Nesterov Momentum Adversarial Perturbations in the Deep Reinforcement Learning Domain

Ezgi Korkmaz [1]

Abstract

Deep reinforcement learning algorithms achieved significant success in the last five years and gave birth to a new research area. Currently, deep reinforcement learning algorithms have been deployed in many different fields from resource management to deep neural architecture design. Exploring the generalization capabilities of deep reinforcement learning algorithms still remains an active area of research. One way to asses the generalization of deep reinforcement learning agents lies in investigating their reactions to adversarial perturbations. In this paper, we propose a new Nesterov momentum based optimization to find adversarial perturbations for the deep reinforcement learning domain. We show in our experiments our Nesterov momentum based approach achieves state-of-the-art results in various games from the Atari environment. We believe our proposed approach can be an important initial step in the robustification of deep reinforcement learning agents.

1. Introduction

Deep Neural Networks (DNN)s have gained significant momentum both in research and applications, and are currently utilized in many different domains such as natural language processing Sutskever et al. (2014), speech recognition Hannun et al. (2014), image recognition Krizhevsky et al. (2012), and self learning systems ((Mnih et al., 2015), Schulman et al. (2017), Lillicrap et al. (2015)). Reinforcement learning algorithms, in particular, have seen dramatic recent improvements with the introduction of DNNs as function approximators (Mnih et al., 2015). Following this initial work, deep reinforcement learning has been applied in a variety of areas including: robotics (Gu et al. (2017), Kalashnikov et al. (2018)), autonomous driving Dosovitsky et al. (2017), auction bidding Wang et al. (2017), deep neural network architecture design Baker et al. (2016), network system control (Jay et al. (2019),Chinchali et al. (2018)), grid operation and security (Duan et al. (2019), Huang et al. (2019)), financial trading Noonan (2017), blockchain protocol security Hou et al. (2019), natural language processing (He et al. (2016), Jaques et al. (2017), Wang et al. (2018)), medical treatment and diagnosis (Yauney & Pratik (2018) , Suchi (2018), Popova et al. (2018), Thananjeyan et al. (2017), Daochang & Jiang (2018), Ghesu et al. (2017)).

Due to the wide applicability of deep reinforcement learning, understanding the robustness and generalization properties of these algorithms is crucial. In particular, a key question is how these algorithms react in the presence of an adversary. Initially, Szegedy et al. (2014) showed that DNNs used for image classification can be made to fail by adding visually imperceptible perturbations to the input image. Further research by Goodfellow et al. (2015) explained the presence of these imperceptible adversarial perturbations by arguing that DNNs learn approximately linear functions. The authors also introduced a new efficient method to compute adversarial perturbations, and demonstrated that including these perturbations in the cost function for DNN training improves robustness. Later Madry et al. (2017) showed that computing optimal adversarial perturbations is a key step in their robust optimization approach to training DNNs resistant to adversarial perturbations. In particular, they demonstrated that better algorithms for computing adversarial examples used in training leads to more robust networks.

We believe adversarial formulations are a way to asses generalization capabilities of deep reinforcement learning agents and are an initial step towards building robust and reliable agents. For these reasons, in this work we focus on adversarial formulations in deep reinforcement learning and make the following contributions:

- We propose Nesterov momentum based optimization to compute adversarial perturbations in the deep reinforcement learning domain.
- We run multiple experiments in the Atari environment

[1]Electrical Engineering and Computer Science School, KTH Royal Institute of Technology, Stockholm, Sweden. Correspondence to: Ezgi Korkmaz <ezgikorkmazk@gmail.com>.

Presented at the ICML 2020 Inductive Biases, Invariances and Generalization in Reinforcement Learning Workshop.

in various games to compare state-of-the-art adversarial formulations and our proposed momentum based approach.

- We demonstrate the momentum based approach can reach higher impact levels with a lower bound on the adversarial perturbation.

