Counterfactual Policy Evaluation and the Conditional Monte Carlo Method

Michel Ma
Department of Computer Science
University of Montreal
michel.ma@umontreal.ca

Pierre-Luc Bacon
Department of Computer Science
University of Montreal
pierre-luc.bacon@mila.quebec

Abstract

We show that the family of hindsight credit assignment algorithms of Harutyunyan et al. (2019) can be derived using a combination of importance sampling and the conditional Monte Carlo method (Hammersley, 1956; Bratley et al., 1987). This new perspective suggests a new interpretation for HCA as a class of off-policy evaluation methods where the off-policy target corresponds to a perturbation to the nominal one. We show that this approach seeks to extrapolate, using importance sampling, what would have been the return associated with actions that have not be taken along a trajectory of the behavior policy. In essence: extracting more information from a fragment of experience using counterfactuals. Coincidentally, this intuition also underlies the tree backup algorithm of Precup et al. (2000) which we derive using the same tools. Our results shed new light onto existing algorithms, and propose new avenues for counterfactual policy evaluation.

1 Introduction

The idea of predicting the outcome of other actions beyond the one actually taken – or *deviations from actuality* (Lewis, 1973) – is a central theme of reinforcement learning (Precup et al., 2000; Littman et al., 2001; Sutton et al., 2011). This is often the case for example in real-world applications for which limited amount of fixed offline data is available (Murphy et al., 2001; Mandel et al., 2014; Saria, 2018). This problem pertains to counterfactual inference (Neyman et al., 1935; Pearl, 1999) and has typically been tackled in reinforcement learning on three fronts: 1) through better generalization (Sutton, 1995) 2) by learning a model (Sutton, 1991) 3) using off-policy methods (Precup et al., 2000) based on importance sampling (Rubinstein, 1981). Our paper focuses on algorithms from the latter.

Counterfactuals can also arise when considering alternative courses of events based on new information retrospectively or *in hindsight*. This is the problem tackled by Harutyunyan et al. (2019) in the context of reinforcement learning who propose a family of *Hindsight Credit Assignment* (HCA) algorithms. While HCA has been derived by its authors using ideas from importance sampling (Rubinstein, 1981), we show how it can be distilled into two main ingredients: 1) a *perturbed* policy as the target 2) the use of the Conditional Monte Carlo Method (CMC) (Hammersley, 1956; Rubinstein, 1981; Bratley et al., 1987) over the importance sampling weights (Hesterberg, 1988; Bucklew, 2005; Rowland et al., 2020; Liu et al., 2020). It is then through the choice of conditioning statistics that the various HCA forms can be obtained. Hence, HCA can be seen as an instance of off-policy learning for a specific choice of target policy.

When no such restriction is made on the target policy, we show that our framework can also be used to elucidate the origins of the tree backup algorithm (Precup et al., 2000). Like HCA, tree backup also extrapolates from all the actions that could have been taken along a trajectory of the behavior policy. However, instead of assuming that the same policy would be followed in those alternatives, it can use a different policy as a target. Our results explain for the first time the statistical concepts

Offline Reinforcement Learning Workshop at Neural Information Processing Systems, 2020.

at play in tree backup without having recourse to eligibility traces, which they have always been associated to (Precup et al., 2000; Munos et al., 2016; Mahmood et al., 2017). More specifically, we show that the absence of the behavior policy at the denominator – a lack of importance ratios (Mahmood et al., 2017) – stems directly from the conditional Monte Carlo method.

2 Background and Notation

We denote a finite Markov Decision Process (MDP) by $(\mathcal{S}, \mathcal{A}, P, p_0, r, \gamma, T)$, where $\mathcal{S}$ is the set of possible states, $\mathcal{A}$ the set of possible actions, $P : \mathcal{S} \times \mathcal{A} \to \text{Dist}(\mathcal{S})$ defines the transition probabilities from state-action pairs to other states, p_0 is the initial state distribution, $r \in \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\gamma \in [0, 1)$ is the discount factor, and $T \in \mathbb{N}$ is the horizon. A randomized stationary policy, which we denote as $\pi(a_t|s_t)$, $\pi : \mathcal{S} \to \text{Dist}(\mathcal{A})$, is a conditional probability density function over actions conditioned on a state. We write the total expected discounted return from a state s under a policy π as:

$$V^\pi(s) := \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t R_t \middle| S_0 = s \right] := \mathbb{E}_\pi \left[G_T | S_0 = s \right],$$

and refer to V^π as the *value function* of π. Rather than conditioning on a state only, we can also condition a policy on a state and an action when evaluating the expected return. We summarize this information in a *action value function* which we write as:

$$Q^\pi(s, a) := \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t R_t \middle| S_0 = s, A_0 = a \right] := \mathbb{E}_\pi \left[G_T \middle| S_0 = s, A_0 = a \right].$$

We use capital letters inside any expectation to denote a random variable. In this case, R_t can also be replaced with $r(S_t, A_t)$ following our definition of an MDP. We also use interchangeably G_T to denote the return $\sum_{t=0}^{T} \gamma^t R_t$.

