Behaviour-Agnostic Reinforcement Learning: *We have data, now what?*

"How does the use of sample-splitting and cross-fitting techniques mitigate the effects of double-dipping in behaviour agnostic reinforcement learning (RL) ?"

Author: Yaren Aslan (y.aslan-1@student.tudelft.nl)
Supervisor: Stephan Bongers, **Responsible Professor:** Frans Oliehoek

01 Background and Objectives

Off-policy Evaluation [1]

- Off-policy evaluation (OPE) refers to the setting where the agent estimates the value of a target policy by referring only to a dataset of experience previously collected by other policies in this environment.
- The objective is to estimate the expected cumulative (discounted) reward of the target policy would achieve if deployed.

Behaviour-agnostic OPE and DICE Estimators [1]

- Behaviour-agnostic OPE denotes an approach where the learning algorithm does not make any assumptions about the behavior policy that generated the dataset.
- Estimators such as from the "DICE" family are employed to display the ratio between the propensity of the target policy to visit distinct state-action pairs compared to their occurrence likelihood in the off-policy data.

Overfitting and Double-dipping [3]

- Double-dipping describes overfitting a model through both building and evaluating the model on the same dataset,
  - leading to low in-sample error but high variance and poor generalizability.
- In DICE estimators, the same dataset is used for both training the neural network on Q-value functions and visitation densities, and estimating the target policy value.

OBJECTIVE 01
Adopt commonly used techniques in doubly robust machine learning methods, namely sample-splitting and cross-fitting for DICE-RL estimators

OBJECTIVE 02
Analyze how effective the employed techniques are in mitigating the risks associated with double dipping for behaviour-agnostic RL

02 Methodology

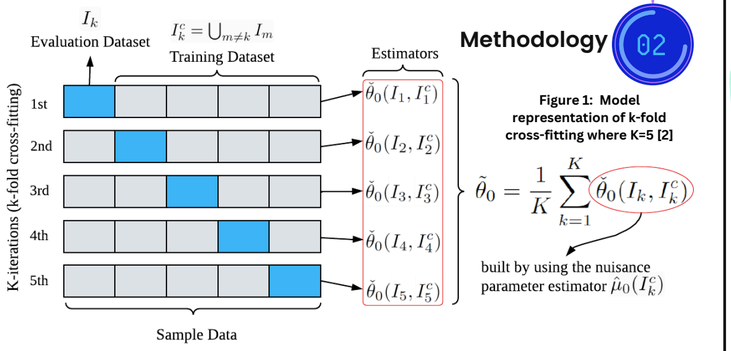


Figure 1: Model representation of k-fold cross-fitting where K=5 [2]

$$\tilde{\theta}_0 = \frac{1}{K}\sum_{k=1}^{K} \check{\theta}_0(I_k, I_k^c)$$

built by using the nuisance parameter estimator $\hat{\mu}_0(I_k^c)$

The true value μ_0 of the nuisance parameter μ is estimated by $\hat{\mu}_0(I^c)$ using the training sample $(W_i)_{i \in I^c}$

The true value θ_0 of the target parameter θ is estimated by estimator $\check{\theta}_0(I, I^c)$ using the evaluation sample $(W_i)_{i \in I}$

distribution correction ratio: $\zeta^*(s,a) = \frac{d^\pi(s,a)}{d_D(s,a)}$

average per-step reward value: $\rho(\pi) = \mathbb{E}_{(s,a,r)\sim d_D}[\zeta^*(s,a)\cdot r]$

03 Experimental Setup

(available at: https://github.com/compScienceYaren/dice_rl)

01
- Created datasets using Frozenlake-v0 each with a behaviour (α=0.0) and target (α=1.0) policy
  - for generalizability purposes, used 20 seeds (ranging from 0 to 19, inclusive)

02
- Implemented k-fold cross-fitting for the DICE estimators, under the name "SplitDICE"
  - training the estimator for 10000 steps
  - estimating the average per-step reward value at 100-step intervals

03
- Experimented naive (no sample-splitting), 2-fold cross-fitting and 5-fold cross-fitting strategies
- Built upon the configurations of BestDICE (as the best performer of all the existing DICE variants)

