UDC 004.8:62-5
DOI: https://doi.org/10.17721/1812-5409.2025/2.24

Maksym SHAMRAI, PhD Student
ORCID ID: 0000-0003-0730-494X
e-mail: m.shamrai@imath.kiev.ua
Institute of mathematics of NAS of Ukraine, Kyiv, Ukraine

NONASYMPTOTIC BOUNDS ON RETURN DEGRADATION FOR OBD-PRUNED NEURAL CONTROLLERS

Deep reinforcement learning (RL) has delivered striking results across domains ranging from games to robotics, yet the resulting controllers frequently comprise millions of parameters – far beyond the memory, latency, and energy budgets of embedded platforms such as quadrotors, mobile manipulators, and on-board microcontrollers. Pruning offers a practical path to deployment by removing parameters while preserving accuracy, but a fundamental question remains open for control:* how much does pruning degrade closed-loop return? *A theory is developed that links* parameter-space perturbations *produced by pruning to* return degradation *in a discounted MDP, without relying on global curvature of the training loss. The starting point is a tight, policy-level inequality: we show that the return gap $|J(\pi') - J(\pi)|$ is controlled by the statewise* total-variation *(TV) distance between the original and pruned policies. This TV-based bound follows directly from the performance-difference lemma and a bounded-advantage argument, and admits a KL variant via Pinsker's inequality. To connect this policy shift to the magnitude of pruning, we provide two complementary routes. First, at a locally optimal policy, a second-order Taylor expansion of the policy probabilities yields an* OBD-style *bound. Second, recognizing that a global Hessian is infeasible for modern models, we invoke a layer-wise robustness theorem for ReLU MLP controllers. Practically, the bound enables pre-pruning budgeting, post-pruning validation, and principled layer allocation. Conceptually, it bridges compression and safe policy improvement: the same TV/KL machinery that underlies trust-region methods now certifies pruning steps in deep RL. Overall, the results provide the first end-to-end, scalable framework to translate pruning actions into* behavior-level *guarantees for deep RL controllers, enabling reliable compression under tight on-board constraints.

K e y w o r d s: ***deep reinforcement learning, neural policies, optimal brain damage pruning, safety certificates, compression.***

AMS 2020 classification: 68T07, 68T05, 93E20, 93B35.

Introduction

From mastering board and video games – Go (Silver et al., 2016), Atari (Mnih et al., 2015), Dota 2 (Berner et al., 2019), and StarCraft II (Vinyals et al., 2019) – to producing agile locomotion in simulation (MuJoCo) (Todorov, Erez, & Tassa, 2012) and dexterous visuomotor manipulation on real robots (Levine et al., 2016; Andrychowicz et al., 2020; Black et al., 2025), recent breakthroughs have been powered by deep reinforcement learning (RL). The policies behind these systems typically contain *hundreds of millions to billions* of parameters, a scale that accelerates learning and robustness but clashes with deployment realities: limited on-board memory, tight real-time latency, and strict energy budgets on platforms such as quadrotors, mobile manipulators, and autonomous vehicles. This mismatch makes model compression – especially parameter pruning – an indispensable step for bringing RL controllers out of the lab and into embedded, safety-critical settings. Second-order pruning methods such as Optimal Brain Damage (OBD) (LeCun, Denker, & Solla, 1989) and modern calibration-based approaches like SparseGPT (Frantar, & Alistarh, 2023) offer effective reductions in model size, yet their impact on *return* remains poorly understood. Motivated by this gap between empirical practice and provable guarantees, we develop bounds that translate parameter perturbations into certified limits on performance, building on the performance-difference framework (Kakade, & Langford, 2002; Schulman et al., 2015) and recent layer-wise robustness analyses (Shamrai, 2025).

