# Stability Analysis of On-Policy Imitation Learning Algorithms Using Dynamic Regret

Jonathan Lee, Michael Laskey, Ajay Kumar Tanwani, Ken Goldberg
Dept. of Electrical Engineering and Computer Sciences
University of California, Berkeley

***Abstract*—On-policy imitation learning algorithms iteratively improve policies by rolling them out and observing loss functions from expert demonstrations. The theoretical properties of on-policy imitation learning algorithms are often studied from an online optimization or game-theoretic perspective. Of interest recently are conditions necessary for guaranteeing the stability of these algorithms. To analyze stability, we advocate the need for a dynamic regret analysis, which measures the loss of a policy compared to the best it could have done on its own distribution. Using this notion of regret, we investigate stability conditions for on-policy algorithms. We also show that in the infinite sample case the average dynamic regret rate of follow-the-leader under a stability condition and online gradient descent under slightly stronger conditions tend to zero in the number of iterations.**

## I. Introduction

A fundamental problem in imitation learning by supervised learning is covariate shift [1], where the distribution of states visited by the learned policy differs from those seen during training time. Algorithms, often inspired by game-theoretic formulations, such as DAGGER [8], AGGREVATED [9], LOKI [3] and GAIL [6] have been proposed to mitigate the covariate shift. These methods have been shown to perform well in practice and their theoretical analyses support these empirical observations. A fundamental area of concern recently has been to determine under what conditions these algorithms are guaranteed to be stable in the sense that they converge to a locally optimal solution.

We study these properties for on-policy methods, which iteratively rollout the current policy and observe a loss from the expert. Specifically, we focus on the stability of in the generic context of follow-the-leader and online gradient descent, which underlie the aforementioned algorithms. Prior work has employed online optimization analyses to achieve regret bounds for these algorithms. In these analyses, the loss function at an iteration of the algorithm is given by the loss on the distribution of states induced by the current policy parameters. In this work, we define the stability of an algorithm on a particular problem as the convergence of the policy parameters to the optimal parameters on policy's own distribution. Recently Cheng and Boots [2] proved results on the convergence of the final policy for DAGGER, a follow-the-leader algorithm, and suggested that the algorithm can be unstable unless certain conditions are met.

Understanding the stability of imitation learning algorithms in general is important because we are concerned with having the policy perform the best it can on its current distribution. Motivated by this, we advocate the need for *dynamic regret* [10, 7, 4] as a metric for stability of imitation learning algorithms. As opposed to the well known *static regret* which measures hindsight performance with respect to the aggregate of the observed losses, dynamic regret measures performance of a policy at each instantaneous iteration. This metric accurately reflects the goal of stability because it compares the current policy against the best it could be on its distribution with respect to the expert. In other fields, dynamic regret is used for portfolio management and network routing analysis where distributions change over time [5].

Dynamic regret provides us with a general and well-studied framework for evaluating stability of on-policy imitation learning algorithms. Like with static regret, algorithms with sublinear dynamic regret rates in the number of iterations are preferred because their *average* regret rates tend to zero, indicating convergence. As we will see in later sections, it also provides the ability to leverage known information about how the policy affects the state distribution, enabling us to obtain precise statements about an algorithm's performance. It is well known that it is not possible to achieve sublinear dynamic regret in general due to the possibility of well-informed or adversarial loss functions [7]. However, in imitation learning, we may use the regularity and predictability of the loss functions to our advantage. Our contributions are the following:

1) We propose using a dynamic regret analysis to evaluate the stability of on-policy imitation learning algorithms.
2) We present average dynamic regret rates for follow-the-leader and online gradient descent.

