UDK 517.54
DOI: 10.37069/1683-4720-2025-39-7

©2025. **М. Шамрай**

CLOSED-FORM ROBUSTNESS BOUNDS FOR SECOND-ORDER PRUNING OF NEURAL CONTROLLER POLICIES

Deep neural policies have unlocked agile flight for quadcopters, adaptive grasping for manipulators, and reliable navigation for ground robots, yet their millions of weights conflict with the tight memory and real-time constraints of embedded microcontrollers. Second-order pruning methods – *Optimal Brain Damage* (OBD) and its variants, including *Optimal Brain Surgeon* (OBS) and the recent SparseGPT – compress networks in a single pass by leveraging the local Hessian, achieving far higher sparsity than magnitude thresholding. Despite their success in vision and language, the consequences of such weight removal on *closed-loop* stability, tracking accuracy, and safety have remained unclear. We present the first mathematically rigorous robustness analysis of second-order pruning in nonlinear discrete-time control. The system evolves under a continuous transition map $f : X \times U \to X$, while the controller is an L-layer multilayer perceptron with ReLU-type activations that are *globally* 1-Lipschitz. Pruning the weight matrix of layer k replaces W_k with $W_k + \delta W_k$, producing the perturbed parameter vector $\widehat{\Theta} = \Theta + \delta\Theta$ and the pruned policy $\pi(\cdot; \widehat{\Theta})$. For every input state $s \in X$ we derive the closed-form inequality $\|\pi(s;\Theta) - \pi(s;\widehat{\Theta})\|_2 \le C_k(s)\,\|\delta W_k\|_2$, where the constant $C_k(s)$ depends *only* on unpruned spectral norms and biases, and can be evaluated in closed form from a single forward pass. When several layers $S \subseteq \{1,\dots,L\}$ are pruned we show the deviations add *linearly*, $\sum_{k\in S} C_k(s)\|\delta W_k\|_2$, yielding an explicit state-dependent or worst-case robustness estimation. All constants are computed *offline*. Therefore, environment roll-outs, back-propagation, or hyper-parameter tuning are *not* necessary, what makes the framework directly usable inside compression pipelines. The derived bounds specify, prior to field deployment, the maximal admissible pruning magnitude compatible with a prescribed control-error threshold. By linking second-order network compression with closed-loop performance guarantees, our work narrows a crucial gap between modern deep-learning tooling and the robustness demands of safety-critical autonomous systems.

MSC: 34N05.

Keywords: *second-order pruning, neural controller robustness, spectral-norm bounds, Lipschitz multilayer perceptron.*

1. Introduction.

Modern autonomous systems, from agile quadcopters to embodied household robots, are increasingly controlled by *neural policies* that map high-dimensional sensor data directly to control actions. Rich function classes such as multilayer perceptrons (MLPs), convolutional networks, or *Vision–Language–Action* (VLA) models achieve impressive closed-loop performance when trained by reinforcement learning (RL) or behaviour cloning [1, 2]. Yet deploying those policies on size, weight, and power-constrained hardware demands aggressive *model compression*. Among the most practical compression

The authors confirm that there is no conflict of interest and acknowledges financial support by the Simons Foundation grant (SFI-PD-Ukraine-00014586, M.S.) and the project 0125U000299 of the National Academy of Sciences of Ukraine. We also express our gratitude to the Armed Forces of Ukraine for their protection, which has made this research possible.

techniques are *second-order pruning* algorithms – Optimal Brain Damage (OBD) [3], Optimal Brain Surgeon (OBS) [4], and their recent large-model successor *SparseGPT* [5] – which remove weights that minimise an analytically estimated activation loss.

While second-order pruning is widely used for large language models, its impact on the *closed-loop behaviour* of a control system is poorly investigated. A small perturbation of the weights can propagate through the policy network, alter the control signal, and ultimately degrade safety or task performance. Precise guarantees on how much pruning a controller can tolerate are therefore crucial for safety-critical applications [6,7,11,13]. To the best of our knowledge, *no prior work provides a rigorous, closed–form upper bound on the control error induced by second–order pruning* in nonlinear systems.