The contributions of the paper are following: (i) *a unified performance–difference bound:* we start from the classical Kakade–Langford lemma and derive a tight upper bound on the return drop in terms of the *total-variation (TV)* distance between the original and pruned policies (Theorem 2); (ii) *Second-order link to OBD:* via a Taylor expansion we connect the TV distance to the *diagonal Hessian scores* used by OBD, producing the first explicit *return guarantee* that scales with the standard OBD metric (Corollary 2); (iii) *scalable, Hessian-free certificate:* to avoid a $d \times d$ global Hessian, we incorporate a recent layer-wise robustness theorem and obtain a *layer-local* bound linear in the ℓ_2 norm of the pruned weights (Corollary 3).

Our analysis bridges two previously disconnected areas: *the neural network compression*, which focuses on reducing model size but seldom reasons about downstream control metrics and the *safe RL*, which studies return-sensitive modifications but rarely addresses large-scale pruning (Gu et al., 2024). The resulting framework enables practitioners to prune deep RL policies *with performance certificates*, paving the way for resource-aware yet reliable autonomy.

1. Preliminaries

Fundamental notations are: *discount factor:* $\gamma \in (0, 1)$ controls the importance of future rewards; $\gamma \approx 1$ encourages long-term planning, whereas small γ focuses on immediate gains; *state & action spaces:* $\mathcal{S}$ and $\mathcal{A}$ are *finite* sets of states and actions; *dynamics and rewards:* $P(\, s' \,|\, s, a)$ is the state-transition kernel and the reward is bounded: $0 \leqslant r(s, a) \leqslant R_{\max} < \infty$ for all (s, a); *norms:* bold symbols (e.g. $\mathbf{x}$) denote vectors, $\|\cdot\|_1$ and $\|\cdot\|_2$ are the usual ℓ_1 (sum) and ℓ_2 (Euclidean) norms; *probability simplices:* $\Delta(\mathcal{A})$ denotes the set of all distributions over $\mathcal{A}$; *we treat probability distributions as column vectors whose entries sum to* 1 (both norm and linear algebra notation apply seamlessly).

Definition 1 (Markov Decision Process). A *Markov Decision Process* (MDP) is the tuple $\langle \mathcal{S}, \mathcal{A}, P, r, \gamma, \rho \rangle$, where ρ is the distribution of the initial state s_0.

Definition 2 (Policy). A *policy* is any measurable mapping $\pi : \mathcal{S} \to \Delta(\mathcal{A}),\ s \mapsto \pi(\cdot \,|\, s)$. We write $a \sim \pi(\cdot \,|\, s)$ when sampling an action from π in state s.

A *pruned policy* π' is obtained from a given neural policy π_θ by setting a subset of network parameters to zero using *Optimal Brain Damage* (OBD) (LeCun, Denker, & Solla, 1989). We use $\pi' = \pi_{\theta'}$ to emphasize the new parameters θ'.

Definition 3 (Value, action-value, and advantage). For any policy π and state-action pair (s, a),

$$V^\pi(s) := \mathbb{E}_\pi\Big[\textstyle\sum_{t=0}^\infty \gamma^t r(s_t, a_t) \;\Big|\; s_0 = s\Big], \quad Q^\pi(s,a) := \mathbb{E}_\pi\Big[\textstyle\sum_{t=0}^\infty \gamma^t r(s_t, a_t) \;\Big|\; s_0 = s, a_0 = a\Big],$$

$$A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s).$$

Intuition. $V^\pi(s)$ is the expected *return* starting from s when following π; $Q^\pi(s,a)$ is the same, assuming we first force action a. The *advantage* A^π measures how much better (or worse) a is compared to the average prescribed by π in s.

Definition 4 (Discounted visitation distribution). The *discounted state visitation* under π is

$$d^\pi(s) \;:=\; (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \,\mathsf{Pr}(s_t = s \mid \pi, \rho),$$

which places more weight on states visited earlier in an episode.

Definition 5 (Expected return (performance)). Given an initial-state distribution ρ, the (discounted) expected return of policy π is

$$J(\pi) \;:=\; \mathbb{E}_{s_0\sim\rho}\big[V^\pi(s_0)\big] \;=\; \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d^\pi,\, a\sim\pi(\cdot|s)}\big[r(s,a)\big],$$

where the second equality is a standard identity obtained by unrolling the definition of d^π in Def. 4. For a pruned policy π', we write $J(\pi')$ analogously.

