Using Confounded Data in Offline RL

Anonymous Authors[1]

Abstract

In this work we consider the problem of confounding in offline RL, also called the delusion problem [12]. While it is known that learning from purely offline data is a hazardous endeavor in the presence of confounding, in this paper we show that offline, confounded data can be safely combined with online, non-confounded data to improve the sample-efficiency of model-based RL. We import ideas from the well-established framework of do-calculus to express model-based RL as a causal inference problem, thus bridging the fields of RL and causality. We propose a latent-based method which we prove is correct and efficient, in the sense that it attains better generalization guarantees thanks to the offline, confounded data (in the asymptotic case), regardless of the expert's behavior. We illustrate the effectiveness of our method on a series of synthetic experiments.

1. Introduction

As human beings, understanding cause and effect is crucial to successfully navigate our environment. This is also true of reinforcement learning (RL) agents, and in particular model-based agents, who must learn the effects of their own actions (interventions) on the environment. From this perspective, offline RL is analogous to the problem of causal inference from observational data, which requires assumptions about the data-generating process [1]. In the context of Markov Decision Processes (MDPs), estimating causal effects from offline data can be shown to be straightforward, due to the absence of confounding. In Partially-Observable MDPs (POMDPs) however, it requires the additional assumption that the data-collection agent did not use any privileged information besides that available to the learning agent. Relaxing this assumption results in confounded data, which off-the-shelf offline RL algorithms [10; 11] should not use.

A typical example is in the context of medicine, when offline data is collected from physicians who may rely on information absent from their patient's medical records, such as their wealthiness or their lifestyle. Suppose that wealthy patients in general get prescribed specific treatments by their physicians, because they can afford it, while being less at risk to develop severe conditions regardless of their treatment, because they can also afford a healthier lifestyle. This creates a spurious correlation called confounding, and will cause a naive recommender system to wrongly infer that costly treatments have positive health effects. Another example is in the context of autonomous driving, when offline data is collected from human drivers who have a wider field of vision than the camera on which the robot driver relies. Suppose human drivers push the brakes when they see a person waiting to cross the street, and only when the person walks in front of the car it enters the camera's field of vision. Then, again, a naive robot might wrongly infer that whenever the brakes are pushed, a person appears in front of the car. In order to minimize collisions with pedestrians, it might get regrettably reluctant to push the brakes.

Of course, in both those situations, the learning agent can infer the right causal model by disregarding the (confounded) offline data altogether, and by relying solely on online data instead, collected from its own interactions. However, in both those situations also, performing interventions for the sole purpose of seeing what happens is impractical, while collecting offline data by observing the behaviour of human agents is much more affordable. The question we address in this paper is, how can confounded, offline data be leveraged by an RL agent who can also collect online data?

To answer this question we import tools and ideas from the well-established field of causality [13] into the model-based RL framework. We formalize model-based RL as a causal inference problem using the framework of do-calculus [15], and we present a generic method for combining online and offline data in model-based RL, with a formal proof of correctness and efficiency even in the presence of confounding. We propose a practical implementation in the tabular setting, and present three experiments on synthetic toy problems that illustrate its effectiveness.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]"behind every causal conclusion there must lie some causal assumption that is not testable in observational studies", Pearl [14].

2. Model-based RL as causal inference

Decision-making problems are inherently causal [5]. Here we follow the *do*-calculus framework of Judea Pearl to reason about causality and confounding[2]. In this framework, queries of form $p(y|do(x), z)$ measure the effect of an action (or intervention) $do(X = x)$ on an outcome event $Y = y$, in the context where another event $Z = z$ is also observed.

