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

Jonathan Lee, Michael Laskey, Ajay Kumar Tanwani and Ken Goldberg

University of California, Berkeley

Motivation

A fundamental problem in imitation learning by supervised learning is covariate shift, 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 [3], AGGREVATED [4], and LOKI [2] have been proposed to mitigate the covariate shift.

An area of interest 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 [1].

We address the question of when certain online imitation learning algorithms lead to policies that perform the best they can on their own distribution.

Contributions

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

On-Policy Imitation Learning

We consider on-policy imitation learning algorithms. At any iteration n for $1 \leq n \leq N$, a policy π_θ parameterized by $\theta_n \in \Theta$ is rolled out, inducing a distribution over trajectories $p(\tau; \pi_\theta)$ and an observed loss function on that distribution

$$f_n(\theta) = \mathbb{E}_{p(\tau;\pi_{\theta_n})} J(\theta, \tau),$$

where J is a loss function defined by the supervisor. Then θ_{n+1} is computed using $f_1, \ldots, f_n$.

Algorithms

- *Follow-The-Leader*: DAGGER is a variant of the follow-the-leader algorithm. Update rule: $\theta_{n+1} = \arg\min_\theta \sum_{m=1}^{n} f_m(\theta)$
- *Online Gradient Descent*: Recently there has be interest an imitation analogue of policy gradients. Online gradient descent underlies such algorithms. Update rule: $\theta_{n+1} = \theta_n - \eta \nabla f_n(\theta)$

Dynamic Regret

$$R_D = \sum_{n=1}^{N} f_n(\theta_n) - \sum_{n=1}^{N} \min_{\theta \in \Theta} f_n(\theta)$$

Dynamic regret is well-studied in online optimization for online problems with changing distributions.

In comparison to the more well known static regret, which compares the algorithm's sequence of parameters to a single fixed parameter, dynamic regret compares the nth policy to the instantaneous best policy on the nth distribution. The advantage 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. Stability properties can be observed in terms of convergence of average dynamic regret.

Dynamic regret rates are often proportional to the amount of variation of the loss functions over iterations, so low regret is not possible to prove in general.

- If the loss functions change in an unpredictable manner, the regret can be large and lead to instability.
- If the variation of the loss functions is small, convergence in average regret can be proved.
- In imitation learning, the distribution of f_n is controlled by θ_n. If θ_n change slowly, then so might f_n.

Dynamic Regret Insight

In imitation learning, the variation of the loss functions $f_1, \ldots, f_N$ is non-adversarial. It is controlled by the on-policy algorithm. Knowledge of the variation of θ_n can be used to make strong guarantees.

Guarantees

For all n, let $f_n(\theta)$ be α-strongly convex and γ-smooth in θ. For all $\theta_1, \theta_2 \in \Theta$, assume the following bound holds:

$$\|\nabla_\theta \mathbb{E}_{p(\tau;\pi_{\theta_1})} J - \nabla_\theta \mathbb{E}_{p(\tau;\pi_{\theta_2})} J\| \leq \beta \|\theta_1 - \theta_2\|$$

Follow-The-Leader

If $\alpha > \beta$, then the average dynamic regret tends towards zero with rate $\frac{1}{N} R_D = O(\max(1/N, N^{2\beta/\alpha - 2}))$.

Online Gradient Descent

If $\alpha > \beta$, $\alpha^2 > 2\beta\gamma$ and the step-size is $\eta = \frac{\alpha(\alpha^2 - 2\gamma\beta)}{2\gamma^2(\alpha^2 - \beta^2)}$, then the average dynamic regret tends towards zero with rate $O(1/N)$.

References

[1] Ching-An Cheng and Byron Boots.
Convergence of value aggregation for imitation learning.
AISTATS, 2018.

[2] Ching-An Cheng, Xinyan Yan, Nolan Wagener, and Byron Boots.
Fast policy learning through imitation and reinforcement.
In *UAI*, 2018.

[3] Stéphane Ross, Geoffrey J Gordon, and J Andrew Bagnell.
A reduction of imitation learning and structured prediction to no-regret online learning.
AISTATS, 2011.

[4] Wen Sun, Arun Venkatraman, Geoffrey J Gordon, Byron Boots, and J Andrew Bagnell.
Differentiable imitation learning for sequential prediction.
In *ICML*, 2017.