Definition 6 (Total variation and KL divergence). For distributions $p, q \in \Delta(\mathcal{A})$,

$$D_{\mathsf{TV}}(p,q) := \tfrac{1}{2}\sum_{a\in\mathcal{A}} |p(a) - q(a)|, \qquad D_{\mathsf{KL}}(p\|q) := \sum_{a\in\mathcal{A}} p(a)\log\frac{p(a)}{q(a)}.$$

Lemma 1 (Performance–difference lemma (Kakade, & Langford, 2002)). *For any two policies π, π',*

$$J(\pi') - J(\pi) \;=\; \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d^{\pi'}}\mathbb{E}_{a\sim\pi'}\big[A^\pi(s,a)\big].$$

We parameterize the policy by $\theta \in \mathbb{R}^d$ and write π_θ. *Pruning* removes parameters to reduce memory and inference cost. Formally, let $m \in \{0,1\}^d$ be a binary *mask* and define the pruned parameters $\theta' := m \odot \theta$ (Hadamard product). The parameter perturbation is $\delta := \theta' - \theta = (m - \mathbf{1}) \odot \theta$, so that $\delta_i = -\theta_i$ for pruned indices and 0 otherwise. For layered networks we also use layerwise notation $\{W_\ell, b_\ell\}_{\ell=1}^L$ and masks M_ℓ so that $\widehat{W}_\ell := M_\ell \odot W_\ell$.

Optimal Brain Damage (OBD) selects weights to prune by minimizing a second-order approximation to the post-pruning loss. If $h_i \geqslant 0$ denotes the ith diagonal entry of the Hessian of the loss (or a suitable layer-local surrogate), the *saliency* of weight θ_i is $s_i \;:=\; h_i\,\theta_i^2$. Under a global sparsity budget $\tau \in (0,1)$ (or a target of k weights), OBD prunes the indices with the *smallest* s_i, yielding θ' that (approximately) minimizes the predicted loss increase subject to the budget. In the sequel we relate this parameter change δ (or layerwise $\|\delta W_k\|_2$) to changes in the *behavior* of the policy, quantified by $D_{\mathsf{TV}}\big(\pi'(\cdot \mid s), \pi(\cdot \mid s)\big)$ and ultimately by the return gap $|J(\pi') - J(\pi)|$.

To connect small parameter perturbations to changes in the controller's outputs, we rely on the following *deterministic* robustness result for single-layer pruning in feed-forward ReLU networks. It upper-bounds how a perturbation of layer k propagates to the policy output via a product of spectral norms of the *unpruned* layers.

Theorem 1 (Robustness of an OBD-pruned policy (Shamrai, 2025)). *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 $s \in X$,

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

$$\leqslant\; \|\delta W_k\|_2 \underbrace{\Big(\sup_{s\in X}\|s\|_2 \prod_{\substack{\ell=1\\ \ell\neq k}}^{L}\|W_\ell\|_2 + \sum_{i=1}^{k-1}\prod_{\substack{\ell=i+1\\ \ell\neq k}}^{L}\|W_\ell\|_2\|b_i\|_2\Big)}_{=:\; C_{\mathsf{max}}}. \tag{2}$$

Consequently, the control-error bound in (1) *can be summarized as* $B_\pi(\delta\Theta) = C_{\max}\|\delta W_k\|_2$.

2. Results

Theorem 2 (Return difference via total variation)**.** *In any finite MDP with* $r \in [0, R_{\max}]$ *and discount* γ*,*

$$|J(\pi') - J(\pi)| \;\leqslant\; \frac{2R_{\max}}{(1-\gamma)^2}\, \mathbb{E}_{s\sim d^{\pi'}}\Big[D_{\mathsf{TV}}\big(\pi'(\cdot\,|\,s),\, \pi(\cdot\,|\,s)\big)\Big].$$