## II. Preliminaries

In this section we introduce the notation, problem statement, and assumptions for imitation learning in a supervised learning setting. Let $s_t \in \mathcal{S}$ and $u_t \in \mathcal{U}$ be the state and control in a Markov decision process. The probability of a trajectory $\tau$ of length $T$ under policy $\pi \in \Pi : \mathcal{S} \mapsto \mathcal{U}$ is given by

$$p_\pi(\tau) = p(s_1) \prod_{t=1}^{T-1} p_\pi(u_t|s_t) p(s_{t+1}|s_t, u_t).$$

In this paper, we consider parametric policies, i.e., there is a convex, normed space of parameters $\Theta$ with diameter $D := \max_{\theta_1,\theta_2 \in \Theta} \|\theta_1 - \theta_2\|$ and the imitation learning objective is to find a parameterized policy $\pi_\theta$ where $\theta \in \Theta$ that reduces some loss with respect to the expert policy $\pi^*$ which may not

necessarily be attainable by $\Theta$. The loss of a policy $\pi$ along a trajectory $\tau$ is a non-negative function $J$ such that

$$J(\tau, \pi) = \sum_{t=1}^{T-1} \ell_t(\pi(s_t), \pi^*(s_t)),$$

where $\ell_t : \mathcal{U} \times \mathcal{U} \mapsto \mathbb{R}_{\geq 0}$ is the instantaneous loss. We consider the average loss of of a parameter $\theta_1 \in \Theta$ over the distribution of trajectories generated by a possibly different policy parameter $\theta_2 \in \Theta$:

$$\mathbb{E}_{p(\tau;\pi_{\theta_1})} J(\tau, \pi_{\theta_2}).$$

Thus, $\theta_1$ controls the distribution of trajectories observed and $\theta_2$ controls the predictions in the instantaneous loss. We refer to the first argument as the distribution-generating parameter and the second argument as the evaluation parameter. For example behavior cloning, which involves sampling from the supervisor's distribution, would correspond to minimizing $\mathbb{E}_{(\tau;\pi^*)} J(\tau, \pi_\theta)$ over $\theta$ via empirical risk minimization.

The objective of imitation learning can be written as:

$$\min_{\theta \in \Theta} \mathbb{E}_{p(\tau;\pi_\theta)} J(\tau, \pi_\theta),$$

This corresponds to rolling out and evaluating on the same policy. It reflects the goal of having the policy do well on its own induced distribution. This objective is challenging and cannot be solved with regular supervised learning on the expert's distribution of trajectories since the distribution of inputs is a function of the hypothesis [1].

This paper will consider iterative on-policy algorithms over $N \in \mathbb{N}$ iterations. Specifically at any iteration $n$ for $1 \leq n \leq N$, the policy parameter $\theta_n$ is rolled out as the distribution-generating parameter and the loss $\mathbb{E}_{p(\tau;\pi_{\theta_n})} J(\tau, \pi_\theta)$, is observed. For convenience denote $f_n(\theta) := \mathbb{E}_{p(\tau;\pi_{\theta_n})} J(\tau, \pi_\theta)$. Let $\nabla_n(\theta) := \nabla f_n(\theta)$ denote the gradient in the evaluation parameter at the $n$th iteration. These loss functions form the sequence of losses used in the dynamic regret metric. In this paper, we say that an algorithm is stable if the difference in performance of the policy on its own distribution compared to the optimal on the same distribution tends to zero.

Next we briefly describe the main assumptions of this paper. The assumptions are stated formally in Section V. For the purpose of analysis, as in prior work in both imitation learning and online optimization, we assume strong convexity and smoothness of the loss function in the evaluation parameter. Strong convexity ensures the loss is curved at least quadratically while smoothness guarantees it is not too curved. As in [2], we also assume a regularity constraint on the $f_n$ with respect to the distribution-generating parameter $\theta_n$. This assumption reflects the sensitivity of the loss with respect to the state distribution.

## III. Algorithms

We now review two common algorithms in online optimization that can be directly applied to imitation learning for parameterized policies. Both algorithms are iterative and take advantage of loss functions observed at each iteration.

### A. Follow-The-Leader

DAGGER is a variant of the follow-the-leader algorithm. We begin with a random initial parameter $\theta_1$. Then at iteration $n$, we roll out the current policy $\pi_{\theta_n}$ and observe loss function on current policy's distribution $f_n(\theta)$. We then update the policy with the rule: $\theta_{n+1} = \arg\min_{\theta \in \Theta} \sum_{m=1}^{n} f_m(\theta)$ and repeat.