2. Related Work and Backgorund

2.1. Crafting Adversarial Perturbations

Szegedy et al. (2014) proposed to minimize the distance between the original image and adversarially produced image to create adversarial perturbations. The authors used box-constrained L-BFGS to solve this optimization problem.

$$\arg\min_{x_{\text{adv}}} = c \cdot \|x_{\text{adv}} - x\| - J(x_{\text{adv}}, y) \tag{1}$$

Here x is the input, y is the output label, and $J(x, y)$ is the cost function for image classification. Goodfellow et al. (2015) introduced the fast gradient method (FGM)

$$x_{\text{adv}} = x + \epsilon \cdot \frac{\nabla_x J(x, y)}{||\nabla_x J(x, y)||_p}, \tag{2}$$

for crafting adversarial examples in image classification by taking the gradient of the cost function $J(x, y)$ used to train the neural network in the direction of the input. Kurakin et al. (2016) proposed an iterative fast gradient method (I-FGM) by applying FGM multiple time with small step size and clipping the pixel values of intermediate updates after each step to guarantee they are in the ϵ-ball.

$$x^0_{\text{adv}} = x, \tag{3}$$

$$x^{N+1}_{\text{adv}} = \text{clip}_\epsilon(x^N_{\text{adv}} + \alpha \text{sign}(\nabla_x J(x^N_{\text{adv}}, y))) \tag{4}$$

Dong et al. (2018) suggested an iterative fast gradient method employing a momentum approach called MI-FGSM to create adversarial perturbations in the image classification domain. Furthermore, the authors increase the transferability of the adversarial examples by utilizing MI-FGSM, and an ensemble of models to produce adversarial perturbations.

Finally, Carlini & Wagner (2017) introduced targeted attacks in the image classification domain based on distance minimization between the adversarial image and the original image while targeting a particular label. In the deep reinforcement learning domain the Carlini & Wagner (2017) formulation is

$$\min_{s_{\text{adv}} \in D_{\epsilon,p}(s)} \quad \|s_{\text{adv}} - s\|_p$$
$$\text{subject to} \quad a^*(s) \neq a^*(s_{\text{adv}}),$$

where s is the unperturbed input, s_{adv} is the adversarially perturbed input, $a^*(s)$ is the action taken in the unperturbed state, and $a^*(s_{\text{adv}})$ is the action taken in the adversarial state. This formulation is basically the minimization of the distance to the adversarial state constrained to states leading to sub-optimal actions as determined by the Q-network.

2.2. Adversarial Deep Reinforcement Learning

Huang et al. (2017) and Kos & Song (2017) concurrently proposed the first adversarial attacks on deep reinforcement learning agents by using adversarial examples crafted by FGSM. Mandlekar et al. (2017) proposed a physically plausible threat model and used FGSM perturbations to make the agent's policy more robust. Pattanaik et al. (2018) proposed a projected gradient based method to maximize the probability of the worst possible action in the given state. Lin et al. (2017) considered the timing perspective of the adversarial attacks by using the Carlini & Wagner (2017) formulation. Pinto et al. (2017) use a two player zero-sum discounted Markov game to model the interaction of the adversary and the agent. The authors train the agent in this game with the adversary to increase robustness of the agent. Gleave et al. (2020) modeled the relationship between the agent and the adversary as a two player Markov game and solved it via Proximal Policy Optimization proposed by Schulman et al. (2017). In this work the authors focus on letting the adversary take natural actions in the environment instead of injecting ℓ_p-norm bounded perturbations.

3. Nesterov Momentum- FGM

In this work we propose a Nesterov momentum-based method Nesterov (1983) to find the adversarial perturbation.

Algorithm 1 Nesterov Momentum-FGM

Input: Loss function J, the bound on perturbation ϵ, actions a, states s, iterations T and decay factor μ.
Output: Adversarially perturbed state s_{adv} with $\|s - s_{\text{adv}}\|_2 \leq \epsilon$
$\alpha = \epsilon/T$; $v_0 = 0$; $s^0_{\text{adv}} = 0$
for $t = 1$ **to** T **do**
  Calculate $\nabla_{s_{\text{adv}}} J(s^t_{\text{adv}} + \mu \cdot v_t, a)$
  $v_{t+1} = \mu \cdot v_t + \dfrac{\nabla_{s_{\text{adv}}} J(s^t_{\text{adv}} + \mu \cdot v_t, a)}{\|\nabla_{s_{\text{adv}}} J(s^t_{\text{adv}} + \mu \cdot v_t, a)\|_1}$
  $s^{t+1}_{\text{adv}} = s^t_{\text{adv}} + \alpha \cdot \dfrac{v_{t+1}}{\|v_{t+1}\|_2}$
end for
Return: $s_{\text{adv}} = s^T_{\text{adv}}$

