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 $x_t \in \mathbb{R}^n$ and control signal $u_t \in \mathbb{R}^m$. 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 forward difference

$$y^{(k)} := \frac{F(hd^{(k)}) - F(0)}{h} \tag{4}$$

with estimator $y^{(k)} = J_{\text{cl}} d^{(k)} + \frac{r^{(k)}}{h}$ and remainder $r^{(k)} = \frac{1}{2}(hd^{(k)})^T \mathcal{H}(hd^{(k)}) = \mathcal{O}(h^2)$. Let $Y = [y^{(1)}, ..., y^{(N)}] \in \mathbb{R}^{n \times N}$, $D = [d^{(1)}, ..., d^{(N)}] \in \mathbb{R}^{n \times N}$ and $R = [\frac{r^{(1)}}{h}, ..., \frac{r^{(N)}}{h}] \in \mathbb{R}^{n \times N}$. Then,

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

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

To show convergence to the equilibrium, we must show that J_{cl} is Hurwitz. Define the estimation error

$$\Delta J_{\text{cl}} := \hat{J}_{\text{cl}} - J_{\text{cl}}. \tag{6}$$

From the estimator model we have

$$\Delta J_{\text{cl}} = R D^T (DD^T)^{-1} \tag{7}$$

with bound $||\Delta J_{\text{cl}}||_2 \leq ||R||_2 ||D^T (DD^T)^{-1}||_2$. Using the singular value decomposition $D = U \Sigma V^T$, the directional term on the right-hand side can be simplified $D^T (DD^T)^{-1} = V \Sigma^{-1} U^T$, yielding

$$||D^T (DD^T)^{-1}||_2 = \frac{1}{\sigma_{\min}(D)}. \tag{8}$$

Where $\{d^{(k)}\}$ is selected as N i.i.d unit norm isotropic random directions, $DD^T \approx \frac{N}{n} I$ results in $\sigma_{\min}(D) \approx \sqrt{\frac{N}{n}}$ with high probability for large N. Next, by assuming $F \in \mathcal{C}^2$, we can bound the second-order Taylor

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

remainder as $||\frac{r^{(k)}}{h}||_2 \leq \frac{1}{2}L_2 h||d^{(k)}||_2$. The constant L_2 can be estimated via the second directional derivative

$$\begin{aligned}\hat{L}_2 &= \sup_{||d||=1} ||D^2 F(0)[d,d]||_2 \\ &\approx \max_k \left|\left|\frac{F(hd^{(k)}) - 2F(0) + F(-hd^{(k)})}{h^2}\right|\right|\end{aligned} \tag{9}$$

To account for numerical error, a small scale may be used $L_2 = \beta\hat{L}_2$, $\beta > 1$. Then,

$$\begin{aligned}||R||_2 &\leq \max_k ||\frac{r^{(k)}}{h}||_2\sqrt{\text{rank}(R)} \\ &\leq \max_k ||\frac{r^{(k)}}{h}||_2\sqrt{\min(n,N)} \\ &\leq \frac{1}{2}L_2 h\sqrt{n}\end{aligned} \tag{10}$$

using the requirement that $\{d^{(k)}\}$ spans $\mathbb{R}^n$. Substituting bounds into Eq. 7, we find a computable bound on the estimation error

$$||\Delta J_{\text{cl}}||_2 \leq \frac{1}{2}L_2 h\frac{n}{\sqrt{N}} \tag{11}$$

From this upper bound we may now derive the conditions to certify that J_{cl} is Hurwitz given $\hat{J}_{\text{cl}}$. Assume J_{cl} is diagonalizable. Let $\hat{J}_{\text{cl}} = V\Lambda V^{-1}$ with eigenvalues $\{\hat{\lambda}_i\}$ and margin $\hat{\lambda} = -\max_i \operatorname{Re}\hat{\lambda}_i$. The Baur-Fike Theorem [1] gives a bound on the distance between an eigenvalue λ_i of $J_{\text{cl}} = \hat{J}_{\text{cl}} - \Delta J_{\text{cl}}$ and $\hat{\lambda}_i$:

$$\begin{aligned}|\lambda_i - \hat{\lambda}_i| &\leq \kappa(V)||\Delta J_{\text{cl}}||_2 \\ &\leq \kappa(V)\frac{1}{2}L_2 h\frac{n}{\sqrt{N}}\end{aligned} \tag{12}$$

with condition number $\kappa(V) = ||V||_2||V^{-1}||_2$ for the matrix of eigenvectors of $\hat{J}_{\text{cl}}$. Therefore, if

$$\kappa(V)\frac{1}{2}L_2 h\frac{n}{\sqrt{N}} \leq \hat{\lambda} \tag{13}$$

then

$$\begin{aligned}\operatorname{Re}\lambda_i &\leq \operatorname{Re}\hat{\lambda}_i + |\lambda_i - \hat{\lambda}_i| \\ &\leq -\hat{\lambda} + \kappa(V)\frac{1}{2}L_2 h\frac{n}{\sqrt{N}} \\ &< 0\forall\lambda_i.\end{aligned} \tag{14}$$

Thus, Eq. 13 is sufficient to conclude that J_{cl} is Hurwitz. We have now certified that the closed loop dynamics are locally exponentially stable based on the sampled response within the perturbation radius h for each $t > t_0$. This proof holds despite only assuming J_{cl} is diagonalizable, $\{d^{(k)}\}$ is full rank, and $F \in \mathcal{C}^2$.

III. System Analysis and Design

IV. Simulation Results

V. Conclusions

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. [2]

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

References

[1] F. L. Bauer and C. T. Fike, “Norms and exclusion theorems,” *Numerische mathematik*, vol. 2, no. 1, pp. 137–141, 1960.

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