Corollary 1 (KL–based version)**.** *Combining Theorem 2 with Pinsker's inequality gives*

$$|J(\pi') - J(\pi)| \;\leqslant\; \frac{2R_{\max}}{(1-\gamma)^2}\, \mathbb{E}_{s\sim d^{\pi'}}\Big[\sqrt{\tfrac{1}{2}D_{\mathsf{KL}}\big(\pi'\|\pi\big)}\Big].$$

A simpler (conservative) worst-case bound is obtained by replacing the expectation with a supremum over s*.*

Corollary 2 (OBD performance bound)**.** *Let* $\pi_{\theta^\star}$ *be a locally optimal network policy with vanishing first derivatives, that is* $\nabla_\theta \pi_{\theta^\star}(a\,|\,s) = 0$*. Prune weights* $\mathcal{P} \subset [d]$ *to obtain* $\theta' = \theta^\star + \delta$*,* $\delta_i = -\theta_i^\star$ *for* $i \in \mathcal{P}$*. If each* $\theta \mapsto \pi_\theta(a\,|\,s)$ *is* $\mathcal{C}^3$ *with bounded third derivatives, then*

$$|J(\pi_{\theta'}) - J(\pi_{\theta^\star})| \;\leqslant\; \frac{R_{\max}}{2(1-\gamma)^2}\, S_{\mathsf{OBD}} + \mathcal{O}(\|\delta\|_2^3),$$

where $S_{\mathsf{OBD}} := \sum_{i\in\mathcal{P}}\Big[\sum_{(s,a)} h_i^{(a,s)}\Big]\theta_i^{\star\,2}$ *is the sum of* diagonal Hessian scores *used by OBD (LeCun, Denker, & Solla, 1989).*

Intuition. S_{OBD} is precisely what OBD minimises; the corollary states that, for small enough pruning, the drop in return is *linear* in that score.

The diagonal entries $h_i^{(a,s)}$ in Corollary 2 come from the *global* Hessian $H_{a,s} \in \mathbb{R}^{d\times d}$, where $d \approx 10^8$ for contemporary vision or language policies. Even storing $H_{a,s}$ is infeasible, let alone computing it for every (s,a). SparseGPT (Frantar, & Alistarh, 2023) circumvents this by optimising a simple quadratic objective $\frac{1}{2}\|XW - Y\|_F^2$ *layer by layer*, whose Hessian $X^\top X$ is small and easy to invert. Thus curvature is captured *locally* while memory scales linearly in the number of parameters.

Corollary 3 (Return bound via layer–local TV (single pruned layer))**.** *Assume the setting of Theorem 1 and that the policy head mapping the final network output to action probabilities is* 1*–Lipschitz in* ℓ_2 *(e.g., a softmax over logits) so that* $\pi(\cdot\mid s), \pi'(\cdot\mid s) \in \Delta(\mathcal{A})$ *with* $A := |\mathcal{A}| < \infty$*. Then*

$$\big|J(\pi') - J(\pi)\big| \;\leqslant\; \frac{R_{\max}\sqrt{A}\, C_{\max}}{(1-\gamma)^2}\, \big\|\delta W_k\big\|_2.$$

The corollary leverages the layer–wise output perturbation control from Theorem 1 to upper bound the *statewise* policy shift in total variation, and then propagates this shift to a guaranteed return bound via Theorem 2. The certificate is: (i) *linear* in the pruning magnitude $\|\delta W_k\|_2$; (ii) *dimension–free* with respect to the total parameter count; and (iii) dependent on the action–space size only through $\sqrt{A}$. A less conservative, distributional variant replaces $\sup_s$ by $\mathbb{E}_{s\sim d^{\pi'}}[\cdot]$, yielding the same scaling but with $C_{\max}$ multiplied by an average instead of a worst–case factor. For multiple pruned layers, a triangle–inequality extension gives $\|\pi' - \pi\|_2 \leqslant \sum_{k\in\mathcal{K}} C_{\max}^{(k)}\|\delta W_k\|_2$, leading to an additive bound on the return gap.