We provide the first *robustness analysis* of OBD/OBS-style weight pruning for deterministic nonlinear control systems. Our goal is to bound, in closed form, the deviation of the pruned control signal.

Contributions.

1. **Formal robustness framework.** Section casts pruning as a perturbation of the parameter vector Θ and formulates the *pruning robustness problem* via the control–error bound (3).

2. **Tight single–layer bound.** Theorem 1 proves that for any 1–Lipschitz (ReLU–type) activation the deviation of a pruned layer k is

$$\left\|\pi(s;\Theta) - \pi(s;\widehat{\Theta})\right\|_2 \;\le\; C_k(s)\,\|\delta W_k\|_2,$$

   where the constant $C_k(s)$ depends only on *unpruned* weights, biases and the input norm.

3. **Extension to multiple layers.** Corollary 1 extends the result to an arbitrary set of pruned layers in an *additive* fashion, yielding the computable bound $B_\pi(\delta\Theta) = \sum_{k\in S} C_{k,\max}\|\delta W_k\|_2$.

4. **Practical implications.** The bounds can be evaluated *offline* from a forward pass and spectral norms alone, enabling principled trade–offs between compression ratio and control robustness without repeated interaction with the physical system.

2. Formal Problem Setup.

2.1 Deterministic nonlinear control problem.

Definition Controlled dynamics. Let the **state space** be $X \subset \mathbb{R}^n$ and the **action space** be $U \subset \mathbb{R}^m$. A trajectory $\{x_t\}_{t\ge 0}$ evolves according to the discrete–time difference equation

$$x_{t+1} \;=\; f\big(x_t, u_t\big), \qquad t = 0, 1, 2, \ldots,$$

where the transition map $f : X \times U \to X$ is continuous, ensuring that a unique trajectory exists for every admissible control sequence $\{u_t\}_{t\ge 0}$ and initial state $x_0 \in X$.

Definition Parametric neural policy. Let $L \in \mathbb{N}$ be the number of affine layers. Collect all weights and biases in the parameter vector

$$\Theta \;=\; \{W_\ell, b_\ell\}_{\ell=1}^{L} \in \mathbb{R}^q.$$

The **policy** $\pi(\,\cdot\,;\Theta) : X \to U$ is the neural network obtained by composing these affine layers with pointwise, 1–Lipschitz activations σ:

$$\pi(x;\Theta) \;:=\; (\sigma \circ A_L) \circ \cdots \circ (\sigma \circ A_1)(x), \qquad A_\ell(z) \;=\; W_\ell z + b_\ell.$$

The control applied at time t is $u_t = \pi(x_t;\Theta)$.

Definition Discounted return. Fix an initial distribution $x_0 \sim \mu$, a bounded reward $r : X \times U \to \mathbb{R}$, and a discount factor $\gamma \in (0,1)$. The expected return of policy $\pi(\cdot;\Theta)$ is

$$J(\Theta) \;:=\; \mathbb{E}_{x_0 \sim \mu}\big[\textstyle\sum_{t=0}^{\infty} \gamma^t\, r\big(x_t, \pi(x_t;\Theta)\big)\big].$$

An **optimal policy** is any maximiser $\Theta^\star \in \arg\max_{\Theta \in \mathbb{R}^q} J(\Theta)$.

2.2 Second–order (OBD) pruning criterion.

Throughout this paper we study the effect of pruning individual weights in Θ by the *Optimal Brain Damage (OBD)* saliency [3]. Consider a single affine layer $A_\ell(z) = W_\ell z + b_\ell$ and an input mini–batch $X_\ell \in \mathbb{R}^{d \times n_\ell}$. The change in pre–activation outputs caused by replacing W_ℓ with $\widehat{W}_\ell = W_\ell + \delta W_\ell$ is measured by

$$E_\ell(W_\ell, \widehat{W}_\ell) \;=\; \big\|W_\ell X_\ell - \widehat{W}_\ell X_\ell\big\|_F^2. \tag{1}$$