In Nesterov momentum the accumulated gradients,

$$\nabla_{s_{\text{adv}}} J(s^t_{\text{adv}} + \mu \cdot v_t, a) \tag{5}$$

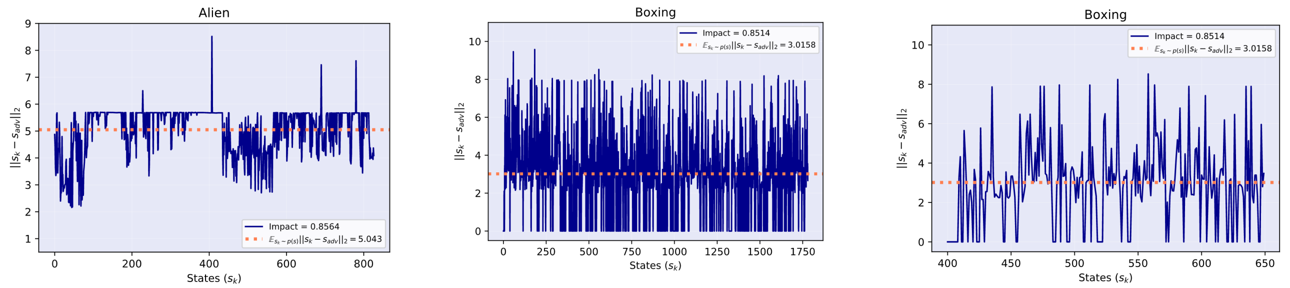

Figure 1. Left: Alien $\|s - s_{\text{adv}}\|_2$ adversarial perturbation computed by Carlini & Wagner (2017) over the states. Middle: Boxing $\|s - s_{\text{adv}}\|_2$ adversarial perturbation computed by Carlini & Wagner (2017) over the states. Right: Zoomed on states between $[400, 650]$ of Boxing $\|s - s_{\text{adv}}\|_2$ adversarial perturbation computed by Carlini & Wagner (2017) over the states.

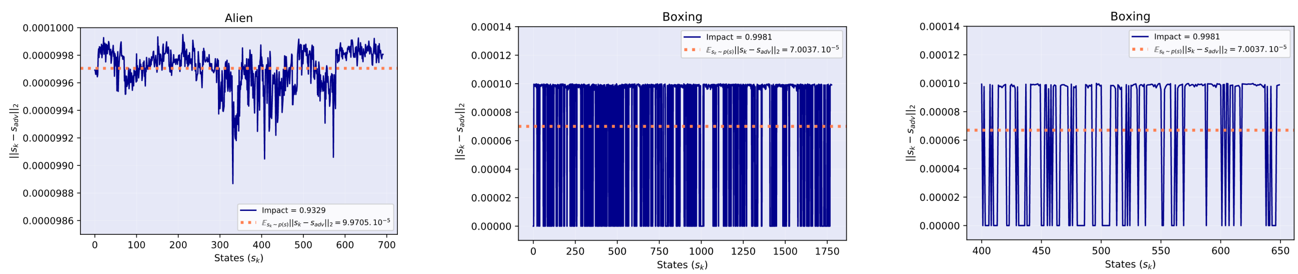

Figure 2. Left: Alien $\|s - s_{\text{adv}}\|_2$ adversarial perturbation computed by Nesterov-MFGM over the states. Middle: Boxing $\|s - s_{\text{adv}}\|_2$ adversarial perturbation computed by Nesterov-MFGM over the states. Right: Zoomed on states between $[400, 650]$ of Boxing $\|s - s_{\text{adv}}\|_2$ adversarial perturbation computed by Nesterov-MFGM over the states.

will be computed at the point $s^t_{\text{adv}} + \mu \cdot v_t$. Note that μ is the decay factor and v_t is the accumulated gradients. The Algorithm itself is presented in 1.

4. Experiments

4.1. Experimental Setup

In our experiments we averaged over 10 episodes for 3 Atari games Bellemare et al. (2013) from the Open AI gym environment Brockman et al. (2016). Our agents are trained with Double-DQN Wang et al. (2016). We normalize the average return of the agents when we calculate the impact of the adversary as follows. Let R_{max} be the average return for the agent who always chooses the best action in a given state, let R_{min} be the average return for the agent who always chooses the worst possible action in a given state, and let R_{a} be the average return of the agent under the influence of the adversary. We define the impact,

$$\text{I} = \frac{\text{R}_{\text{max}} - \text{R}_{\text{a}}}{(\text{R}_{\text{max}} - \text{R}_{\text{min}})} \tag{6}$$

Intuitively this normalization measures how much the adversary degrades the performance of the agent when compared with a worst-case agent which always chooses the worst possible action. It is important to take this worst-case agent into account, because the agent still collects non-zero stochastic rewards from the environment even when always choosing the worst possible action.

