k and T.

ACKNOWLEDGMENTS

This research was supported by the Koret Foundation grant for Smart Cities and Digital Living 2030. TZB was supported by the Eric and Wendy Schmidt Center at the Broad Institute.

REFERENCES

[1] Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert E Schapire. 2017. Optimal and efficient contextual bandit algorithms. In *COLT*.

[2] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. 2014. Taming the monster: A fast and simple algorithm for contextual bandits. In *International conference on machine learning*. PMLR, 1638–1646.

[3] Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. 2019. MNL-bandit: A dynamic learning approach to assortment selection. *Operations Research* 67, 5 (2019), 1453–1485.

[4] Arya Akhavan, Karim Lounici, Massimiliano Pontil, and Alexandre B Tsybakov. 2024. A conversion theorem and minimax optimality for continuum contextual bandits. *arXiv preprint arXiv:2406.05714* (2024).

[5] Jose A Ayala-Romero, Andres Garcia-Saavedra, and Xavier Costa-Perez. 2024. Risk-aware continuous control with neural contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 20930–20938.

[6] Tavor Z. Baharav, Spyros Dragazis, and Aldo Pacchiano. 2025. The Good, the Bad, and the Sampled: a No-Regret Approach to Safe Online Classification. *arXiv preprint arXiv:2510.01020* (2025). arXiv:2510.01020 https://arxiv.org/abs/2510.01020

[7] Hamsa Bastani and Mohsen Bayati. 2020. Online decision making with high-dimensional covariates. *Operations Research* 68, 1 (2020), 276–294.

[8] Hamsa Bastani, Pavithra Harsha, Georgia Perakis, and Divya Singhvi. 2018. Sequential learning of product recommendations with customer disengagement. *Available at SSRN* 3240970 (2018).

[9] Felix Berkenkamp, Andreas Krause, and Angela P Schoellig. 2023. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *Machine learning* 112, 10 (2023), 3713–3747.

[10] Marcello Fiducioso, Sebastian Curi, Benedikt Schumacher, Markus Gwerder, and Andreas Krause. 2019. Safe contextual Bayesian optimization for sustainable room temperature PID control tuning. *arXiv preprint arXiv:1906.12086* (2019).

[11] Dylan Foster and Alexander Rakhlin. 2020. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *International conference on machine learning*. PMLR, 3199–3210.

[12] Dylan J Foster and Akshay Krishnamurthy. 2021. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems* 34 (2021), 18907–18919.

[13] Apostolos Galanopoulos, Jose A Ayala-Romero, Douglas J Leith, and George Iosifidis. 2021. AutoML for video analytics with edge computing. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 1–10.

[14] Yonatan Gur, Ahmadreza Momeni, and Stefan Wager. 2022. Smoothness-adaptive contextual bandits. *Operations Research* 70, 6 (2022), 3198–3216.

[15] Yichun Hu, Nathan Kallus, and Xiaojie Mao. 2020. Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. In *Conference on Learning Theory*. PMLR, 2007–2010.

[16] Nathan Kallus and Madeleine Udell. 2020. Dynamic assortment personalization in high dimensions. *Operations Research* 68, 4 (2020), 1020–1037.

[17] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. 2008. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*. 681–690.

[18] Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, and Chicheng Zhang. 2020. Contextual Bandits with Continuous Actions: Smoothing, Zooming, and Adapting. *Journal of Machine Learning Research* 21, 137 (2020), 1–45.

[19] Wenhao Li, Ningyuan Chen, and L Jeff Hong. 2019. A dimension-free algorithm for contextual continuum-armed bandits. *arXiv preprint arXiv:1907.06550* (2019).

[20] Vianney Perchet and Philippe Rigollet. 2013. The multi-armed bandit problem with covariates. (2013).

[21] J Xavier Salvat, Jose A Ayala-Romero, Lanfranco Zanzi, Andres Garcia-Saavedra, and Xavier Costa-Perez. 2023. Open radio access networks (O-RAN) experimentation platform: Design and datasets. *IEEE Communications Magazine* 61, 9 (2023), 138–144.

[22] David Simchi-Levi and Yunzong Xu. 2022. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research* 47, 3 (2022), 1904–1931.

[23] Aleksandrs Slivkins. 2019. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272* (2019).

[24] Ambuj Tewari and Susan A Murphy. 2017. From ads to interventions: Contextual bandits in mobile health. In *Mobile health: sensors, analytic methods, and applications*. Springer, 495–517.

[25] Nirandika Wanigasekara and Christina Lee Yu. 2019. Nonparametric contextual bandits in an unknown metric space. *Advances in Neural Information Processing Systems* 32 (2019), 14684–14694.

[26] Zhijin Zhou, Yingfei Wang, Hamed Mamani, and David G Coffey. 2019. How do tumor cytogenetics inform cancer treatments? dynamic risk stratification and precision medicine using multi-armed bandits. *Dynamic Risk Stratification and Precision Medicine Using Multi-armed Bandits (June 17, 2019)* (2019).