Learning-augmented Control via Online Adaptive Policy Selection: No Regret via Contractive Perturbations

Yiheng Lin
California Institute of Technology
Pasadena, USA

James A. Preiss
California Institute of Technology
Pasadena, USA

Emile Anand
California Institute of Technology
Pasadena, USA

Yingying Li
California Institute of Technology
Pasadena, USA

Yisong Yue
California Institute of Technology
Pasadena, USA

Adam Wierman
California Institute of Technology
Pasadena, USA

ACM Reference Format:
Yiheng Lin, James A. Preiss, Emile Anand, Yingying Li, Yisong Yue, and Adam Wierman. 2023. Learning-augmented Control via Online Adaptive Policy Selection: No Regret via Contractive Perturbations. In *Proceedings of ACM Conference (Conference'23).* ACM, Orlando, FL, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 MOTIVATION AND BACKGROUND

Machine-learned control policies have had great empirical success as dynamical systems become increasingly complicated in modern applications ranging from autonomous driving to power systems. Compared with traditional approaches, ML-based policies offer more flexibility to leverage the underlying data distribution and additional information like future predictions. However, when deployed in changing environments, the performance of ML-based policies may vary as the environment changes: When the current environment is aligned with the data distribution that the ML-based policy is trained on, they can achieve near-optimal performance. Otherwise, the ML-based policy may perform much worse than a robust classic controller.

Learning-augmented algorithms provides a promising way to address the challenge of unreliable ML-based policy. A common objective is to achieve near-optimal performance when the ML-based policy is near-optimal (*consistency*) and robust performance when the ML-based policy is unreliable (*robustness*). Towards this goal, a widely-used approach is to combine a robust controller with the ML-based policy [2–4]. Many existing results require the user to know which controller will be robust and decide the parameters of the learning-augmented controller before deployment to achieve some target trade-off between consistency and robustness [3, 6, 7]. However, a limitation of this approach is that the pre-designed trade-off and controller parameters may turn out to be too conservative or optimistic when the performance of ML-based policies is time-varying. Some works on learning-augmented control also study the regret/competitive difference against a static hindsight optimal parameter [4, 5], but the results on adaptive/dynamic regret is still missing. This motivates us to ask the following question:
Can we select the parameters of a learning-augmented controller online that adapts to the current performance of the ML-based policy/advise with provable regret guarantees?

We provide an affirmative answer to this question in this project from the perspective of online policy selection. We propose a novel algorithm, *Gradient-based Adaptive Policy Selection (GAPS)*, for general policy selection problems with provable adaptive/regret guarantees. We also provide a concrete application example for GAPS on learning-augmented Model Predictive Control (MPC).

2 PROBLEM SETTING

We consider a discrete-time online optimal control problem with time-varying dynamics $x_{t+1} = g_t(x_t, u_t)$ and stage costs $f_t(x_t, u_t)$ in a finite horizon T, where x_t and u_t denote the state and control input respectively. The learning-augmented controller is a controller class with parameter $\theta_t \in \Theta$ and it commits the control input $u_t = \pi_t(x_t, \theta_t)$ at time step t. Suppose we design the parameter set Θ such that for some fixed (can be unknown) parameters $\theta^{(c)}$ and $\theta^{(r)}$ in Θ, the consistent controller can be written as $\pi_t(\cdot, \theta^{(c)})$ and the robust controller is $\pi_t(\cdot, \theta^{(r)})$. Different benchmark policy sequences correspond to different objectives for learning-augmented control. For example, to achieve consistency and robustness simultaneously, it suffices for the online controller to achieve no regret against the controller $\pi_t(\cdot, \theta^*)$ with the static hindsight optimal parameter θ^*. In contrast, it is more challenging to adapt to the timely performance of the ML-based policy/advice because the controller must compete against the controller $\pi_t(\cdot, \theta_t^*)$ with time-varying hindsight optimal parameter sequence $\{\theta_t^*\}_{t\in[T]}$.