2.1 Conditional Importance Sampling

Importance sampling is a popular tool in reinforcement learning for off-policy evaluation. That is, estimating either the value function or action-value function of a target policy π under samples from some other behavior policy μ. The term importance sampling, borrowed from the simulation literature, in the context of RL simply refers to a change of variable:

$$\mathbb{E}_\pi \left[G_T \right] = \mathbb{E}_\mu \left[G_T \frac{\mathbb{P}_\pi(\tau)}{\mathbb{P}_\mu(\tau)} \right] = \mathbb{E}_\mu \left[G_T \prod_{k=0}^{T} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} \right] := \mathbb{E}_\mu \left[G_T \mathcal{P}_{0:T}^{\pi,\mu} \right] \quad ,$$

where τ is a shorthand for the trajectory $S_0, A_0, S_1, A_1, \ldots, S_T, A_T$ and $\mathcal{P}_{i:j}^{\pi,\mu}$ is the importance sampling ratio $\prod_{k=i}^{j} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}$. It is well-known that importance sampling is prone to high variance in high dimensional settings (Rubinstein, 1981; Bratley et al., 1987; L'Ecuyer, 1994), which is especially pronounced in reinforcement learning when learning over long horizons (Liu et al., 2018, 2020).

In an attempt to reduce the variance of such estimators, a new line of work Hallak and Mannor (2017); Liu et al. (2018); Gelada and Bellemare (2019); Xie et al. (2019) of *marginalized* (Xie et al., 2019) estimators have been proposed. In Liu et al. (2020), the authors showed that such estimators stem from an application of the conditional Monte Carlo (CMC) method (Bratley et al., 1987), which arises naturally from the law of total variance. Liu et al. (2020) shows that given any statistic ϕ_T dependent on the entire trajectory, we have:

$$V^\pi(s) = \mathbb{E}_\mu \left[G_T \mathcal{P}_{0:T}^{\pi,\mu} \middle| S_0 = s \right] = \mathbb{E}_\mu \left[\mathbb{E}_\mu \left[G_T \mathcal{P}_{0:T}^{\pi,\mu} \middle| S_0 = s, \phi_T \right] \middle| S_0 = s \right] \quad , \tag{1}$$

and by the law of total variance:

$$\text{Var}\left(\mathbb{E}_\mu \left[G_T \mathcal{P}_{0:T}^{\pi,\mu} \middle| S_0 = s, \phi_T \right] \right) = \text{Var}\left(G_T \mathcal{P}_{0:T}^{\pi,\mu} \right) - \mathbb{E}_\mu \left[\text{Var}\left(G_T \mathcal{P}_{0:T}^{\pi,\mu} \right) \middle| S_0 = s, \phi_T \right] \quad .$$

Because the second term is always non-negative, we have that

$$\text{Var}\left(\mathbb{E}_\mu \left[G_T \mathcal{P}_{0:T}^{\pi,\mu} \middle| S_0 = s, \phi_T \right] \right) \leq \text{Var}\left(G_T \mathcal{P}_{0:T}^{\pi,\mu} \right) \quad ,$$

and the variance is either reduced or kept the same. The extended conditional Monte Carlo (ECMC) (Bratley et al., 1987) method allows us to condition on a stage-dependent variable ϕ_t instead:

$$V^\pi(s) = \mathbb{E}_\mu \left[\sum_{t=0}^{T} \gamma^t R_t \mathcal{P}_{0:T}^{\pi,\mu} \middle| S_0 = s \right] = \mathbb{E}_\mu \left[\sum_{t=0}^{T} \gamma^t \mathbb{E}_\mu \left[R_t \mathcal{P}_{0:T}^{\pi,\mu} \middle| S_0 = s, \phi_t \right] \middle| S_0 = s \right] . \quad (2)$$

Unlike CMC, ECMC does not guarantee the variance remain unchanged (Bratley et al., 1987) or is reduced, because the variance of a summation is not distributive, and involves covariance terms between each summed random variable. Liu et al. (2020) give sufficient conditions for variance reduction using ECMC.

3 Counterfactual Policy Evaluation

Consider a perturbation of a policy π, denoted π^a to be identical to π, except that π^a performs action a as its first action every time instead. From this definition, we can rewrite $Q^\pi(s,a)$ as $V^{\pi^a}(s)$ to be equivalent expressions. We can then observe that the on-policy evaluation problem for Q^π is equivalent to the *off-policy value evaluation* problem:

$$Q^\pi(s,a) = V^{\pi^a}(s) = \mathbb{E}_\pi \left[\mathcal{P}_{0:T}^{\pi^a,\pi} \sum_{t=0}^{T} \gamma^t R_t \middle| S_0 = s \right] . \quad (3)$$

From this expression, we can obtain the classic Monte Carlo estimator for $Q^\pi(s,a)$:

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[\mathcal{P}_{0:T}^{\pi^a,\pi} \sum_{t=0}^{T} \gamma^t R_t \middle| S_0 = s \right] = \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t R_t \middle| S_0 = s, A_0 = a \right] .$$