### B. Online Gradient Descent

Recently there has be interest in "imitation gradients" [3], an imitation analogue of policy gradients in reinforcement learning. Online gradient descent underlies such algorithms. We begin with a random initial parameter $\theta_1$ and stepsize $\eta$. At iteration $n$, we roll out $\pi_{\theta_n}$ and observe loss $f_n(\theta)$. Then we update the policy with $\theta_{n+1} = \theta_n - \eta \nabla f_n(\theta)$ and repeat.

## IV. Dynamic Regret

In order to show stability of an on-policy algorithm, we are interested in showing that the policies generated by the algorithm perform well on the loss on their own induced state distributions. To measure this, we turn to the dynamic regret, defined as

$$R_D(\theta_1, \ldots, \theta_N) := \sum_{n=1}^{N} f_n(\theta_n) - \sum_{n=1}^{N} \min_{\theta \in \Theta} f_n(\theta). \quad (1)$$

In comparison to the more well known static regret, which compares the algorithm's sequence of parameters to the single fixed parameter, dynamic regret compares the $n$th policy to the instantaneous best policy on the $n$th distribution. The advantage of the dynamic regret metric is that the optima track the changes in state distribution so that a policy's performance is always evaluated with respect to the most relevant state distribution, which is the current one. Thus we can examine stability properties of an algorithm by observing the convergence of the *average* dynamic regret, defined as $\frac{1}{N} R_D$.

The dynamic regret of an algorithm is fundamentally dependent on the change in the loss functions over iterations, often expressed in terms of quantities called variations. If the loss functions change in an unpredictable manner, we can expect large variation terms leading to large regret and instability. This is the reason that sublinear dynamic regret bounds cannot be obtained in general using only the assumptions commonly used for static regret [10]. If instead the loss functions change smoothly and slowly, we expect the variation to small, even bounded in some cases. The ability to evaluate regret in terms of the changes in the loss functions allows us to acquire precise bounds and statements about stability.

In imitation learning, the variation of the loss functions is related to the amount change in the state distribution induced by the sequence of policies. Thus if these changes are controlled to some extent, we can prove stability or, at the very least, conditions for stability. The next section will use these notions to show dynamic regret bounds and conditions for stability of follow-the-leader and online gradient descent.

## V. Guarantees

We now formally state the assumptions first introduced in Section II. We begin with assumptions on the loss in the evaluation parameter.

*Assumption 5.1 (Strong Convexity):* For all $n \in \mathbb{N}$ and $\theta_1, \theta_2 \in \Theta$, $\exists \alpha > 0$ such that

$$f_n(\theta_2) \geq f_n(\theta_1) + \langle \nabla_n(\theta_1), \theta_2 - \theta_1 \rangle + \frac{\alpha}{2}\|\theta_1 - \theta_2\|^2.$$

*Assumption 5.2 (Smoothness and Bounded Gradient):* For all $n \in \mathbb{N}$ and $\theta_1, \theta_2 \in \Theta$, $\exists \gamma > 0$ such that

$$\|\nabla_n(\theta_1) - \nabla_n(\theta_2)\| \leq \gamma\|\theta_1 - \theta_2\|$$

and $\exists G > 0$ such that $\|\nabla_n(\theta_1)\| \leq G$.

Finally, we state the regularity constraint on the loss as a function of the distribution-generating parameter, in line with Cheng and Boots [2].

*Assumption 5.3 (Dynamics Regularity):* For all $n, m \in \mathbb{N}$ and $\theta_n, \theta_m, \theta \in \Theta$, $\exists \beta > 0$ such that

$$\|\nabla_n(\theta) - \nabla_m(\theta)\| \leq \beta\|\theta_n - \theta_m\|.$$

We now present our main novel results about well-studied algorithms from online optimization in the context of imitation learning for the infinite sample case. Let $\theta_n^* = \arg\min_{\theta \in \Theta} f_n(\theta)$ be the optimal parameter at iteration $n$. We begin with a result concerning a stability constant $\lambda := \frac{\beta}{\alpha}$. [2].