2.1. POMDPs

We consider episodic Partially-Observable Markov Decision Processes (POMDPs) of the form $M = (\mathcal{S}, \mathcal{O}, \mathcal{A}, p_{init}, p_{trans}, p_{obs}, r)$, with hidden states $s \in \mathcal{S}$, observations $o \in \mathcal{O}$, actions $a \in \mathcal{A}$, initial and transition state distributions $p_{init}(s_0)$ and $p_{trans}(s_{t+1}|s_t, a_t)$, observation distribution $p_{obs}(o_t|s_t)$, and reward function $r : \mathcal{O} \to \mathbb{R}$. For simplicity we assume episodes $\tau = (o_0, a_0, \ldots, o_T)$ of finite length $|\tau| = T$, and we introduce the concept of a history at time t, $h_t = (o_0, a_0, \ldots, o_t)$. The control mechanism is represented as a stochastic policy π, which together with the POMDP dynamics p_{init}, p_{trans} and p_{obs} defines a probability distribution over trajectories, $p(\tau)$. In this work we consider two types of control policies.

Definition 2.1 (Standard policy)**.** In *standard control*, actions are decided based only on the available information from the past, according to a *standard policy* $\pi_{std}(a_t|h_t)$.

This corresponds to the standard POMDP regime, where the goal is to find an optimal standard policy, $\pi^\star_{std}$, which maximizes the expected long-term reward.

Definition 2.2 (Privileged policy)**.** In *privileged control*, actions can be decided based on the hidden state as well, according to a *privileged policy* $\pi_{std}(a_t|h_t, s_t)$.

This allows us to consider situations in offline RL where episodes are collected from an external agent who uses privileged information, in the extreme case the entire POMDP hidden state. This is the origin of the confounding problem.

2.2. Model-based RL

In model-based RL we are interested in the causal effect of immediate interventions ($do(a_t)$) on the system, given that past observations ($o_{0\to t}$) and past interventions ($do(a_{0\to t-1})$) have already happened. This is captured in $p(o_{t+1}|o_{0\to t}, do(a_{0\to t}))$[3], the POMDP causal transition model. Together with the initial distribution $p(o_0)$, it allows for the evaluation of any standard policy $\pi_{std}(a_t|h_t)$. We will see that, depending on whether a standard or privileged policy was used during the data-collection process, estimating this causal model from offline data can be straightforward (no confounding) or impossible (confounding).

[2]We refer the reader to [15] for an introduction to *do*-calculus.

[3]This notation can be found also in [12]

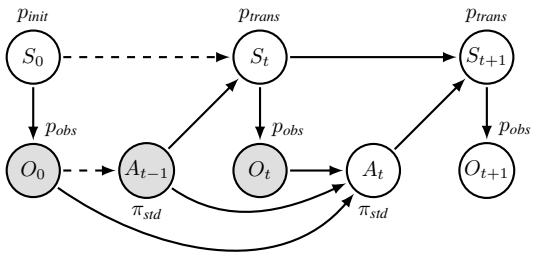


Figure 1. The standard POMDP setting (no confounding).

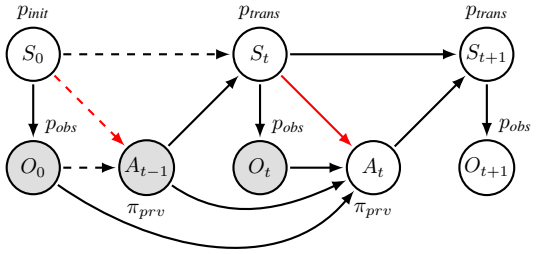


Figure 2. The privileged POMDP setting (confounding).

2.3. The interventional regime

Consider a dataset of episodes collected in a standard control setting, $\mathcal{D}_{int} \sim p_{std}(\tau)$, from policy π_{std}. The corresponding causal DAG is depicted in Figure 1. By applying *do*-calculus, the causal model is trivially identifiable as

$$p(o_{t+1}|o_{0\to t}, do(a_{0\to t})) = p_{std}(o_{t+1}|h_t, a_t).$$

Assuming that the control policy is strictly positive ($\pi_{std}(a_t|h_t) > 0$, $\forall a_t, h_t$), an unbiased estimator can be obtained from $\mathcal{D}_{int}$ via log-likelihood maximization,

$$\hat{q} = \arg\max_{q\in\mathcal{Q}} \sum_{\tau}^{\mathcal{D}_{int}} \sum_{t=0}^{|\tau|-1} \log q(o_{t+1}|h_t, a_t). \qquad (1)$$

This corresponds to a regular model-based RL approach.