This expression is only able to update $Q^\pi(s,a)$ based on samples that begin with s, a. As mentioned earlier, we would like to leverage the possibility of using counterfactuals. Let us define a new set of estimators, called *counterfactual on-policy evaluation* which combines equations (1) (CMC) or (2) (ECMC) with equation (3):

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[\mathbb{E}_\pi \left[\mathcal{P}_{0:T}^{\pi^a,\pi} G_T | S_0 = s, \phi_T \right] \middle| S_0 = s \right] , \quad (4)$$

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t \mathbb{E}_\pi \left[R_t \mathcal{P}_{0:T}^{\pi^a,\pi} | S_0 = s, \phi_t \right] \middle| S_0 = s \right] . \quad (5)$$

A trivial application of the PDIS estimator on equation (5), setting ϕ_t to be the history of the trajectory $\tau_{0:t}$ and R_t, yields the same equation as the classic Monte Carlo estimator above, and thus no improvement. Let us consider instead the SIS estimator on equation (5), setting ϕ_t to $\{S_t, A_t, R_t\}$:

$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t \mathbb{E}_\pi \left[R_t \mathcal{P}_{0:T}^{\pi^a,\pi} \middle| S_0 = s, S_t, A_t, R_t \right] \middle| S_0 = s \right] \\ &= \mathbb{E} \left[R_0 \middle| S_0 = s, A_0 = a \right] + \mathbb{E}_\pi \left[\sum_{t>0}^{T} \gamma^t \frac{\mathbb{P}_\pi(A_0 = a | S_0, S_t)}{\pi(a|S_0)} R_t \middle| S_0 = s \right] . \end{aligned} \quad (6)$$

We may also consider ϕ_T to be the return G_T in the CMC scenario, equation (4), to obtain:

$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}_\pi \left[\mathbb{E}_\pi \left[\mathcal{P}_{0:T}^{\pi^a,\pi} G_T \middle| S_0 = s, G_T \right] \middle| S_0 = s \right] \\ &= \mathbb{E}_\pi \left[G_T \frac{\mathbb{P}_\pi(A_0 = a | S_0, G_T)}{\pi(a|S_0)} \middle| S_0 = s \right] . \end{aligned} \quad (7)$$

Equations (6) and (7) (proven in the appendix) require knowledge of some other functions, namely $r(s,a)$ for any state s and action a, $\mathbb{P}_\pi(A_0 = a | S_0, S_t)$ for any action a, states S_0 and S_t, and $\mathbb{P}_\pi(A_0 = a | S_0, G_T)$ for any action a, state S_0 and return G_T. Provided that these functions can be learned or are given, the above estimators allow for the on-policy evaluation of Q^π through the use

of counterfactuals. That is, $Q^\pi(s,a)$ can be estimated for all actions a given a single sample drawn from $S_0 = s$ and π.

In fact, these equations correspond to the expressions presented in (Harutyunyan et al., 2019). Seeing it as a case of conditional importance sampling can offer insight into the original paper by Harutyunyan et al. (2019), offering an explanation to their results that goes beyond their intuitions and heuristics. To paraphrase Harutyunyan et al. (2019): *"We hope to improve the efficiency of credit assignment by changing measures from policy π to a future conditional distribution"*. In reality, we now know that any improvements are a result of ECMC or CMC, and unlikely due to explicit better credit assignment.

4 Connection to Tree Backup

A popular method for off-policy evaluation that has long been seen as a separate approach to importance sampling is the tree backup algorithm originally proposed in (Precup et al., 2000). We consider here the full Monte Carlo version of n-step tree backup, eg where $n = T$. However, our results are can be extended to any other values of n.

$$Q^\pi(s,a) = \mathbb{E}_\mu \left[\sum_{t=0}^{T} \gamma^t \prod_{k=1}^{t} \pi(A_k|S_k)(R_t + \gamma \sum_{\alpha \neq A_{t+1}} \pi(\alpha|S_{t+1}) Q^\pi(S_{t+1}, \alpha)) \middle| S_0 = s, A_0 = a \right] . \tag{8}$$

By convention when $k > j$, $\prod_{k=i}^{j} X_k = 1$. At first glance it can be dismissed as something unrelated to importance sampling since the behavior policy is nowhere to be seen in the algorithm. Additionally, empirical evidence suggests that the tree backup estimator outperforms IS and PDIS. Intuitively, the tree backup algorithm omits any explicit product of μ by considering possible *"what-if?"* scenarios, or counterfactuals, thus circumventing the curse of horizon, but theoretical justifications are lacking.

Using the same framework developed above for *counterfactual policy evaluation* with no restrictions on the target policy, the tree backup algorithm can be re-derived as the basic IS estimator improved by ECMC. The proof is found in the appendix.

