Sensitivity Analysis for Neural Network Controllers

Kyle Morgenstein[1,2]
kjm3887

***Abstract*—**

I. INTRODUCTION

Reinforcement learning (RL) has become the standard control technique in legged locomotion due to its robustness, expressivity, and ease of deployment. Despite these attributes, RL control policies still fail catastrophically when evaluated on out of distribution (OOD) inputs. Due to the black box nature of RL policies, safety efforts largely focus on observing potentially destabilizing changes in the action space of the policy, and triggering safe modes when risk thresholds are reached (e.g. over-current protection). Efforts to quantify the distribution of valid inputs during training may result in more proactive runtime anomaly detection, but such efforts provide only weak guarantees of stability given the distribution shift between simulation-based training and hardware deployment. In this work we propose a more rigorous treatment of anomaly detection using tools from nonlinear sensitivity analysis. Treating the trained policy as an artifact, we aim to exploit the structure of the learning-based controller to provide stronger guarantees to prevent catastrophic failure at runtime.

II. SYSTEM MODELING

Consider a nonlinear, time varying system

$$\dot{x}_t = f(x_t) + g(x_t)u_t \tag{1}$$

with state $x_t \in \mathbb{R}^m$ and control signal $u_t \in \mathbb{R}^n$. We seek to understand the sensitivity of the closed loop dynamics without assuming knowledge of f or g. Let $u_t = \pi(z_t)$ be the output from a neural network controller tracking a reference signal such that the error dynamics

$$\begin{aligned} z_t &:= x_t - x_t^{\text{ref}} \\ \dot{z}_t &= F(z_t, \pi(z_t)) \end{aligned} \tag{2}$$

has an equilibrium at $F(0, \pi(0)) = 0 \forall t$. The linearized closed loop system Jacobian

$$J_{\text{cl}} = \left.\frac{\partial F}{\partial z_t}\right|_{z_t=0} \tag{3}$$

and can be estimated as follows:

For $k = 1, ..., N$ select perturbation direction $d^{(k)} \in \mathbb{R}^n$ with $||d^{(k)}|| = 1$ and radius $h \in \mathbb{R}$. Then define the finite difference

$$y^{(k)} := \frac{F(hd^{(k)}) - F(-hd^{(k)})}{2h} \tag{4}$$

with estimator $y^{(k)} = J_{\text{cl}}d^{(k)} + \epsilon^{(k)}$, $\epsilon^{(k)} \sim \mathcal{N}(0, \Sigma)$. Let $Y = [y^{(1)}, ..., y^{(N)}] \in \mathbb{R}^{n\times N}$, $D = [d^{(1)}, ..., d^{(N)}] \in \mathbb{R}^{n\times N}$ and $E = [\epsilon^{(1)}, ..., \epsilon^{(N)}] \in \mathbb{R}^{n\times N}$. Then,

$$\begin{aligned} Y &= J_{\text{cl}}D + E \\ \hat{J}_{\text{cl}} &= YD^T(DD^T)^{-1} \end{aligned} \tag{5}$$

if DD^T is invertible (i.e. $\{d^{(k)}\}$ spans $\mathbb{R}^n$). A valid choice of $\{d^{(k)}\}$ requires $N \geq n$.

Divergence happens near singularity of open loop dynamics and sensitive regions of policy

Show with $M\ddot{q} = \tau$

Assume not near singularity

Then sensitive only if policy is sensitive

Do policy analysis

III. SYSTEM ANALYSIS AND DESIGN

IV. SIMULATION RESULTS

V. CONCLUSIONS

This template provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in

[1]Aerospace Engineering, UT Austin `kylem@utexas.edu`
[2]Apptronik, Inc. `kylemorgenstein@apptronik.com`

italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

VI. MATH

A. *Equations*

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled. Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in

$$\alpha + \beta = \chi \tag{1}$$

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

B. *Figures and Tables*

Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence. [1]

TABLE I

AN EXAMPLE OF A TABLE

One	Two
Three	Four

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an document, this method is somewhat more stable than directly inserting a picture.

Fig. 1. Inductance of oscillation winding on amorphous magnetic core versus DC bias magnetic field

APPENDIX

Appendixes should appear before the acknowledgment.

ACKNOWLEDGMENT

REFERENCES

[1] R. Hermann and A. Krener, “Nonlinear controllability and observability,” *IEEE Transactions on Automatic Control*, vol. 22, no. 5, pp. 728–740, 1977.