2.4. The observational regime

Consider now a dataset of episodes collected in a privileged POMDP setting, $\mathcal{D}_{obs} \sim p_{prv}(\tau)$, from a privileged policy π_{prv}. The corresponding causal DAG is depicted in Figure 2. This time, *do*-calculus tells us that the causal model is not identifiable, because of the POMDP hidden states s_t which act as confounding variables.

2.5. Connection to online and offline RL

To relate the concepts of interventional and observational data to online and offline RL, the question to ask is, when the training samples were collected, could the control policy have used privileged information, besides the history h_t?

In online RL, the learning agent explicitly controls the data-collection policy, so by design it can not rely on privileged

information. Hence, data collected in an online RL setting can be safely treated as interventional.

In offline RL, the learning agent might have limited knowledge about the data-collection policy, sometimes no knowledge at all. In some settings it can be shown that the policy could not have used any privileged information, in which case the offline data can be treated as interventional. For example, with human replays from Atari video games, it is hard to imagine a human player having access to more information from the machine's internal state than the regular video and audio outputs from the game (the history h_t). But in more general settings, access to privileged information can not be dismissed. This is particularly true with human demonstrations collected in the wild, such as in the context of autonomous driving, medical recommender systems (examples in Section 1), or question answering systems [12]. In such settings the offline data must be treated as observational, and naively applying off-the-shelf offline RL algorithms ([10; 11]) is a risky endeavour.

3. Combining observational and interventional data

What about situations where both interventional and observational data is available (e.g., a mix of online and offline data)? Often, collecting a large observational dataset from demonstrations is relatively cheap (recording human drivers in the wild), while collecting an interventional dataset is expensive (recording a robot driver in the wild). How to leverage the observational data then, despite confounding?

3.1. Problem statement

We consider two datasets $\mathcal{D}_{int}$ and $\mathcal{D}_{obs}$ of POMDP episodes, sampled respectively using a standard and a privileged policy. We then ask the following question: can $\mathcal{D}_{obs}$ improve the estimator (1) obtained from $\mathcal{D}_{int}$ only?

3.2. The augmented learning procedure

Since both datasets are sampled from the same POMDP (p_{init}, p_{trans}, p_{obs}) controlled in different ways, we introduce a regime indicator [5] variable $I \in \{0, 1\}$ so that $\mathcal{D}_{obs} \sim p(\tau|i = 0)$ and $\mathcal{D}_{int} \sim p(\tau|i = 1)$, with the augmented control policy $\pi(a_t|h_t, s_t, i) = \pi_{prv}(a_t|h_t, s_t)$ when $i = 0$ and $\pi_{std}(a_t|h_t)$ when $i = 1$. This results in the augmented data-generating process from Figure 3.

In order to learn the causal transition model $p(o_{t+1}|o_{0\to t}, do(a_{0\to t})$ we propose the following two-step procedure, which relies on fitting a latent probabilistic model $\hat{q}$ to the training data. Our latent model is constrained to respect the structure of our augmented POMDP, with a latent variable $z_t \in \mathcal{Z}$ to substitute the hidden state $s_t \in \mathcal{S}$.

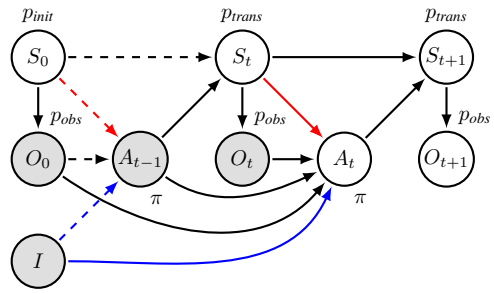


Figure 3. Augmented POMDP setting, with a policy regime indicator I taking values in $\{0, 1\}$ (1=interventional, no confounding, 0=observational, potential confounding), such that $\pi(a_t|h_t, s_t, i = 1) = \pi(a_t|h_t, i = 1)$, enforcing the contextual independence $A_t \perp\!\!\!\perp S_t \mid H_t, I = 1$.