Theorem 1. *(tree backup is an instance of conditional importance sampling)*

$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}_\mu \left[\sum_{t=0}^{T} \gamma^t \mathbb{E}_\mu \left[R^t \mathcal{P}_{1:t}^{\pi,\mu} \middle| S_0 = s, A_0 = a, A_1, A_2, \ldots, A_t \right] \middle| S_0 = s, A_0 = a \right] \\ &= \mathbb{E}_\mu \left[\sum_{t=0}^{T} \gamma^t \prod_{k=1}^{t} \pi(A_k|S_k)(R_t + \gamma \sum_{\alpha \neq A_{t+1}} \pi(\alpha|S_{t+1}) Q^\pi(S_{t+1}, \alpha)) \middle| S_0 = s, A_0 = a \right] . \end{aligned}$$

From Theorem 1, we can achieve a better understanding of the real benefits of the tree backup algorithm, just as Liu et al. (2020) did with PDIS and SIS. Seen as part of the *counterfactual policy evaluation* estimators, we can better understand how tree backup achieves better empirical results through counterfactual reasoning by leveraging the extended conditional Monte Carlo method.

5 Conclusion

In this work, we present two novel applications of conditional importance sampling, which was formalized by Liu et al. (2020). First, we propose a new use of CIS in the on-policy setting, to leverage the use of counterfactuals. Following this novel idea of "counterfactual on-policy evaluation", we recover the expressions derived in (Harutyunyan et al., 2019) through the lens of the (extended) conditional Monte Carlo method. Second, in the off-policy evaluation setting, we cast the tree backup algorithm as an instance of CIS, demonstrating that while not obvious, it does perform some form of importance sampling. In both cases, conditional importance sampling is used to inject counterfactual reasoning in importance sampling based estimators. This framework goes beyond heuristics to provide a theoretically just way to reduce the variance of IS estimators. These findings lay the groundwork for a better understanding of when and how such algorithms could be used for better policy evaluation.

References

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, and Rémi Munos. Hindsight credit assignment. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 12467–12476, 2019.

J. M. Hammersley. Conditional monte carlo. *J. ACM*, 3(2):73–76, April 1956. ISSN 0004-5411.

Paul Bratley, Bennett L. Fox, and Linus E. Schrage. *A Guide to Simulation (2Nd Ed.)*. Springer-Verlag, Berlin, Heidelberg, 1987. ISBN 0-387-96467-3.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In Pat Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pages 759–766. Morgan Kaufmann, 2000.

David Lewis. *Counterfactuals*. Harvard University Press, 1973.

Michael L. Littman, Richard S. Sutton, and Satinder P. Singh. Predictive representations of state. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 1555–1561. MIT Press, 2001.

Richard S. Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M. Pilarski, Adam White, and Doina Precup. Horde: a scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In Liz Sonenberg, Peter Stone, Kagan Tumer, and Pinar Yolum, editors, *10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011), Taipei, Taiwan, May 2-6, 2011, Volume 1-3*, pages 761–768. IFAAMAS, 2011.

S A Murphy, M J van der Laan, and J M Robins and. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, December 2001.

Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In Ana L. C. Bazzan, Michael N. Huhns, Alessio Lomuscio, and Paul Scerri, editors, *International conference on Autonomous Agents and Multi-Agent Systems, AAMAS '14, Paris, France, May 5-9, 2014*, pages 1077–1084. IFAAMAS/ACM, 2014.

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

J. Neyman, K. Iwaszkiewicz, and St. Kolodziejczyk. Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, 2(2):107, 1935.

Judea Pearl. *Synthese*, 121(1/2):93–149, 1999.

Richard S. Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In David S. Touretzky, Michael Mozer, and Michael E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8, NIPS, Denver, CO, USA, November 27-30, 1995*, pages 1038–1044. MIT Press, 1995.

Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *SIGART Bull.*, 2(4):160–163, 1991. URL `https://doi.org/10.1145/122344.122377`.

Reuven Y. Rubinstein. *Simulation and the Monte Carlo Method*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1981. ISBN 0471089176.

Timothy C. Hesterberg. *Advances in Importance Sampling*. PhD thesis, Stanford University, August 1988.

James A Bucklew. Conditional importance sampling estimators. *IEEE transactions on information theory*, 51(1):143–153, 2005.

Mark Rowland, Anna Harutyunyan, Hado van Hasselt, Diana Borsa, Tom Schaul, Rémi Munos, and Will Dabney. Conditional importance sampling for off-policy learning. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, vol-

ume 108 of *Proceedings of Machine Learning Research*, pages 45–55. PMLR, 2020. URL http://proceedings.mlr.press/v108/rowland20b.html.

Yao Liu, Pierre-Luc Bacon, and Emma Brunskill. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling, 2020.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1046–1054, 2016.

Ashique Rupam Mahmood, Huizhen Yu, and Richard S. Sutton. Multi-step off-policy learning without importance sampling ratios. *CoRR*, abs/1702.03006, 2017. URL http://arxiv.org/abs/1702.03006.

Pierre L'Ecuyer. Efficiency improvement and variance reduction. In *Proceedings of the 26th Conference on Winter Simulation*, WSC '94, pages 122–132, San Diego, CA, USA, 1994. Society for Computer Simulation International. ISBN 0-7803-2109-X.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 5361–5371, 2018.

Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1372–1383, 2017.

Carles Gelada and Marc G. Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.*, pages 3647–3655, 2019.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, 2019.

A Proofs

A.1 Proof of theorem 1

Lemma 1. *Given any state s, action a, target policy π, and behavior policy μ for which $\mu(a_t|s_t) > 0$ for all a_t and s_t:*

$$Q^\pi(s,a) = \mathbb{E}_\mu\left[R_0 + \sum_{t=1}^{T} \gamma^T R_t \mathcal{P}_{1:t}^{\pi,\mu} \middle| S_0 = s, A_0 = a\right] = \mathbb{E}_\mu\left[\sum_{t=0}^{T} \gamma^t R_t \mathcal{P}_{1:t}^{\pi,\mu} \middle| S_0 = s, A_0 = a\right]$$

Proof. By per decision importance sampling:

$$V^\pi(s) = \mathbb{E}_\mu\left[\sum_{t=0}^{T} \gamma^t R_t \prod_{k=0}^{t} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} \middle| S_0 = s\right]$$

By the law of total expectation

$$= \mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\sum_{t=0}^{T} \gamma^t R_t \prod_{k=0}^{t} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} \middle| S_0 = s, A_0\right] \middle| S_0 = s\right]$$

$$= \sum_{a\in\mathcal{A}} \mu(a|s)\mathbb{E}_\mu\left[\sum_{t=0}^{T} \gamma^t R_t \prod_{k=0}^{t} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} \middle| S_0 = s, A_0 = a\right]$$

$$= \sum_{a\in\mathcal{A}} \mu(a|s)\mathbb{E}_\mu\left[\frac{\pi(A_0|S_0)}{\mu(A_0|S_0)}\sum_{t=0}^{T} \gamma^t R_t \prod_{k=1}^{t} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} \middle| S_0 = s, A_0 = a\right]$$

$$= \sum_{a\in\mathcal{A}} \mu(a|s)\frac{\pi(a|s)}{\mu(a|s)}\mathbb{E}_\mu\left[\sum_{t=0}^{T} \gamma^t R_t \prod_{k=1}^{t} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} \middle| S_0 = s, A_0 = a\right]$$

$$= \sum_{a\in\mathcal{A}} \pi(a|s)\mathbb{E}_\mu\left[\sum_{t=0}^{T} \gamma^t R_t \prod_{k=1}^{t} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} \middle| S_0 = s, A_0 = a\right]$$

By definition, $V^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s)Q^\pi(s,a)$, therefore

$$Q^\pi(s,a) = \mathbb{E}_\mu\left[\sum_{t=0}^{T} \gamma^t R_t \mathcal{P}_{1:t}^{\pi,\mu} \middle| S_0 = s, A_0 = a\right]$$

□

Lemma 2. *Consider the MDP setting, where random variables $S_i, A_i, \ldots, S_t, A_t$ are actions and states visited by a policy μ. Let $\mu(a|s)$ be the conditional probability of taking action a at state s, then:*

$$\mathbb{E}_\mu\left[\sum_{t=i}^{T} f(S_i, A_i, \ldots, S_t, A_t)\right]$$

$$= \mathbb{E}_\mu\left[\sum_{\alpha\neq A_i} \mu(\alpha|S_i)\mathbb{E}_\mu\left[\sum_{t=i}^{T} f(S_i, A_i, \ldots, S_t, A_t) \middle| A_i = \alpha\right] + \mu(A_i|S_i)\sum_{t=i}^{T} f(S_i, A_i, \ldots, S_t, A_t)\right]$$

Proof.

$$\mathbb{E}_\mu \left[\sum_{t=i}^{T} f(S_i, A_i, \ldots, S_t, A_t) \right]$$

By the law of total expecation

$$= \mathbb{E}_\mu \left[\mathbb{E}_\mu \left[\sum_{t=i}^{T} f(S_i, A_i, \ldots, S_t, A_t) \Big| A_i \right] \right]$$

$$= \sum_{\alpha \in \mathcal{A}} \mathbb{P}_\mu(A'_i = \alpha) \mathbb{E}_\mu \left[\sum_{t=i}^{T} f(S_i, A_i, \ldots, S_t, A_t) | A_i = \alpha \right]$$

$$= \mathbb{E}_\mu \left[\sum_{\alpha \in \mathcal{A}} \mathbb{P}_\mu(A'_i = \alpha) \mathbb{E}_\mu \left[\sum_{t=i}^{T} f(S_i, A_i, \ldots, S_t, A_t) | A_i = \alpha \right] \right]$$

$$= \mathbb{E}_\mu \left[\sum_{\alpha \neq A_i} \mathbb{P}_\mu(A'_i = \alpha) \mathbb{E}_\mu \left[\sum_{t=i}^{T} f(S_i, A_i, \ldots, S_t, A_t) | A_i = \alpha \right] + \mathbb{P}_\mu(A'_i = A_i) \mathbb{E}_\mu \left[\sum_{t=i}^{T} f(S_i, A_i, \ldots, S_t, A_t) | A_i = A_i \right] \right]$$

$$= \mathbb{E}_\mu \left[\sum_{\alpha \neq A_i} \mu(\alpha | S_i) \mathbb{E}_\mu \left[\sum_{t=i}^{T} f(S_i, A_i, \ldots, S_t, A_t) | A_i = \alpha \right] + \mu(A_i | S_i) \sum_{t=i}^{T} f(S_i, A_i, \ldots, S_t, A_t) \right]$$