The comparison of MI-FGM and our proposed optimization algortihm Nesterov-MFGM for adversarial formulation in deep reinforcement learning is shown in Table 1. It can be seen that our proposed algorithm has a higher mean and lower standard deviation impact on the agent compared to the state-of-the-art.

Table 1. Attack impacts for MI-FGM and Nesterov-MFGM with ℓ_2 norm bound, $\epsilon = 10^{-4}$.

Games	MI-FGM	Nesterov-MFGM
Alien	0.8240±0.23	0.9423±0.03
Bankheist	0.9030±0.05	0.9468±0.02
Boxing	0.5904±0.29	0.6798±0.26

4.2. ℓ_2-norm Bound Comparison

In this section we will compare the state-of-the-art targeted attack Carlini & Wagner (2017) and our proposed Nesterov-MFGM adversary. The adversarial perturbation in ℓ_2-norm, $\|s - s_{\text{adv}}\|_2$, over the states computed by Carlini & Wagner (2017) formulation is shown in Figure 2. Table 2 and Table 3 show the corresponding impact value for the given perturbation profile, and the mean of the ℓ_2-norm adversarial perturbation over the states. Note that in Table 2 and 3 Carlini & Wagner (2017) achieves less impact yet requires orders of magnitude larger adversarial perturbations.

Table 2. Attack impacts and $\|s - s_{\text{adv}}\|_2$ for Carlini & Wagner (2017) and Nesterov-MFGM in Alien.

Alien	Carlini & Wagner (2017)	Nesterov-MFGM
Impact	0.8240	0.9329
$\|s - s_{\text{adv}}\|_2$	5.043	$9.9705 \cdot 10^{-5}$

Table 3. Attack impacts and $\|s - s_{\text{adv}}\|_2$ for Carlini & Wagner (2017) and Nesterov-MFGM in Boxing.

Boxing	Carlini & Wagner (2017)	Nesterov-MFGM
Impact	0.8564	0.9981
$\|s - s_{\text{adv}}\|_2$	3.0158	$7.0037 \cdot 10^{-5}$

5. Conclusion

In this paper we proposed a new optimization method based on Nesterov momentum for computing adversarial examples for deep reinforcement learning. We show in various games from the Atari environment that our proposed approach achieves higher impact compared to the state-of-the-art. Investigating the computation of adversarial perturbations is an important first step in designing robust deep reinforcement learning algorithms. Furthermore, we believe our optimization method can be instrumental in adversarial training algorithms for deep reinforcement learning.

References

Baker, B., Gupta, O., Naik, N., and Raskar, R. Designing neural network architectures using reinforcement learning. *arXiv preprint arXiv:1611.02167*, 2016.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research.*, pp. 253279, 2013.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv:1606.01540*, 2016.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. *In 2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017.

Chinchali, S., Hu, P., Chu, T., Sharma, M., Bansal, M., Misra, R., Pavone, M., and Katti, S. Cellular network traffic scheduling with deep reinforcement learning. *In Thirty-Second AAAI Conference on Artificial Intelligence.*, 2018.

Daochang, L. and Jiang, T. Deep reinforcement learning for surgical gesture segmentation and classification. *In International conference on medical image computing and computer-assisted intervention.*, pp. 247–255.Springer, Cham, 2018.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., and Li, J. Boosting adversarial attacks with momentum. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

Dosovitsky, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. Carla: An open urban driving simulator. *In Proceedings of the Conference on Robot Learning (CoRL)*, 78:1–16, 2017.

Duan, J., Shi, D., Diao, R., Li, H., Wang, Z., Zhang, B., Bian, D., and Yi, Z. Deep-reinforcement-learning-based autonomous voltage control for power grid operations. *IEEE Transactions on Power Systems*, 2019.

Ghesu, F.-C., Georgescu, B., Zheng, Y., Grbic, S., Maier, A., Hornegger, J., and Comaniciu, D. Multi-scale deep reinforcement learning for real-time 3d-landmark detection in ct scans. *IEEE transactions on pattern analysis and machine intelligence 41*, pp. 176–189, 2017.

Gleave, A., Dennis, M., Wild, C., Neel, K., Levine, S., and Russell, S. Adversarial policies: Attacking deep reinforcement learning. *International Conference on Learning Representations ICLR*, 2020.