*Lemma 5.4:* Given the assumptions, the following equality holds on the difference between consecutive optimal parameters for follow-the-leader and online gradient descent algorithms at any $n$:

$$\|\theta_{n+1}^* - \theta_n^*\| \leq \lambda\|\theta_{n+1} - \theta_n\|.$$

*Proof:* By strong convexity of $f_{n+1}$, we have $\frac{\alpha}{2}\|\theta_{n+1}^* - \theta_n^*\|^2 \leq f_{n+1}(\theta_n^*) - f_{n+1}(\theta_{n+1}^*) \leq \|\nabla_{n+1}(\theta_n^*)\|\|\theta_n^* - \theta_{n+1}\| - \alpha/2\|\theta_{n+1}^* - \theta_n^*\|^2$. Then by rearranging terms, $\|\theta_{n+1}^* - \theta_n^*\| \leq \frac{1}{\alpha}\|\nabla_{n+1}(\theta_n^*) - \nabla_n(\theta_n^*)\| \leq \frac{\beta}{\alpha}\|\theta_{n+1} - \theta_n\|$, where the last inequality uses Assumption 5.3. ∎

This lemma suggests that in the case where $\lambda < 1$, we know with certainty that $\|\theta_{n+1}^* - \theta_n^*\| < \|\theta_{n+1} - \theta_n\|$. In other words, the optimal parameters cannot runaway faster than the algorithm's parameters. This intuition is also consistent with the findings of prior work [2], which shows that convergence of the $N$th policy can be guaranteed when $\lambda < 1$ for follow-the-leader.

### A. Follow-The-Leader

We now introduce a dynamic regret corollary to Theorem 2 of Cheng and Boots [2].

*Corollary 5.5:* For follow-the-leader under the assumptions, if $\lambda < 1$, then the average dynamic regret tends towards zero in $N$.

*Proof:* The proof is immediate from the result of Theorem 2 of Cheng and Boots [2]. We have $f_n(\theta_n) - f_n(\theta_n^*) \leq \frac{\left(\lambda e^{1-\lambda} G\right)^2}{2\alpha n^{2(1-\lambda)}}$. Summing from 1 to $N$, we get $\sum_{n=1}^N f_n(\theta_n) - \sum_{n=1}^N f_n(\theta_n^*) \leq \sum_{n=1}^N \frac{\left(\lambda e^{1-\lambda} G\right)^2}{2\alpha n^{2(1-\lambda)}} = O(\max(1, N^{2\lambda-1}))$. Then the average dynamic regret is $\frac{1}{N} R_D = O(\max(1/N, N^{2\lambda-2}))$, which goes to zero. ∎

### B. Online Gradient Descent

For the analysis of dynamic regret bounds for online gradient descent, we require a stronger condition that $\alpha^2 > 2\gamma\beta$. Written another way, the condition is $2\lambda < \psi$ where $\lambda$ is the stability constant and $\psi = \frac{\alpha}{\gamma}$ is the condition number of $f_n$. So we require that the problem is both stable and well-conditioned.

In this proof, we will make use of a variation known as the path variation, which measures the amount of change in the optimal parameters of the loss functions.

*Definition 5.6 (Path Variation):* For a sequence of optimal parameters from $m$ to $n$ given by $\theta_{m:n}^* := (\theta_i^*)_{m \leq i \leq n}$, the path variation is defined as:

$$V(\theta_{m:n}^*) := \sum_{i=m}^{n-1} \|\theta_i^* - \theta_{i+1}^*\|.$$

*Theorem 5.7:* For online gradient descent under the assumptions, if $\lambda < 1$, $2\lambda < \psi$ and $\eta = \frac{\alpha(\alpha^2 - 2\gamma\beta)}{2\beta^2(\alpha^2 - \gamma^2)}$, then the average dynamic regret tends towards zero in $N$.

Before directly proving this theorem, we establish several supporting results based on the path variation.