Learning Our learning problem formulates as standard likelihood maximization,

$$\hat{q} = \arg\max_{q\in\mathcal{Q}} \sum_{(\tau)}^{\mathcal{D}_{obs}} \log q(\tau|i = 0) + \sum_{(\tau)}^{\mathcal{D}_{int}} \log q(\tau|i = 1). \quad (2)$$

In practice, the learnable model q decomposes into only four components: the initial latent model $q(z_0)$, the observation model $q(o_t|z_t)$, the latent transition model $q(z_{t+1}|z_t, a_t)$ and the privileged agent's behaviour model $q(a_t|z_t, i = 0)$.

Inference In a second step, we extract $\hat{q}(o_{t+1}|o_{0\to t}, do(a_{0\to t})$ by applying do-calculus on the DAG from Figure 3, with z_t instead of s_t. The procedure conveniently unrolls as a forward algorithm, and relies on the recurrent computation of $\hat{q}(z_t|h_t, i = 1)$, a.k.a. the agent's belief state at time t [2; 17].

Guarantees We now show that our two-step approach is correct and efficient (proofs deferred to Appendix A.3).

Proposition 3.1. *Assuming $|\mathcal{Z}| \geq |\mathcal{S}|$, $\hat{q}(o_{t+1}|o_{0\to t}, do(a_{0\to t})$ is an unbiased estimator of $p(o_{t+1}|o_{0\to t}, do(a_{0\to t})$.*

Proposition 3.2. *For any $\mathcal{D}_{int}$, the estimator $\hat{q}(o_{t+1}|o_{0\to t}, do(a_{0\to t})$ offers strictly better generalization guarantees when $|\mathcal{D}_{obs}| \to \infty$ than when $|\mathcal{D}_{obs}| = 0$.*

3.3. Relation to large sequence models

A recent trend in RL is to apply large sequence models to estimate the environment's dynamics in a model-based fashion [16; 8], or to parameterize a goal-conditioned policy in a model-free fashion [4; 18]. While large sequence models appear promising for efficiently combining offline and online data, they remain vulnerable to confounding, as pinpointed by Ortega et al. [12]. Because it follows a generic model-based approach, our method could be easily combined with a large sequence model to address large-scale RL scenarios, while being robust to confounding.

4. Experiments

We compare our *augmented* method against two baselines: *no obs*, where $\mathcal{D}_{obs}$ is not used at all during training, and *naive*, where $\mathcal{D}_{obs}$ is naively combined with $\mathcal{D}_{int}$ as if there was no confounding. For each method we employ the same tabular model and training procedure to learn $\hat{q}$, only the datasets used in Equation (2) change. We then train an agent via actor-critic using the POMDP causal transition model $\hat{q}$ recovered by each method, and we evaluate it in the real test environment. The whole procedure is similar to that of [7].

We report experiments on two synthetic toy problems, *tiger* and *sloppy gridworld*. For each toy problem, we collect a fixed observational dataset $\mathcal{D}_{obs}$ of 2^{13} (8192) samples, and we vary the size of the interventional dataset $\mathcal{D}_{int}$ on a logarithmic scale from 2^2 (4) to 2^{13} (8192). We repeat each experiment 20 times from random seeds.

Tiger is a classic, small-scale POMDP from Cassandra et al. [3] with $|\mathcal{S}| = 6$ hidden states and time horizon $T = 50$. The learning agent can listen to collect noisy information about the tiger's position, or open a door (left or right), which triggers a +10 (treasure) or -100 (tiger) reward and resets the tiger to a random position. The privileged agent has full knowledge of the tiger's position, and acts as follows

	action		
tiger	listen	open left	open right
Left	0.05	0.3	0.65
right	0.05	0.8	0.15

$$\pi_{prv}(action|tiger)$$

As can be seen in Figure 4, the *no obs* method learns a good policy around $|\mathcal{D}_{int}| = 2^5$, while our *augmented* method immediately starts with a good policy at $|\mathcal{D}_{int}| = 2^2$ thanks to the observational data. The *naive* method is hurt by confounding, and only starts overcoming the initial bias at around $|\mathcal{D}_{int}| = 2^{11}$.