A second–order Taylor expansion of E_ℓ around W_ℓ yields

$$E_\ell(W_\ell + \delta W_\ell) \;\approx\; \tfrac{1}{2}\,\mathrm{vec}(\delta W_\ell)^\top H_\ell\, \mathrm{vec}(\delta W_\ell), \qquad H_\ell := \nabla^2_{W_\ell} E_\ell\,.$$

Pruning a *single* weight $w_q \in \Theta$ (that is, setting $\widehat{w}_q = 0$) gives the OBD saliency

$$\Delta E_q \;=\; \frac{1}{2}\,\frac{w_q^2}{(H_\ell^{-1})_{qq}}, \tag{2}$$

which serves as an importance score for that weight. Aggregating (2) over the layers to which OBD is applied produces a pruned parameter vector $\widehat{\Theta} \in \mathbb{R}^q$ and the perturbation $\delta\Theta := \widehat{\Theta} - \Theta$.

2.3 Pruning robustness problem.

Our objective is to quantify **how sensitive the closed–loop behaviour is to the OBD perturbation** $\delta\Theta$. Specifically, we seek a computable bound $B_\pi : \mathbb{R}^q \to \mathbb{R}_+$ such that

$$\big\|\pi(x;\Theta) - \pi(x;\widehat{\Theta})\big\|_2 \;\le\; B_\pi(\delta\Theta), \qquad \forall\, x \in X. \tag{3}$$

Although inequality (3) applies to *any* feed–forward network, we focus on the control setting because the resulting guarantees translate directly into performance and safety margins for autonomous systems.

Remark. For clarity, we treat each weight matrix W_ℓ and bias b_ℓ as a distinct block within Θ. Hence any scalar weight w_q appearing in (2) is an element of Θ, and replacing w_q by 0 modifies Θ exactly as required in the control bound (3).

3. Background.

All robustness estimates in this paper rest on *Lipschitz control* of the policy network. We collect the required analytic facts in this section.

3.1 Lipschitz mappings.

Definition Global Lipschitz continuity. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called L–**Lipschitz** with respect to the Euclidean norm if

$$\|f(x) - f(y)\|_2 \;\le\; L\,\|x - y\|_2, \qquad \forall\, x, y \in \mathbb{R}^n. \tag{4}$$

The smallest such constant is denoted $L(f)$.

Theorem Rademacher [14]. *Every locally Lipschitz map $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable almost everywhere and $L(f) = \operatorname{ess\,sup}_x \|Df(x)\|_2$, where $\|\cdot\|_2$ is the operator norm induced by ℓ_2.*

3.2 Lipschitz multilayer perceptrons.

Definition MLP with 1–Lipschitz activations. Fix $L \in \mathbb{N}$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ satisfy $|\sigma(a) - \sigma(b)| \le |a - b|$. An L**-layer MLP** is the map

$$f(x;\Theta) := (\sigma \circ A_L) \circ \cdots \circ (\sigma \circ A_1)(x), \qquad A_\ell(z) = W_\ell z + b_\ell,$$

with parameters $\Theta = \{W_\ell, b_\ell\}_{\ell=1}^{L}$ and $W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$.

Proposition Spectral-norm Lipschitz bound [12]. For the MLP of Definition ,

$$L\big(f(\,\cdot\,;\Theta)\big) \;\le\; \prod_{\ell=1}^{L} \|W_\ell\|_2. \tag{5}$$

Interpretation for pruning analysis. Equation (5) expresses the *global* Lipschitz constant of a neural policy directly in terms of the spectral (operator-norm) factors that appear in the OBD parameter vector Θ. Hence perturbing any weight matrix $W_k \mapsto W_k + \delta W_k$ alters both $L\big(f(\cdot;\Theta)\big)$ and the saliency (2), allowing us to translate the activation-level OBD error into a control-signal bound (3) in later sections.

4. Results.

Proposition 1. Non-expansiveness of ReLU-type activations. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be an activation satisfying

1. *zero anchor:* $\varphi(0) = 0$,
2. *unit Lipschitz:* $|\varphi(a) - \varphi(b)| \le |a - b|$ for all $a, b \in \mathbb{R}$.