*Lemma 5.8:* For a sequence of predictions made by the online gradient descent algorithm $\theta_{1:N}$ and a sequence of optimal parameters $\theta_{1:N}^*$, the following inequality holds on the path variation:

$$V(\theta_{1:N}^*) \leq \eta\frac{\beta\gamma}{\alpha}\sum_{n=1}^N \|\theta_n - \theta_n^*\|.$$

*Proof:* From Lemma 5.4, we have $\|\theta_{n+1}^* - \theta_n^*\| \leq \frac{\beta}{\alpha}\|\theta_{n+1} - \theta_n\| = \frac{\beta}{\alpha}\|\eta\nabla_n(\theta_n)\| = \eta\frac{\beta}{\alpha}\|\nabla_n(\theta_n) - \nabla_n(\theta_n^*)\| \leq \eta\frac{\beta\gamma}{\alpha}\|\theta_n - \theta_n^*\|$, where the final inequality uses Assumption 5.2. Then the result follows immediately. ∎

*Lemma 5.9:* Let $\rho = (1 - \alpha\eta + \gamma^2\eta^2)^{1/2}$, which is always nonnegative for any positive choice of $\eta$ because $\gamma \geq \alpha$ by definition. Then the following inequality holds

$$\sum_{n=1}^N \|\theta_n - \theta_n^*\| \leq \|\theta_1 - \theta_1^*\| + \sum_{n=1}^N \rho\|\theta_n - \theta_n^*\| + V(\theta_{1:N}^*).$$

*Proof:* By strong convexity we have the following: $0 \leq 2(f_n(\theta_n) - f_n(\theta_n^*)) \leq 2\langle \nabla_n(\theta_n), x_n - \theta_n^* \rangle - \alpha\|\theta_n^* - \theta_n\|^2$. By the update rule given in the algorithm:

$$\begin{aligned}\|\theta_{n+1} - \theta_n^*\|^2 &= \|\theta_n - \eta\nabla_n(\theta_n) - \theta_n^*\|^2 \\ &= \|\eta\nabla_n(\theta_n)\|^2 + \|\theta_n - \theta_n^*\|^2 \\ &\quad - 2\eta\langle \nabla_n(\theta_n), \theta_n - \theta_n^* \rangle. \end{aligned} \tag{2}$$

By rearranging the terms in (2) and combining with the very first inequality, we arrive at the following:

$$\|\theta_{n+1} - \theta_n^*\|^2 \leq (1 - \alpha\eta)\|\theta_n - \theta_n^*\|^2 + \|\eta\nabla_n(\theta_n)\|^2.$$

Using Assumption 5.2 and the fact that $\nabla_n(\theta_n^*) = 0$:

$$\begin{aligned}\|\theta_{n+1} - \theta_n^*\|^2 &\leq \|\theta_n - \theta_n^*\|^2 - \alpha\eta\|\theta_n - \theta_n^*\|^2 \\ &\quad + \eta^2\|\nabla_n(\theta_n) - \nabla_n(\theta_n^*)\|^2 \\ &\leq \left(1 - \alpha\eta + \gamma^2\eta^2\right)\|\theta_n - \theta_n^*\|^2. \end{aligned} \tag{3}$$

Then let $\rho = \left(1 - \alpha\eta + \gamma^2\eta^2\right)^{1/2}$. Following from [7], consider the series:

$$\begin{aligned}\sum_{n=1}^{N} \|\theta_n - \theta_n^*\| &= \|\theta_1 - \theta_1^*\| + \sum_{n=2}^{N} \|\theta_n - \theta_{n-1}^* + \theta_{n-1}^* - \theta_n^*\| \\ &\leq \|\theta_1 - \theta_1^*\| + \sum_{n=2}^{N} \|\theta_n - \theta_{n-1}^*\| + V(\theta_{1:N}^*) \\ &\leq \|\theta_1 - \theta_1^*\| + \sum_{n=1}^{N} \rho\|\theta_n - \theta_n^*\| + V(\theta_{1:N}^*),\end{aligned}$$

where the second line uses the definition of the path variation and the third line uses (3). ∎

*Proof of Theorem 5.7:* We begin by bounding the result from Lemma 5.9 above by Lemma 5.8:

$$\sum_{n=1}^{N} \|\theta_n - \theta_n^*\| \leq \|\theta_1 - \theta_1^*\| + \left(\rho + \eta\frac{\beta\gamma}{\alpha}\right) \sum_{n=1}^{N} \|\theta_n - \theta_n^*\|.$$