Discussions and conclusions

We studied the effect of pruning on the *behavior* of deep RL policies, providing performance guarantees that are both interpretable and scalable. Our analysis starts from a policy-level quantity and ends with a certified bound on the return gap, avoiding any direct reasoning about nonconvex training losses. We derived a TV-based performance bound (Theorem 2) that upper-bounds $|J(\pi') - J(\pi)|$ through the state-wise policy shift $D_{\mathsf{TV}}(\pi', \pi)$, with a KL variant via Pinsker (Corollary 1), connected TV to second-order parameter perturbations at an optimal policy, giving an OBD-style certificate (Corollary 2), eliminated the intractable global Hessian by invoking a layer-wise robustness result (Theorem 1), which yields a tight, *layer-local* estimate of probability shift; combined with Theorem 2, this produces a practical, dimension-free return bound (Corollary 3). Together, these results form a pipeline

$$\text{pruning magnitude } (\|\delta W_k\|_2) \;\Rightarrow\; D_{\mathsf{TV}}(\pi', \pi) \;\Rightarrow\; |J(\pi') - J(\pi)|$$

that scales to modern policies without global curvature estimation.

The bound depends on (a) the pruning magnitude $\|\delta W_k\|_2$, (b) a computable robustness constant $C_{\max}$ (product of layer spectral norms and input bounds), and (c) the action-space factor $\sqrt{A}$. Each quantity can be estimated once per model (via power iteration and a calibration bound on $\|s\|_2$), enabling *pre-pruning* budgeting and *post-pruning* validation with negligible overhead. The certificate is additive across pruned layers by triangle inequality, which supports iterative or structured pruning schedules.

Our guarantees are worst-case and may be conservative if $C_{\max}$ is dominated by a small number of large spectral norms or by a loose input bound $\sup_{s\in X}\|s\|_2$. The $\sqrt{A}$ dependence is inevitable for $\ell_2 \to \ell_1$ conversion, but can overstate degradation in highly peaked policies. Finally, the theory assumes a 1-Lipschitz policy head (e.g., softmax or bounded continuous control) and does not exploit environment-specific mixing or structure beyond bounded rewards.

Acknowledgments. The author expresses his gratitude to the Armed Forces of Ukraine for their protection, which has made this research possible.

Sources of funding. The author 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.

References

Andrychowicz, O. M., Baker, B., Chociej, M., Józefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., Schneider, J., Sidor, S., Tobin, J., Welinder, P., Weng, L., & Zaremba, W. (2020). Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1), 3–20. https://doi.org/10.1177/0278364919887447

Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., Józefowicz, R., Gray, S., Olsson, C., Pachocki, J., Petrov, M., de Oliveira Pinto, H. P., Raiman, J., Salimans, T., Schlatter, J., … Zhang, S. (2019). *Dota 2 with large scale deep reinforcement learning.* https://doi.org/10.48550/arXiv.1912.06680

Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Galliker, M. Y., Ghosh, D., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., LeBlanc, D., Levine, S., … Zhilinsky, U. (2025). $\pi_{0.5}$*: a vision-language-action model with open-world generalization.* https://doi.org/10.48550/arXiv.2504.16054

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.* JMLR.org. https://dl.acm.org/doi/10.5555/3618408.3618822

Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., & Knoll, A. (2024). A review of safe reinforcement learning: Methods, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 11216-11235. https://doi.org/10.1109/TPAMI.2024.3457538

Kakade, S., & Langford, J. (2002). Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning* (p. 267–274). Morgan Kaufmann Publishers Inc. https://dl.acm.org/doi/10.5555/645531.656005

LeCun, Y., Denker, J., & Solla, S. (1989). Optimal brain damage. In Touretzky (Ed.), *Advances in Neural Information Processing Systems* (Vol. 2). Morgan-Kaufmann. https://shorturl.at/Sr2Ve

Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(1), 1334–1373. https://dl.acm.org/doi/10.5555/2946645.2946684

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M. A., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529–533. https://doi.org/10.1038/nature14236

Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. In Bach & Blei (Eds.), *Proceedings of the 32nd International Conference on Machine Learning* (Vol. 37, pp. 1889–1897). PMLR. https://proceedings.mlr.press/v37/schulman15.html