□

Back to Theorem 1

Consider first conditioning on A_1

$$Q^\pi(s, a) = \mathbb{E}_\mu \left[R_0 + \sum_{t=1}^{T} \gamma^t R_t \mathcal{P}_{1:t}^{\pi,\mu} \Big| S_0 = s, A_0 = a \right]$$

By Lemma 2 on the second term, where $i = 1$

$$= \mathbb{E}_\mu \Bigg[R_0 + \sum_{\alpha \neq A_1} \mu(\alpha | S_1) \mathbb{E}_\mu \left[\sum_{t=1}^{T} \gamma^t R_t \mathcal{P}_{1:t}^{\pi,\mu} \Big| S_0 = s, A_0 = a, A_1 = \alpha \right]$$

$$+ \mu(A_1 | S_1) \sum_{t=1}^{T} \gamma^t R_t \mathcal{P}_{1:t}^{\pi,\mu} \Big| S_0 = s, A_0 = a \Bigg]$$

$$= \mathbb{E}_\mu \Bigg[R_0 + \sum_{\alpha \neq A_1} \mu(\alpha | S_1) \frac{\pi(\alpha | S_1)}{\mu(\alpha | S_1)} \mathbb{E}_\mu \left[\sum_{t=1}^{T} \gamma^t R_t \mathcal{P}_{2:t}^{\pi,\mu} \Big| S_0 = s, A_0 = a, A_1 = \alpha \right]$$

$$+ \mu(A_1 | S_1) \frac{\pi(A_1 | S_1)}{\mu(A_1 | S_1)} \sum_{t=1}^{T} \gamma^t R_t \mathcal{P}_{2:t}^{\pi,\mu} \Big| S_0 = s, A_0 = a \Bigg]$$

$$= \mathbb{E}_\mu \Bigg[R_0 + \sum_{\alpha \neq A_1} \pi(\alpha | S_1) \mathbb{E}_\mu \left[\sum_{t=1}^{T} \gamma^t R_t \mathcal{P}_{2:t}^{\pi,\mu} \Big| S_0 = s, A_0 = a, A_1 = \alpha \right]$$

$$+ \pi(A_1 | S_1) \sum_{t=1}^{T} \gamma^t R_t \mathcal{P}_{2:t}^{\pi,\mu} \Big| S_0 = s, A_0 = a \Bigg]$$

And by Lemma 1

$$= \mathbb{E}_\mu \Bigg[R_0 + \sum_{\alpha \neq A_1} \pi(\alpha | S_1) \gamma Q^\pi(S_1, \alpha) + \gamma \pi(A_1 | S_1) \sum_{t=1}^{T} \gamma^{t-1} R_t \mathcal{P}_{2:t}^{\pi,\mu} \Big| S_0 = s, A_0 = a \Bigg]$$

Now, since $\pi(A_1|S_1)$ is independent of any future actions or states, we may do the same thing onto the last term, conditioning on A_2 now to obtain:

$$
\begin{aligned}
Q^\pi(s,a) &= \mathbb{E}_\mu\left[R_0 + \sum_{t=1}^{T}\gamma^t R_t \mathcal{P}_{1:t}^{\pi,\mu}\middle| S_0 = s, A_0 = a\right] \\
&= \Bigg[R_0 + \sum_{\alpha\neq A_1}\pi(\alpha|S_1)\gamma Q^\pi(S_1,\alpha) \\
&+ \gamma\pi(A_1|S_1)\Bigg(R_1 + \sum_{\alpha\neq A_2}\gamma\pi(\alpha|S_2)Q^\pi(S_2,\alpha) + \gamma\pi(A_2|S_2)\sum_{t=2}^{T}\gamma^{t-2}R_t\mathcal{P}_{3:t}^{\pi,\mu}\Bigg)\Bigg|S_0 = s, A_0 = a\Bigg] \\
&= \mathbb{E}_\mu\Bigg[\sum_{t=0}^{1}\gamma^t\prod_{k=0}^{t}\pi(A_k|S_k)(R_t + \gamma\sum_{\alpha\neq A_{t+1}}\pi(\alpha|S_{t+1})Q^\pi(S_{t+1},\alpha)) \\
&+ \gamma^2\prod_{k=1}^{2}\pi(A_k|S_k)\sum_{t=2}^{T}\gamma^{t-2}R_t\mathcal{P}_{3:t}^{\pi,\mu}\Bigg|S_0 = s, A_0 = a\Bigg]
\end{aligned}
$$