An an example of our problem setting is learning-augmented MPC, inspired by [5]. In this example, the dynamical system is $g_t(x_t, u_t) = A_t x_t + B_t u_t + w_t$, with (A_t, B_t) known, and the stage costs are $f_t(x_t, u_t) = x_t^\top Q_t x_t + u_t^\top R_t u_t$. At time t, the controller observes $\{Q_{t:t+k-1}, R_{t:t+k-1}, w_{t:t+k-1|t}\}$, where $w_{\tau|t}$ denotes the ML-based prediction of the future disturbance w_τ available at time step t. Learning-augmented MPC commits the first entry of

$$\begin{aligned} \arg\min_{u_{t:t+k-1|t}} \ & \sum_{\tau=t}^{t+k-1} f_\tau(x_{\tau|t}, u_{\tau|t}) + x_{t+k|t}^\top \tilde{Q} x_{t+k|t} \\ \text{s. t. } & x_{t|t} = x_t, \\ & x_{\tau+1|t} = A_\tau x_{\tau|t} + B_\tau u_{\tau|t} + \lambda_t^{[\tau-t]} w_{\tau|t} : \ t \le \tau < t+k, \end{aligned} \tag{1}$$

Conference'23, June 2023, Orlando, FL, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

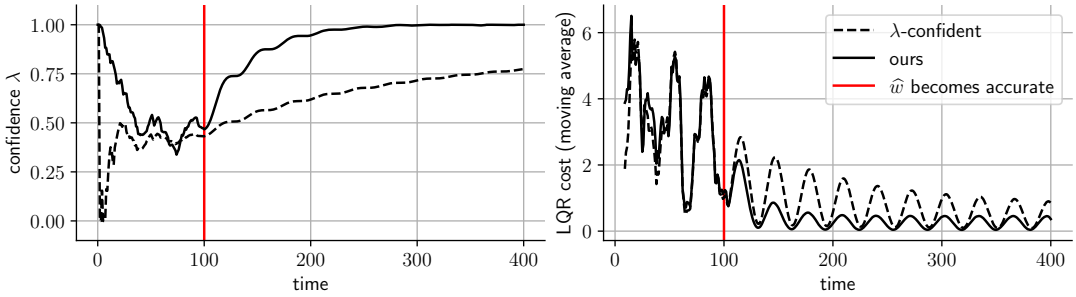

Figure 1: Comparison of GAPS with Li et al. [5] for predictive control of scalar LQR system with sinusoidal disturbance. The disburbance prediction noise is reduced by $100\times$ after time $t = 100$. GAPS adapts more quickly to the change in accuracy.

where $\theta_t = (\lambda_t^{[0]}, \lambda_t^{[1]}, \ldots, \lambda_t^{[k-1]}) \in \Theta = [0,1]^k$ and $\tilde{Q}$ is a fixed positive-definite matrix. Intuitively, $\lambda_t^{[\tau]}$ represents how much we trust the ML-based prediction τ steps into the future. Under this parameterization, the consistent parameter is $\theta^{(c)} = (1, \ldots, 1)$ and the robust parameter is $\theta^{(r)} = (0, \ldots, 0)$, which means completely "trust" or "distrust" the predictions.

3 APPROACH AND CONTRIBUTIONS

In our work, we propose a general exponentially decaying, or "contractive", perturbation property on the policy-induced closed-loop dynamics. The contractive perturbation property we introduce generalizes a key property of disturbance-action controllers [1, 8] and includes other important policy classes such as MPC [5]. We leverage this property to propose Gradient-based Adaptive Policy Selection (GAPS, Algorithm 1). The design of GAPS is inspired by online gradient descent with a novel way to estimate the gradient. Our algorithm can be implemented efficiently for general policy classes in time-varying nonlinear systems when the Jacobians of past dynamics are known.

Our theoretical analysis shows GAPS can translate the regret bounds in online optimization to online policy selection under assumptions that generalize many previous works on online control (e.g., [1, 8]). When applied to the learning-augmented MPC setting in (1), GAPS achieves a dynamic regret bound against a fully time-varying benchmark that depends on the path length.

Theorem 3.1 (Informal). *Under a set of assumptions that are satisfied by learning-augmented MPC in* (1)*, GAPS with step size* $\eta_t = \eta = O(1/\sqrt{T})$ *can achieve a dynamic regret of* $O\left(\sqrt{PT}\right)$ *against the time-varying hindsight optimal parameter sequence* $\{\theta_t^*\}_{t\in[T]}$ *with path length constraint* $\sum_{t=1}^{T-1} \left\|\theta_t^* - \theta_{t-1}^*\right\| \le P$.