Shamrai, M. (2025). *Closed-form robustness bounds for second-order pruning of neural controller policies.* https://doi.org/10.48550/arXiv.2507.02953

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587), 484–489. https://doi.org/10.1038/nature16961

Todorov, E., Erez, T., & Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems* (p. 5026-5033). https://doi.org/10.1109/IROS.2012.6386109

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A. J., Chung, J., Choi, D., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., … Silver, D. (2019). Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782), 350–354. https://doi.org/10.1038/s41586-019-1724-z

Отримано редакцією журналу / Received: 13.05.25
Прорецензовано / Revised: 24.09.25
Схвалено до друку / Accepted: 10.10.25

Максим ШАМРАЙ, асп.
ORCID ID: 0000-0003-0730-494X
e-mail: m.shamrai@imath.kiev.ua
Інститут математики НАН України, Київ, Україна

НЕАСИМПТОТИЧНІ МЕЖІ ПОГІРШЕННЯ ДИСКОНТОВАНОЇ КУМУЛЯТИВНОЇ ВИНАГОРОДИ ДЛЯ OBD-ПРОРІДЖЕНИХ НЕЙРОННИХ КОНТРОЛЕРІВ

Останнім часом глибоке навчання з підкріпленням (RL) продемонструвало приголомшливі результати в галузях від ігор до робототехніки, однак отримані контролери часто містять мільйони параметрів – значно більше за обмеження пам'яті, затримки й енергоспоживання вбудованих платформ на кшталт квадрокоптерів, мобільних маніпуляторів і бортових мікроконтролерів. Проріджения (pruning) пропонує практичний шлях до розгортання, вилучаючи параметри без втрати точності, але для систем керування лишається відкрите фундаментальне запитання: **наскільки прорідження погіршує замкнену (closed-loop) винагороду?** ***У пропонованій роботі розроблено теорію, що пов'язує збурення у просторі параметрів, спричинені прорідженням, із погіршенням винагороди в дисконтовній МППР (MDP), не покладаючись на глобальну кривизну функції втрат під час навчання. Відправною точкою є «щільна» нерівність на рівні політики: показано, що розрив у винагороді*** $|J(\pi') - J(\pi)|$ ***визначається*** **варіаційною** ***(TV) відстанню між початковою та проріджeною політиками для кожного стану. Ця межа на основі TV безпосередньо випливає з леми про різницю продуктивності й оцінки обмеженої переваги, а також має KL-варіант через нерівність Пінскера. Щоб пов'язати цю зміну політики з величиною прорідження, ми пропонуємо два комплементарні шляхи. По-перше, у разі локально оптимальної політики розклад Тейлора другого порядку для ймовірностей політики дає*** **межу у стилі OBD.** ***По-друге, визнаючи, що глобальний гессіан є непрактичним для сучасних моделей, ми залучаємо теорему про пошарову робастність для контролерів ReLU MLP. На практиці запропонована межа уможливлює планування бюджету перед прорідженням, перевірку після прорідження та принциповий розподіл ступеня прорідження між шарами. На концептуальному рівні вона поєднує компресію та безпечне поліпшення політики: той самий апарат TV/KL, що лежить в основі trust-region методів, тепер надає сертифікацію кроків прорідження у глибокому RL. Загалом отримані результати пропонують першу наскрізну, масштабовану базу, що перетворює дії з прорідження на*** **гарантії на рівні поведінки** ***для глибоких RL-контролерів, забезпечуючи надійну компресію за жорстких бортових обмежень.***

Ключові слова: ***глибоке навчання з підкріпленням, нейронні політики, прорідження optimal brain damage, сертифікати безпеки, стиснення.***

Автор заявляє про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження(у зборі, аналізі чи інтерпретації даних, якщо це мало місце), у написанні рукопису та в рішенні про публікацію результатів.

The author declares no conflicts of interest. The funders had no role in the design of the study (in the collection, analyses or interpretation of data if applicable), in the writing of the manuscript as well as in the decision to publish the results.