Continuously applying the same logic up to A_T results in:

$$
Q^\pi(s,a) = \mathbb{E}_\mu\left[\sum_{t=0}^{T}\gamma^t\prod_{k=0}^{t}\pi(A_k|S_k)(R_t + \gamma\sum_{\alpha\neq A_{t+1}}\pi(\alpha|S_{t+1})Q^\pi(S_{t+1},\alpha))\middle|S_0 = s, A_0 = a\right]
$$

Notice that the conditional trick is not applied to terms in the time steps before A_i, therefore:

$$
\begin{aligned}
Q^\pi(s,a) &= \mathbb{E}_\mu\left[\sum_{t=0}^{T}\gamma^t\mathbb{E}_\mu\left[R^t\mathcal{P}_{1:t}^{\pi,\mu}\middle|S_0 = s, A_0 = a, A_1, A_2, \ldots, A_t\right]\middle|S_0 = s, A_0 = a\right] \\
&= \mathbb{E}_\mu\left[\sum_{t=0}^{T}\gamma^t\prod_{k=1}^{t}\pi(A_k|S_k)(R_t + \gamma\sum_{\alpha\neq A_{t+1}}\pi(\alpha|S_{t+1})Q^\pi(S_{t+1},\alpha))\middle|S_0 = s, A_0 = a\right] \quad \square
\end{aligned}
$$

A.2 Proof of equation 6

For any state s, action a, and policy π for which $\pi(a_t|s_t) > 0$ for all t, we have

$$
\begin{aligned}
Q^\pi(s,a) &= \mathbb{E}_\pi\left[\sum_{t=0}^{T}\gamma^t R_t\mathcal{P}_{0:T}^{\pi^a,\pi}\middle|S_0 = s\right] \\
&= \mathbb{E}_\pi\left[\sum_{t=0}^{T}\gamma^t\mathbb{E}_\pi\left[R_t\mathcal{P}_{0:T}^{\pi^a,\pi}\middle|S_0 = s, S_t, A_t, R_t\right]\middle|S_0 = s\right] \\
&\text{Since } \mathcal{P}_{0:T}^{\pi^a,\pi} \text{ is conditionally independent of } R_t \text{ given } S_t, A_t \\
&= \mathbb{E}_\pi\left[\sum_{t=0}^{T}\gamma^t R_t\mathbb{E}_\pi\left[\mathcal{P}_{0:T}^{\pi^a,\pi}\middle|S_0 = s, S_t, A_t\right]\middle|S_0 = s\right]
\end{aligned}
$$

Let's evaluate the inner expectation:

$$\mathbb{E}_\pi\left[\mathcal{P}_{0:T}^{\pi^a,\pi}\middle|S_0=s,S_t,A_t\right]$$
$$=\sum_{s_0,a_0,\ldots,s_T,a_T}\mathcal{P}_{0:T}^{\pi^a,\pi}\mathbb{P}_\pi(s_0,a_0,\ldots,s_T,a_T|S_0=s,S_t,A_t)$$
$$=\sum_{s_0\in\mathcal{S}}\sum_{a_0\in\mathcal{A}}\frac{\pi^a(a_0|s_0)}{\pi(a_0|s_0)}\sum_{s_1,a_1,\ldots,s_T,a_T}\mathcal{P}_{1:T}^{\pi^a,\pi}\mathbb{P}_\pi(s_0,a_0|S_0=s,S_t,A_t)\mathbb{P}_\pi(\tau_{1:T}|S_0=s,S_t,A_t,s_0,a_0)$$

Because $\pi^a(a_t|s_t)=\pi(a_t|s_t)$ for $t>0$

$$=\sum_{s_0\in\mathcal{S}}\sum_{a_0\in\mathcal{A}}\frac{\pi^a(a_0|s_0)}{\pi(a_0|s_0)}\mathbb{P}_\pi(s_0,a_0|S_0=s,S_t,A_t)\sum_{s_1,a_1,\ldots,s_T,a_T}\mathbb{P}_\pi(s_1,a_1,\ldots,s_T,a_T|S_0=s,S_t,A_t,s_0,a_0)$$
$$=\sum_{s_0\in\mathcal{S}}\sum_{a_0\in\mathcal{A}}\frac{\pi^a(a_0|s_0)}{\pi(a_0|s_0)}\mathbb{P}_\pi(s_0,a_0|S_0=s,S_t,A_t)$$
$$=\sum_{a_0\in\mathcal{A}}\frac{\pi^a(a_0|s)}{\pi(a_0|s)}\mathbb{P}_\pi(a_0|S_0=s,S_t,A_t)$$

Because $\pi^a(a_0|s)=1$ only when $a_0=a$, and 0 otherwise

$$=\frac{\mathbb{P}_\pi(A_0=a|S_0=s,S_t,A_t)}{\pi(a|s)}$$

Now consider when $t=0$, because A_0 is independent of S_0 given A_0:

$$\frac{\mathbb{P}_\pi(A_0=a|S_0=s,S_0,A_0)}{\pi(a|s)}=\frac{\mathbb{P}_\pi(A_0=a|A_0)}{\pi(a|s)}$$