Define the component-wise map

$$\sigma : \mathbb{R}^n \to \mathbb{R}^n, \qquad \sigma(x)_i \;=\; \varphi(x_i),\; i = 1, \ldots, n.$$

Then σ is ℓ_2–*non-expansive*:

$$\|\sigma(x)\|_2 \;\le\; \|x\|_2, \qquad \forall\, x \in \mathbb{R}^n. \tag{6}$$

Proof. For any $x \in \mathbb{R}^n$ we have, by the scalar Lipschitz property with $b = 0$, $|\varphi(x_i)| \le |x_i|$. Squaring, summing and taking the square root yields

$$\|\sigma(x)\|_2^2 \;=\; \sum_{i=1}^{n} \varphi(x_i)^2 \;\le\; \sum_{i=1}^{n} x_i^2 \;=\; \|x\|_2^2,$$

from which (6) follows. □

Condition (6) is satisfied by many ReLU-type activations that are ubiquitous in deep learning:

- **ReLU** $\varphi(x) = \max\{0, x\}$ [15];
- **Leaky-ReLU** $\varphi(x) = \max\{x, \alpha x\}$ with $0 < \alpha \le 1$ [16];
- **PReLU** (parametric ReLU) with learnable $\alpha \in (0, 1]$ [17];
- **ELU** $\varphi(x) = \max\{x, \alpha(e^x - 1)\}$ for $0 < \alpha \le 1$ [8];
- **GELU** (Gaussian-error linear unit), a smooth approximation to ReLU whose derivative is bounded by 1 [9].

All these functions obey $\varphi(0) = 0$ and have slope bounded by 1, hence are 1-Lipschitz, therefore Proposition 1 applies.

Theorem 1. Robustness of an OBD-pruned policy. *Let $\pi(\,\cdot\,;\Theta)$ be the L-layer MLP controller*

$$x_0 = s, \qquad x_\ell = \sigma\big(W_\ell x_{\ell-1} + b_\ell\big),\ \ell = 1, \dots, L, \quad \Pi(s;\Theta) = x_L,$$

with ReLU-type activations σ_l and weights $\Theta = \{W_\ell, b_\ell\}_{\ell=1}^{L}$. Suppose layer k is pruned, giving $\widehat{W}_k = W_k + \delta W_k$ and $\widehat{\Theta} = \Theta + \delta\Theta$. Then, for every input state $s \in X$,

$$\big\|\pi(s;\Theta) - \pi(s;\widehat{\Theta})\big\|_2 \;\le\; \|\delta W_k\|_2 \Big(\|s\|_2 \prod_{\substack{\ell=1 \\ \ell\neq k}}^{L} \|W_\ell\|_2 + \sum_{i=1}^{k-1} \prod_{\substack{\ell=i+1 \\ \ell\neq k}}^{L} \|W_\ell\|_2 \|b_i\|_2 \Big) \tag{7}$$

$$\le\; \|\delta W_k\|_2 \underbrace{\Big(\sup_{s\in X} \|s\|_2 \prod_{\substack{\ell=1 \\ \ell\neq k}}^{L} \|W_\ell\|_2 + \sum_{i=1}^{k-1} \prod_{\substack{\ell=i+1 \\ \ell\neq k}}^{L} \|W_\ell\|_2 \|b_i\|_2 \Big)}_{=:\, C_{\max}}. \tag{8}$$

Consequently, the control-error bound in (3) *can be chosen as $B_\pi(\delta\Theta) = C_{\max}\|\delta W_k\|_2$.*

Proof. Because σ_l are 1-Lipschitz, $\|\sigma_l(u)-\sigma_l(v)\|_2 \le \|u-v\|_2$. Write $\delta x_\ell := x_\ell - \widehat{x}_\ell$, then for the last layer

$$\|\pi(s;\Theta) - \pi(s;\widehat{\Theta})\|_2 = \|x_L - \widehat{x}_L\|_2 = \|\delta x_L\|_2,$$

and

$$\begin{aligned}\|\delta x_L\|_2 &= \|\sigma_L(W_L x_{L-1} + b_L) - \sigma_L(W_L \widehat{x}_{L-1} + b_L)\|_2 \\ &\le \|W_L x_{L-1} - W_L \widehat{x}_{L-1}\|_2 \le \|W_L\|_2 \|\delta x_{L-1}\|_2 \\ &\Rightarrow \|\delta x_L\|_2 \le \|W_L\|_2 \|\delta x_{L-1}\|_2.\end{aligned}$$