Goodfellow, I., Shelens, J., and Szegedy, C. Explaning and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

Gu, S., Holly, Ethan andLillicrap, T., and Levine, S. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. *In 2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3389–3396, 2017.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Greg, D., Else, E., Prenger, R., Satheesh, S., Shubho, S., Coates, A., and Ng, A. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

He, J., Chen, J., He, X., Gao, J., Li, L., Deng, L., and Ostendorf, M. Deep reinforcement learning with a natural language action space. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1621–1630, 2016.

Hou, C., Zhou, M., Ji, Y., Daian, P., Tramer, F., Fanti, G., and Juels, A. Squirrl: Automating attack discovery on blockchain incentive mechanisms with deep reinforcement learning. *arXiv preprint arXiv:1912.01798*, 2019.

Huang, Q., Huang, R., Hao, W., Jie, T., Fan, R., and Huang, Z. Adaptive power system emergency control using deep reinforcement learning. *IEEE Transactions on Smart Grid (2019).*, 2019.

Huang, S., Papernot, N., Goodfellow, Ian an Duan, Y., and Abbeel, P. Adversarial attacks on neural network policies. *Workshop Track of the 5th International Conference on Learning Representations*, 2017.

Jaques, N., Gu, S., Bahdanau, D., Miguel Hernndez-Lobato, J., E. Turner, R., and Eck, D. Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control. *In Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1645–1654, 2017.

Jay, N., Rotman, N., Godfrey, Brighten Schapira, M., and Tamar, A. A deep reinforcement learning perspective on internet congestion control. *In International Conference on Machine Learning*, pp. 3050–3059, 2019.

Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., and Levine, S. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*, 2018.

Kos, J. and Song, D. Delving into adversarial attacks on deep policies. *International Conference on Learning Representations*, 2017.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXivpreprint arXiv:1509.02971*, 2015.

Lin, Y.-C., Zhang-Wei, H., Liao, Y.-H., Shih, M.-L., Liu, i.-Y., and Sun, M. Tactics of adversarial attack on deep reinforcement learning agents. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 3756–3762, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mandlekar, A., Zhu, Y., Garg, A., Fei-Fei, L., and Savarese, S. Adversarially robust policy learning: Active construction of physically-plausible perturbations. *In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3932–3939, 2017.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, a. G., Graves, A., Riedmiller, M., Fidjeland, A., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518:529533, 2015.

Nesterov, Y. E. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *In Dokl. akad. nauk Sssr*, 269:543–547, 1983.

Noonan, L. Jpmorgan develops robot to execute trades. *Financial Times*, pp. 19281937, July 2017.

Pattanaik, A., Tang, Z., Liu, S., and Gautham, B. Robust deep reinforcement learning with adversarial attacks. *In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2040–2042, 2018.

Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. Robust adversarial reinforcement learning. *International Conference on Learning Representations ICLR*, 2017.

Popova, M., Isayev, O., and Tropsha, A. Deep reinforcement learning for de novo drug design. *Science advances 4*, 78, 2018.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv:1707.06347v2 [cs.LG].*, 2017.

Suchi, S. Individualized sepsis treatment using reinforcement learning. *Nature medicine 24*, (11):1641–1642, 2018.

Sutskever, I., Vinyals, O., and Le, Q. V. . Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 2014.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *In Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Thananjeyan, B., Garg, A., Krishnan, S., Chen, C., Miller, L., and Goldberg, K. Multilateral surgical pattern cutting in 2d orthotropic gauze with deep reinforcement learning policies for tensioning. *In 2017 IEEE International Conference on Robotics and Automation (ICRA).*, pp. 2371–2378, 2017.

Wang, L., Zhang, D., Gao, L., Song, J., Guo, L., and Tao Shen, H. Mathdqn: Solving arithmetic word problems via deep reinforcement learning. *In Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Wang, Y., Liu, J., Liu, Y., Hao, J., He, Y., Hu, J., P. Yan, W., and Li, M. Ladder: A human-level bidding agent for large-scale real-time online auctions. *arXiv preprint arXiv:1708.05565*, 2017.

Wang, Z., Schaul, T., Hessel, M., Van Hasselt, H., Lanctot, M., and De Freitas, N. Dueling network architectures for deep reinforcement learning. *Internation Conference on Machine Learning ICML.*, pp. 19952003, 2016.

Yauney, G. and Pratik, S. Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection. *In Machine Learning for Healthcare Conference*, pp. 161–226, 2018.