By rearranging the terms and bounding by the diameter of $\mathcal{X}$:

$$\sum_{n=1}^{N} \|\theta_n - \theta_n^*\| \leq \frac{D}{1 - \rho - \eta\frac{\beta\gamma}{\alpha}}.$$

It can be shown that, under the assumptions, the choice of $\eta = \frac{\alpha(\alpha^2 - 2\beta\gamma)}{2\gamma^2(\alpha^2 - \beta^2)}$ ensures that $\left(1 - \rho - \eta\frac{\beta\gamma}{\alpha}\right)$ is positive. By the $G$-Lipschitz continuity of $f_n$, we have

$$\sum_{n=1}^{N} f_n(\theta_n) - \sum_{n=1}^{N} f_n(\theta_n^*) \leq \frac{GD}{1 - \rho - \eta\frac{\beta\gamma}{\alpha}},$$

and so $R_D(\theta_1, \ldots, \theta_N) = O(1)$. So we have $\frac{1}{N}R_D = O(1/N)$ which goes to zero. ∎

## VI. Discussion and Future Work

There are several important implications of these theorems. The first is that there is an inherent property of stability in on-policy algorithms. Even in the case considered in this paper with assumptions of strong convexity and smoothness and regularity of the dynamics, these algorithms may be at risk of instability when the optima of the loss functions can move faster than the policy parameters, suggesting that careful regularization such as that described by Cheng and Boots [2] may be necessary for on-policy algorithms to stabilize them.

Secondly, the use of dynamic regret to compare the performance of a policy with the best on the same distribution allows us to obtain precise bounds and insights to this stability. When the $O(1)$ bound is attained, the dynamic regret upper bound is constant meaning that it does not grow with the number of iterations. It is clear that in such cases the path variation is also constant in the number of iterations which suggests the series converges and thus the distances between consecutive parameter solutions tend to zero.

These observations make sense intuitively when considering the proofs. We want the parameter vector go towards a solution as in Lemma 5.9 and simultaneously we want consecutive solutions to move slowly as in Lemma 5.8. While these may seem to be competing objectives, we can leverage the regularity of the state distribution to combine the two yielding improved regret rates when the stability conditions are met.

Finally the presented results have implications beyond strictly on-policy algorithms. The GAIL algorithm iteratively updates the policy based on a loss parameterized by an adversarial discriminator. Between iterations, the discriminator is updated, altering the loss function with respect to the change in policy parameters. We believe that similar stability results could exist for this algorithm and similar ones.

## References


[1] J Andrew Bagnell. An invitation to imitation. Technical report, Carnegie Mellon Univ Pittsburgh PA Robotics Inst, 2015.

[2] Ching-An Cheng and Byron Boots. Convergence of value aggregation for imitation learning. *International Conference on Artificial Intelligence and Statistics*, 2018.

[3] Ching-An Cheng, Xinyan Yan, Nolan Wagener, and Byron Boots. Fast policy learning through imitation and reinforcement. In *Conference on Uncertainty in Artificial Intelligence*, 2018.

[4] Eric C Hall and Rebecca M Willett. Online convex optimization in dynamic environments. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):647–662, 2015.

[5] Elad Hazan and Comandur Seshadhri. Adaptive algorithms for online decision problems. *Electronic colloquium on computational complexity (ECCC)*, 2007.

[6] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pages 4565–4573, 2016.

[7] Aryan Mokhtari, Shahin Shahrampour, Ali Jadbabaie, and Alejandro Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pages 7195–7201. IEEE, 2016.

[8] Stéphane Ross, Geoffrey J Gordon, and J Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. *International Conference on Artificial Intelligence and Statistics*, 2011.

[9] Wen Sun, Arun Venkatraman, Geoffrey J Gordon, Byron Boots, and J Andrew Bagnell. Deeply aggrevated: Differentiable imitation learning for sequential prediction. In *International Conference on Machine Learning*, 2017.

[10] Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *International Conference on Machine Learning*, 2016.