Iterating the argument down to layer k yields

$$\begin{aligned}\|\delta x_L\|_2 &\le \Big(\prod_{\ell=k+1}^{L} \|W_\ell\|_2\Big)\|W_k x_{k-1} - \widehat{W}_k x_{k-1}\|_2 \\ &= \Big(\prod_{\ell=k+1}^{L} \|W_\ell\|_2\Big)\|\delta W_k x_{k-1}\|_2 \\ &\le \Big(\prod_{\ell=k+1}^{L} \|W_\ell\|_2\Big)\|\delta W_k\|_2 \|x_{k-1}\|_2.\end{aligned}$$

Because σ_l are ReLU-type by Proposition 1

$$\begin{aligned}\|x_{k-1}\|_2 &= \|\sigma_{k-1}(W_{k-1}x_{k-2} + b_{k-1})\|_2 \\ &\le \|W_{k-1}x_{k-2} + b_{k-1}\|_2 \\ &\le \|W_{k-1}\|_2 \|x_{k-2}\|_2 + \|b_{k-1}\|_2 \\ &\le \cdots \le \\ &\le \|s\|_2 \prod_{\ell=1}^{k-1} \|W_\ell\|_2 + \sum_{i=1}^{k-2} \prod_{\ell=i+1}^{k-1} \|W_\ell\|_2 \|b_i\|_2 + \|b_{k-1}\|_2.\end{aligned}$$

Substituting gives

$$\begin{aligned}&\|\pi(s;\Theta) - \pi(s;\widehat{\Theta})\|_2 \\ &\le \Big(\prod_{\ell=k+1}^{L} \|W_\ell\|_2\Big)\|\delta W_k\|_2 \Big(\|s\|_2 \prod_{\ell=1}^{k-1} \|W_\ell\|_2 + \sum_{i=1}^{k-2} \prod_{\ell=i+1}^{k-1} \|W_\ell\|_2 \|b_i\|_2 + \|b_{k-1}\|_2\Big) \\ &\le \|\delta W_k\|_2 \Big(\|s\|_2 \prod_{\substack{\ell=1 \\ \ell\neq k}}^{L} \|W_\ell\|_2 + \sum_{i=1}^{k-1} \prod_{\substack{\ell=i+1 \\ \ell\neq k}}^{L} \|W_\ell\|_2 \|b_i\|_2\Big),\end{aligned}$$

which completes the bound 7. □

Corollary 1. Robustness under pruning *multiple* layers. *Let the assumptions of Theorem 1 hold and let* $S = \{k_1, \dots, k_m\} \subseteq \{1, \dots, L\}$ *be an index set of layers pruned by OBD. For every* $k \in S$ *write* $\widehat{W}_k = W_k + \delta W_k$ *and* $\widehat{\Theta} = \Theta + \delta\Theta$ *with* $\delta\Theta = \{\delta W_k\}_{k\in S}$. *Define, for each* $k \in S$ *and input state* $s \in X$,

$$C_k(s) \;:=\; \|s\|_2 \prod_{\ell \neq k} \|W_\ell\|_2 + \sum_{i=1}^{k-1} \Big(\prod_{\substack{\ell=i+1 \\ \ell \neq k}}^{L} \|W_\ell\|_2 \Big) \|b_i\|_2, \qquad C_{k,\max} := \sup_{s\in X} C_k(s). \tag{9}$$

Then, for every $s \in X$,

$$\big\|\pi(s;\Theta) - \pi(s;\widehat{\Theta})\big\|_2 \;\le\; \sum_{k\in S} \|\delta W_k\|_2 \; C_k(s) \tag{10}$$

$$\le\; \sum_{k\in S} \|\delta W_k\|_2 \; C_{k,\max}. \tag{11}$$

Consequently, a valid control-error budget in (3) *is* $B_\pi(\delta\Theta) = \sum_{k\in S} C_{k,\max} \, \|\delta W_k\|_2$.