And when $t>0$, because A_0 is independent of A_t given S_t for $t>0$:

$$\frac{\mathbb{P}_\pi(A_0=a|S_0=s,S_t,A_t)}{\pi(a|s)}=\frac{\mathbb{P}_\pi(A_0=a|S_0=s,S_t)}{\pi(a|s)}$$

So, back to the original expression:

$$\begin{aligned}Q^\pi(s,a)&=\mathbb{E}_\pi\left[\sum_{t=0}^{T}\gamma^tR_t\mathbb{E}_\pi\left[\mathcal{P}_{0:T}^{\pi^a,\pi}\middle|S_0=s,S_t,A_t\right]\middle|S_0=s\right]\\&=\mathbb{E}_\pi\left[R_0\frac{\mathbb{P}_\pi(A_0=a|A_0)}{\pi(a|s)}\middle|S_0=s\right]+\mathbb{E}_\pi\left[\sum_{t>0}^{T}\gamma^tR_t\frac{\mathbb{P}_\pi(A_0=a|S_0,S_t)}{\pi(a|S_0)}\middle|S_0=s\right]\\&=\mathbb{E}_\pi\left[R_0\middle|S_0=s,A_0=a\right]+\mathbb{E}_\pi\left[\sum_{t>0}^{T}\gamma^tR_t\frac{\mathbb{P}_\pi(A_0=a|S_0,S_t)}{\pi(a|S_0)}\middle|S_0=s\right]\quad\square\end{aligned}$$

A.3 Proof of equation 7

For any state s, action a, and policy π for which $\pi(a_t|s_t)>0$ for all t, we have:

$$\begin{aligned}Q^\pi(s,a)&=\mathbb{E}_\pi\left[G_T\mathcal{P}_{0:T}^{\pi^a,\pi}\middle|S_0=s\right]\\&=\mathbb{E}_\pi\left[\mathbb{E}_\pi\left[G_T\mathcal{P}_{0:T}^{\pi^a,\pi}\middle|S_0=s,G_T\right]\middle|S_0=s\right]\\&\text{Since }\mathcal{P}_{0:T}^{\pi^a,\pi}\text{ is conditionally independent of }G_T\text{ given }G_T\\&=\mathbb{E}_\pi\left[G_T\mathbb{E}_\pi\left[\mathcal{P}_{0:T}^{\pi^a,\pi}\middle|S_0=s,G_T\right]\middle|S_0=s\right]\end{aligned}$$

Let's evaluate the inner expectation:

$$\mathbb{E}_\pi \left[\mathcal{P}_{0:T}^{\pi^a,\pi} | S_0 = s, G_T \right]$$

$$= \sum_{s_0,a_0,\ldots,s_T,a_T} \mathcal{P}_{0:T}^{\pi^a,\pi} \mathbb{P}_\pi(s_0, a_0, \ldots, s_T, a_T | S_0 = s, G_T)$$

Because $\pi^a(a_t|s_t) = \pi(a_t|s_t)$ for $t > 0$

$$= \sum_{s_0 \in \mathcal{S}} \sum_{a_0 \in \mathcal{A}} \frac{\pi^a(a_0|s_0)}{\pi(a_0|s_0)} \sum_{s_1,a_1,\ldots,s_T,a_T} \mathbb{P}_\pi(s_0, a_0 | S_0 = s, G_T) \mathbb{P}_\pi(s_1, a_1, \ldots, s_T, a_T | S_0 = s, G_T, s_0, a_0)$$

$$= \sum_{s_0 \in \mathcal{S}} \sum_{a_0 \in \mathcal{A}} \frac{\pi^a(a_0|s_0)}{\pi(a_0|s_0)} \mathbb{P}_\pi(s_0, a_0 | S_0 = s, G_T) \sum_{s_1,a_1,\ldots,s_T,a_T} \mathbb{P}_\pi(s_1, a_1, \ldots, s_T, a_T | S_0 = s, G_T, s_0, a_0)$$

$$= \sum_{s_0 \in \mathcal{S}} \sum_{a_0 \in \mathcal{A}} \frac{\pi^a(a_0|s_0)}{\pi(a_0|s_0)} \mathbb{P}_\pi(s_0, a_0 | S_0 = s, G_T)$$

$$= \sum_{a_0 \in \mathcal{A}} \frac{\pi^a(a_0|s)}{\pi(a_0|s)} \mathbb{P}_\pi(a_0 | S_0 = s, G_T)$$

Because $\pi^a(a_0|s) = 1$ only when $a_0 = a$, and 0 otherwise

$$= \frac{\mathbb{P}_\pi(A_0 = a | S_0 = s, G_T)}{\pi(a|s)}$$

Back to the original expression:

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[G_T \mathbb{E}_\pi \left[\mathcal{P}_{0:T}^{\pi^a,\pi} \middle| S_0 = s, G_T \right] \middle| S_0 = s \right]$$

$$= \mathbb{E}_\pi \left[G_T \frac{\mathbb{P}_\pi(A_0 = a | S_0, G_T)}{\pi(a|S_0)} \middle| S_0 = s \right] \quad \square$$