Theorem 3.1 implies that in learning-augmented MPC, GAPS can adapt to the real-time quality of ML-based predictions by adjusting the "trust-level" $\{\lambda^{[\tau]}\}_{\tau\in[k]}$ for future disturbance predictions with different look-ahead lengths. We verify our theoretical result by numerically comparing GAPS with the Self-Tuning λ-Confident Control algorithm proposed in [5]. The simulation results show that GAPS can adapt to a change of prediction quality much quicker than the benchmark for comparison (see Figure 1). Our ongoing research focuses on relaxing the assumptions on known dynamics and other assumptions that limits the type of dynamical systems and policy classes our result can apply to.

Algorithm 1 Gradient-based Adaptive Policy Selection (GAPS)

Require: Step size $\{\eta_t\}$, buffer length B, initial parameter θ_0.

for $t = 0, \ldots, T-1$ **do**
  Observe the current state x_t.
  Pick control action $u_t = \pi_t(x_t, \theta_t)$.
  Incur the stage cost $f_t(x_t, u_t)$ and observe f_t, g_{t-1}.
  $\frac{\partial x_t}{\partial \theta_{t-1}} \leftarrow \frac{\partial g_{t-1}}{\partial u_{t-1}}\Big|_{x_{t-1},u_{t-1}} \cdot \frac{\partial \pi_{t-1}}{\partial \theta_{t-1}}\Big|_{x_{t-1},\theta_{t-1}}$.
  $\frac{\partial x_t}{\partial x_{t-1}} \leftarrow \frac{\partial g_{t-1}}{\partial x_{t-1}}\Big|_{x_{t-1},u_{t-1}} + \frac{\partial g_{t-1}}{\partial u_{t-1}}\Big|_{x_{t-1},u_{t-1}} \cdot \frac{\partial \pi_{t-1}}{\partial x_{t-1}}\Big|_{x_{t-1},\theta_{t-1}}$.
  for $b = 2, \ldots, B-1$ **do**
    Use the buffer to compute $\frac{\partial x_t}{\partial \theta_{t-b}} \leftarrow \frac{\partial x_t}{\partial x_{t-1}} \cdot \frac{\partial x_{t-1}}{\partial \theta_{t-b}}$.
  end for
  Compute the approximated gradient

$$G_t = \left(\frac{\partial f_t}{\partial x_t}\Big|_{x_t,u_t} + \frac{\partial f_t}{\partial u_t}\Big|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial x_t}\Big|_{x_t,\theta_t}\right) \cdot \sum_{b=1}^{B-1} \frac{\partial x_t}{\partial \theta_{t-b}} + \frac{\partial f_t}{\partial u_t}\Big|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial \theta_t}\Big|_{x_t,\theta_t} .$$

  Perform the projected gradient update $\theta_{t+1} = \prod_\Theta(\theta_t - \eta_t G_t)$.
  Empty the buffer, and save $\left[\frac{\partial u_t}{\partial \theta_t}, \frac{\partial x_t}{\partial \theta_{t-1}}, \ldots, \frac{\partial x_t}{\partial \theta_{t-B+1}}\right]$.
end for

REFERENCES

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

[2] Antonios Antoniadis, Christian Coester, Marek Elias, Adam Polak, and Bertrand Simon. 2020. Online metric algorithms with untrusted predictions. In *International Conference on Machine Learning*. PMLR, 345–355.

[3] Nicolas Christianson, Tinashe Handina, and Adam Wierman. 2022. Chasing convex bodies and functions with black-box advice. In *Conference on Learning Theory*. PMLR, 867–908.

[4] Tongxin Li, Ruixiao Yang, Guannan Qu, Yiheng Lin, Steven Low, and Adam Wierman. 2022. Equipping Black-Box Policies with Model-Based Advice for Stable Nonlinear Control. *arXiv preprint arXiv:2206.01341* (2022).

[5] Tongxin Li, Ruixiao Yang, Guannan Qu, Guanya Shi, Chenkai Yu, Adam Wierman, and Steven Low. 2021. Robustness and consistency in linear quadratic control with predictions. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 6 (2021), 1–35. Issue 1.

[6] Manish Purohit, Zoya Svitkina, and Ravi Kumar. 2018. Improving online algorithms via ML predictions. *Advances in Neural Information Processing Systems* 31 (2018).

[7] Daan Rutten, Nicolas Christianson, Debankur Mukherjee, and Adam Wierman. 2023. Smoothed Online Optimization with Unreliable Predictions. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 7, 1 (2023), 1–36.

[8] Max Simchowitz, Karan Singh, and Elad Hazan. 2020. Improper learning for non-stochastic control. In *Conference on Learning Theory*. PMLR, 3320–3436.