Front. Comput. Sci., 2019, 13(5): 911–912
https://doi.org/10.1007/s11704-019-9901-7

PERSPECTIVE

A perspective on off-policy evaluation in reinforcement learning

Lihong LI (✉)

Google Brain, Kirkland, WA 98033, USA

The goal of reinforcement learning (RL) is to build an autonomous agent that takes a sequence of actions to maximize a utility function by interacting with an external, unknown environment. It is a very general learning paradigm that can model a wide range of problems, such as games, robotics, autonomous driving, humancomputer interactions, recommendation, healthcare, and many others. In recent years, powered by advances in deep learning and computing power, RL has seen great successes, with AlphaGo/AlphaZero as a prominent example. Such impressive outcomes have sparked fast growing interests in using RL to solve real-life problems.

In this article, I will argue that we must address the *evaluation* problem before RL can be widely adopted in real-life applications. In RL, the quality of a policy is often measured by the average reward received if the policy is followed by the agent to select actions. If the environment is *simulatable*, as in computer games, evaluation can be done simply by running the policy. However, for most real-life problems like autonomous driving and medical treatment, running a new policy in the *actual* environment can be expensive, risky and/or unethical. Creating a simulated environment for policy evaluation is common practice, but building a high-fidelity simulator can often be harder than finding an optimal policy itself (consider building a simulated patient that covers all possible medical conditions). Therefore, RL practitioners often find themselves stuck in a painful dilemma: in order to deploy a new policy, they have to show it is of sufficient quality, but the only reliable way to do so appears to be deploying the policy!

The problem The challenge above inspires the need for *off-policy evaluation* — evaluating a policy (the “target policy”) only using historical data collected by a *different* policy (the “behavior policy”), *without* actually running the target policy. The problem may sound simple, but in fact has remained to be one of the key and fundamental topics in RL research in the last decades.

It is helpful to compare with supervised learning (SL) to understand the challenges. Suppose we are building a spam detector. Evaluation in this case is straightforward: given a spam classifier, we may measure its accuracy (or other metrics of our choice) on a labeled dataset, and a classifier is considered better if its accuracy is higher. The case of RL is trickier. RL data is often in the form of a trajectory — a sequence of state-action-reward tuples where states depend on actions chosen earlier in the sequence. Therefore, if the policy “deviates” from the trajectory at some point (that is, choosing a different action than the one in the data), all future states and rewards will change but they are not observed in the data. In other words, unlike SL, data in RL only provide *partial* information for evaluation. Off-policy evaluation therefore requires to reason about *counterfactual* outcomes to answer what-if questions [1], and is closely related to causal inference.

The contextual bandits case Off-policy evaluation is easier in an important subclass of RL problems known as contextual bandits, where the agent’s actions do *not* affect future states. However, only the reward of the chosen action is observed in the data, so the need for counterfactual reasoning still exists. Contextual bandits are useful for modeling many important applications such as recommendation, advertising and Web search, where the reward may correspond to user clicks, video viewing time, or revenue [1–3].

Received April 4, 2019

E-mail: lihong@google.com

A powerful class of methods based on inverse propensity scoring (IPS) have proved effective in practice [1,2,4]. They use importance sampling to correct distribution mismatch between the observed data (collected by the behavior policy) and the desired but unobserved data (required by the target policy). The target policy's quality is often estimated by an importance-weighted average of rewards in the data. Under mild assumptions, IPS estimates are unbiased and converge to the target policy's true value as data size increases. A major difficulty in applying IPS methods is their high variance. Many approaches were proposed to reduce variance, possibly at the cost of slightly increased bias, in order to obtain a more accurate estimate [e.g., 2,5,6].

The general RL case IPS methods may be extended to the general case where the agent's actions affect future states. Conceptually, the only change is to apply importance sampling to the whole trajectory [e.g., 7–9]. Unfortunately, the variance of such an estimator can explode exponentially in the trajectory length, a phenomenon called *the curse of horizon* [10]. As a result, these methods have not been widely used in practice.

Recently, a new class of approaches were proposed to compute importance weights on states, not on trajectories, thus avoiding an explicit dependence on the trajectory length. Promising results were obtained for the first such algorithm [10], and stronger algorithms are being developed.

Conclusions Off-policy evaluation for contextual bandits has been successfully used in Web applications, and played a key role in enabling the deployment of bandit algorithms in these problems. The same can happen to general RL scenarios, where reliable off-policy evaluation is expected to unleash the power of RL. It gives a cheap and safe way to benchmark RL algorithms.

Many research opportunities exist, and we name a few to conclude the article. First, theoretical understanding of the problem's statistical nature is relatively limited, especially for general RL [6,8]. Second, most algorithmic developments in this area can be understood as balancing the well-known bias-variance trade-off. Other than the general techniques discussed here, one may identify useful structures in concrete applications to reduce variance, by decreasing the effective number of actions. Third, our discussion has focused on off-policy evaluation. A natural, and more challenging, next step is off-policy *optimization*, which requires to optimize a policy using historical data collected by the behavior policy.

References

1. Bottou L, Peters J, Quiñonero-Candela J, Charles D X, Chickering D M, Portugaly E, Ray D, Simard P, Snelson E. Counterfactual reasoning and learning systems: the example of computational advertising. Journal of Machine Learning Research, 2013, 14(1): 3207–3260
2. Hofmann K, Li L, Radlinski F. Online evaluation for information retrieval. Foundations and Trends in Information Retrieval, 2016, 10(1): 1–117
3. Li L, Chu W, Langford J, Schapire R E. A contextual-bandit approach to personalized news article recommendation. In: Proceedings of the 19th International Conference on World Wide Web. 2010, 661–670
4. Dudík M, Langford J, Li L. Doubly robust policy evaluation and learning. In: Proceedings of the 28th International Conference on Machine Learning. 2011, 1097–1104
5. Swaminathan A, Joachims T. The selfnormalized estimator for counterfactual learning. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. 2015, 3231–3239
6. Wang Y X, Agarwal A, Dudík M. Optimal and adaptive off-policy evaluation in contextual bandits. In: Proceedings of the 34th International Conference on Machine Learning. 2017, 3589–3597
7. Jiang N, Li L. Doubly robust off-policy evaluation for reinforcement learning. In: Proceedings of the 33rd International Conference on Machine Learning. 2016, 652–661
8. Li L, Munos R, Szepesvári C. Toward minimax off-policy value estimation. In: Proceedings of the 18th International Conference on Artificial Intelligence and Statistics. 2015, 608–616
9. Precup D, Sutton R S, Singh S P. Eligibility traces for off-policy policy evaluation. In: Proceedings of the 17th International Conference on Machine Learning. 2000, 759–766
10. Liu Q, Li L, Tang Z, Zhou D. Breaking the curse of horizon: infinite-horizon off-policy estimation. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2018, 5361–5371

Lihong Li is a research scientist at Google Brain, USA. Previously, he held research positions at Yahoo! Research (Silicon Valley) and Microsoft Research (Redmond). His main research interests are in reinforcement learning, including contextual bandits, and other related problems in AI. His work has found applications in recommendation, advertising, Web search and conversation systems, and has won best paper awards at ICML, AISTATS and WSDM. He serves as area chair or senior program committee member at major AI/ML conferences such as AAAI, ICLR, ICML, IJCAI and NIPS/NeurIPS.