Proof. Order the pruned layers as $k_1 < \cdots < k_m$ and construct an intermediate parameter sequence $\Theta^0 := \Theta,\ \Theta^j := \Theta^{j-1} + \delta W_{k_j},\ j = 1, \dots, m$, so that $\Theta^m = \widehat{\Theta}$.

Applying Theorem 1 to the pair (Θ^{j-1}, Θ^j) and layer k_j gives

$$\big\|\pi(s;\Theta^{j-1}) - \pi(s;\Theta^j)\big\|_2 \;\le\; \|\delta W_{k_j}\|_2 \, C_{k_j}(s).$$

Summing these m inequalities and using the triangle inequality yields the first bound in (10). Taking the supremum over s establishes the second bound. □

Corollary 1 shows that control robustness degrades *additively* with respect to the spectral-norm perturbations $\{\delta W_k\}_{k\in S}$. The constants $C_k(s)$ depend only on the unpruned weights, biases, and the input magnitude, making them computable *before* pruning takes place. In practice one may evaluate the tighter, state-dependent bound $C_k(s)$ on a validation set or deploy the uniform constant $C_{k,\max}$ for worst-case guarantees.

5. Conclusion.

We have provided the first closed-form robustness guarantees for second-order network pruning in deterministic, nonlinear discrete-time control. By combining classical Lipschitz bounds with a layer-local analysis of the OBD saliency, we showed that the deviation of the pruned control signal is proportional to the spectral-norm perturbation introduced in each affected layer, and that these deviations accumulate additively across layers. All constants appearing in the bounds depend only on unpruned spectral norms, biases and input magnitude, so they can be evaluated offline from a single forward pass and require no roll-outs, no retraining, no additional hyperparameters.

Limitations. Our analysis is restricted to deterministic dynamics, ReLU-type 1-Lipschitz activations and spectral-norm estimates. It therefore ignores stochastic disturbances, data-dependent curvature. The bounds apply to the instantaneous control signal, not yet to the long-horizon value function or trajectory-level safety constraints.

Future directions. Promising extensions include

- lifting the theory to stochastic or adversarial disturbances,
- deriving return-level or constraint-violation bounds via contraction arguments,
- using the bounds to devise *robustness-aware* pruning schedules that maximise compression under a certified control-error budget.

Bridging these directions will further tighten the link between modern network compression and the rigorous assurance required for autonomous systems.

References

1. Mnih, V., Kavukcuoglu, K., Silver, D., *et al.* (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533.
2. Black, K., Brown, N., Driess, D., *et al.* (2024). π_0: A Vision–Language–Action Flow Model for General Robot Control. *arXiv preprint* arXiv:2410.24164.
3. LeCun, Y., Denker, J., Solla, S. (1989). Optimal Brain Damage. *Advances in Neural Information Processing Systems, 2.*
4. Hassibi, B., Stork, D. (1992). Second-order derivatives for network pruning: Optimal Brain Surgeon. *Advances in Neural Information Processing Systems, 5.*
5. Frantar, E., Alistarh, D. (2023). SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 10323–10337).
6. Gu, S., Yang, L., Du, Y., *et al.* (2022). A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint* arXiv:2205.10330.
7. Nadizar, G., Medvet, E., Pellegrino, F. A., *et al.* (2021). On the effects of pruning on evolved neural controllers for soft robots. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (pp. 1744–1752).
8. Clevert, D. A., Unterthiner, T., Hochreiter, S. (2015). Fast and accurate deep network learning by Exponential Linear Units (ELUs). *arXiv preprint* arXiv:1511.07289.
9. Hendrycks, D., Gimpel, K. (2016). Gaussian Error Linear Units (GELUs). *arXiv preprint* arXiv:1606.08415.
10. Brohan, A., Brown, N., Carbajal, J., *et al.* (2022). RT-1: Robotics Transformer for real-world control at scale. *arXiv preprint* arXiv:2212.06817.
11. Barbara, N. H., Wang, R., Manchester, I. R. (2025). On Robust Reinforcement Learning with Lipschitz-bounded Policy Networks. *arXiv preprint* arXiv:2405.11432.
12. Virmaux, A., Scaman, K. (2018). Lipschitz regularity of deep neural networks: Analysis and efficient estimation. *Advances in Neural Information Processing Systems, 31.*
13. Zhang, B., Jiang, D., He, D., Wang, L. (2022). Rethinking Lipschitz neural networks and certified robustness: A Boolean function perspective. *arXiv preprint* arXiv:2210.01787.
14. Evans, L. C. (2018). *Measure Theory and Fine Properties of Functions.* Routledge.
15. Nair, V., Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning* (pp. 807–814).
16. Maas, A. L., Hannun, A. Y., Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. *Proc. ICML, 30*(1), 3.
17. He, K., Zhang, X., Ren, S., Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1026–1034).