Figure 2: Naive DICE

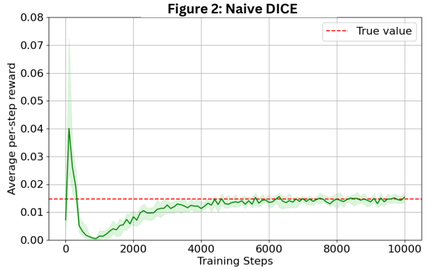
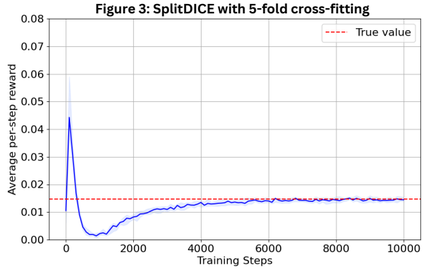
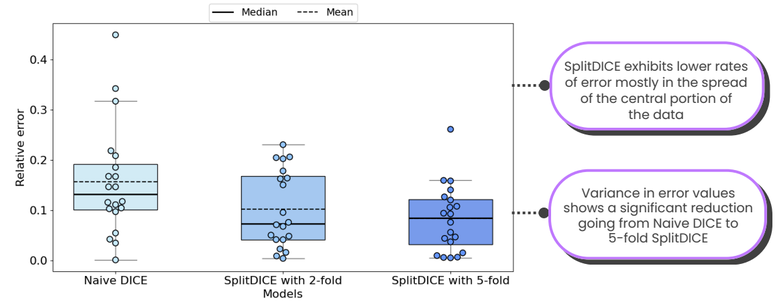


Figure 3: SplitDICE with 5-fold cross-fitting



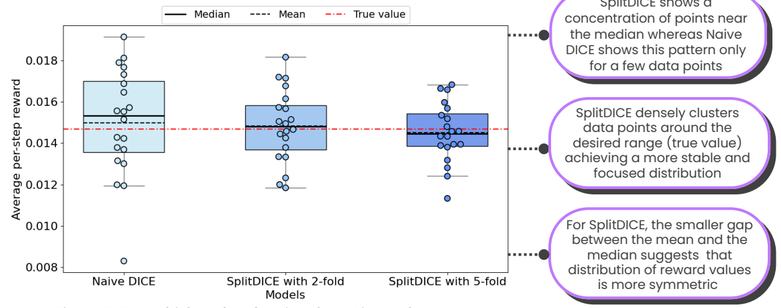
04 Analysis



SplitDICE exhibits lower rates of error mostly in the spread of the central portion of the data

Variance in error values shows a significant reduction going from Naive DICE to 5-fold SplitDICE

Figure 4: Box-whisker plot showing the relative error between the final estimated average per-step reward value and the ground truth



SplitDICE shows a concentration of points near the median whereas Naive DICE shows this pattern only for a few data points

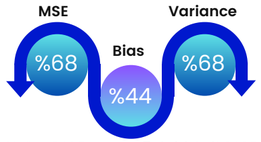
SplitDICE densely clusters data points around the desired range (true value) achieving a more stable and focused distribution

For SplitDICE, the smaller gap between the mean and the median suggests that distribution of reward values is more symmetric

Figure 5: Box-whisker plot showing the estimated average per-step reward value (calculated at step=10000)

Table 1: Performance measures for the considered estimators

Scenarios	MSE	Bias	Variance
Naive estimator	7.773960e-06	0.000309	0.035533
2-fold cross-fit	3.424816e-06	0.000144	0.015937
5-fold cross-fit	2.487964e-06	0.000172	0.011163



Naive DICE versus 5-fold SplitDICE (with 5-fold showing enhanced performance)

05 Conclusion and Future Work

REMARK 01
5-fold SplitDICE has **lower rates** of **relative error** than the Naive DICE at a **significance level of %10.**

REMARK 02
In overall performance measures for **MSE, bias and variance** generalized from the final estimate value of average per-step reward show a **descending trend** from Naive DICE to 5-fold SplitDICE.

REMARK 03
Variance calculated at the end of 10000 steps from all the observations show a **descending trend** going from Naive DICE to 2-fold SplitDICE and from 2-fold SplitDICE to 5-fold SplitDICE.

Complex Environments

Limitation: The datasets were generated using the Frozenlake environment.
Improvement: *Testing SplitDICE methods in more challenging environments such as Reacher, Cartpole etc. to improve scalability.*

Variants of the DICE Family

Limitation: The results obtained from SplitDICE builds upon the configurations of BestDICE.
Improvement: *Applying other variants of DICE such as DualDICE, GradientDICE etc. to test out varying configurations of primal/dual regularization and redundant constraints.*

References

[1] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-Policy Evaluation via the Regularized Lagrangian. July 2020.
[2] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1):C1–C68, 2018.
[3] Tali M. Ball, Lindsay M. Squeglia, Susan F. Tapert, and Martin P. Paulus. Double dipping in machine learning: Problems and solutions. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 5(3):261–263, March 2020.