M. Shamrai

Аналітичні межі робастності при прунінгу другого порядку політик нейронних кон-

тролерів.

Глибокі нейронні політики відкрили можливості гнучкого польоту для квадрокоптерів, адаптивного захоплення для маніпуляторів та надійної навігації для наземних роботів, проте їхні мільйони ваг суперечать обмеженням по пам'яті та роботі в реальному часі у вбудованих мікроконтролерах. Методи прунінгу другого порядку – *Optimal Brain Damage* (OBD) та його варіанти, включаючи *Optimal Brain Surgeon* (OBS) та нещодавній SparseGPT – стискають мережі за один прохід, використовуючи локальну матрицю Гессе, досягаючи набагато більшої розрідженості, ніж метод порогового значення величин. Незважаючи на їхній успіх у комп'ютерному зорі та обробці природної мови, наслідки такого видалення ваг для стабільності *замкнутого циклу*, точності відстеження та безпеки залишаються неясними. Ми представляємо перший математично строгий аналіз робастності прунінгу другого порядку в нелінійному дискретному керуванні. Система розвивається під дією неперервного перехідного відображення $f : X \times U \to X$, тоді як контролер є L-шаровим багатошаровим перцептроном з активаціями типу ReLU, які є *глобально* 1-ліпшицевими. Прунінг вагової матриці шару k замінює W_k на $W_k + \delta W_k$, створюючи збурений вектор параметрів $\widehat{\Theta} = \Theta + \delta\Theta$ та запрунену політику $\pi(\cdot; \widehat{\Theta})$. Для кожного вхідного стану s в X ми виводимо аналітичну нерівність $|\pi(s; \Theta) - \pi(s; \widehat{\Theta})\|_2 \leq C_k(s) \, \|\delta W_k\|_2$, де константа $C_k(s)$ залежить *тільки* від незапрунених спектральних норм та зміщень і може бути аналітично оцінена з одного прямого проходу. Коли кілька шарів $S \subseteq \{1, \ldots, L\}$ пруняться, ми показуємо, що відхилення додаються *лінійно*, $\sum_{k \in S} C_k(s) \|\delta W_k\|_2$, що призводить до явної оцінки робастності, що залежить від стану або найгіршого випадку. Усі константи обчислюються *офлайн*. Таким чином, розгортання середовища, зворотне поширення або налаштування гіперпараметрів *не* є необхідними, що робить фреймворк безпосередньо придатним для використання всередині алгоритмів стиснення. Отримані межі визначають, до розгортання в польових умовах, максимально допустиму величину прунінгу, сумісну із заданим порогом помилки керування. Пов'язуючи стиснення мережі другого порядку з гарантіями продуктивності замкнутого циклу, наша робота звужує критичний розрив між сучасними інструментами глибокого навчання та вимогами до надійності критично важливих для безпеки автономних систем.

Ключові слова: *прунінг другого порядку, робастність нейронного контролера, межі спектральної норми, Ліпшицевий багатошаровий перцептрон.*

Institute of Mathematics
National Academy of Sciences of Ukraine
Kyiv, Ukraine
m.shamrai@imath.kiev.ua

Received 27.05.25