Sloppy gridworld inspired from Alt et al. [1] constitutes a more challenging POMDP with $|\mathcal{S}| = 21$ hidden states and time horizon $T = 20$. Here the agent starts on the top-left corner of a small 5x5 grid, and tries to get to a target at the bottom of a large wall. The actions are *top*, *right*, *bottom*, *left* and *idle*, which correctly move the agent with 50% chances, or move the agent randomly otherwise. The privileged agent has full knowledge of its position at each time step, and uses a shortest-path algorithm to decide on its next action. The learning agent is only revealed its position with 20% chances at each time step, and is blind otherwise (a dummy position is revealed). Here again, from Figure 4 it is clear that our *augmented* method benefits from the observational data compared to the *no obs* method, while the *naive* method is hurt due to the confounding. In Figure 5 we present heatmaps of the test-time trajectories of the RL

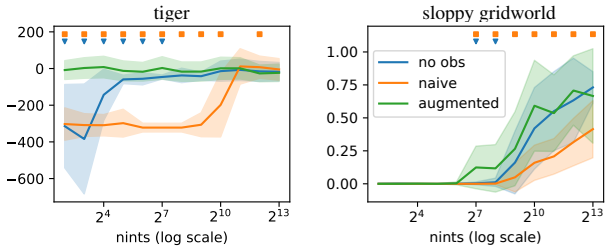


Figure 4. Performance of each RL agent on our two toy problems (the higher the better) as the number of interventional data increases. We report the average cumulative reward (mean $\pm$ std) obtained on the real environment. Little markers indicate a significant difference between our *augmented* method and the baselines *no obs* (down triangles) and *naive* (squares), using a two-sided Wilcoxon signed-rank test [6] with $\alpha < 5\%$.

agents obtained at the $|\mathcal{D}_{int}| = 2^7$ mark (128 samples), for each method. At this point, only our *augmented* method starts passing the wall and reaching the objective (positive reward), while the two other methods struggle escaping the initial position.

More experimental details are available in Appendix A.1 and Appendix A.2, which also presents a third experiment in the simple binary bandit setting. There we also investigate the robustness of our method to various types of confounding, that is, different privileged policies (expert behaviours).

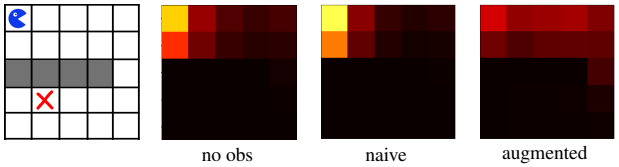


Figure 5. Focus on the sloppy gridworld experiment at the $|\mathcal{D}_{int}| = 2^7$ mark. **Left**: the initial grid. **Right**: a heatmap of the tiles visited by the RL agents at test time. At this point, only the *augmented* method has learned how to pass the wall.

5. Conclusion

In this paper we have presented a simple, generic method for combining interventional and observational (potentially confounded) data in model-based reinforcement learning for POMDPs. We have demonstrated that our method is correct and efficient in the asymptotic case (infinite observational data), and we have illustrated its effectiveness on two synthetic toy problems. A future direction is to investigate this method in the high-dimensional POMDP setting, where learning a latent-based transition model is more challenging. We hope that this work will help bridge the gap between the fields of RL and causality, and will convince the RL community that causality is an adequate tool to reason about observational data, which is plentiful in the world.

References

[1] Alt, B., Schultheis, M., and Koeppl, H. Pomdps in continuous time and discrete spaces. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13151–13162. Curran Associates, Inc., 2020.

[2] Cassandra, A. R. *Exact and approximate algorithms for partially observable Markov decision processes*. PhD thesis, Brown University, 1998.

[3] Cassandra, A. R., Kaelbling, L. P., and Littman, M. L. Acting optimally in partially observable stochastic domains. In *AAAI*, 1994.

[4] Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 15084–15097. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/7f489f642a0ddb10272b5c31057f0663-Paper.pdf.

[5] Dawid, P. Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, 9(1):39–77, 2021. doi: doi:10.1515/jci-2020-0008.

[6] Demsar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, 2006. URL http://dblp.uni-trier.de/db/journals/jmlr/jmlr7.html#Demsar06.

[7] Hafner, D., Lillicrap, T. P., Norouzi, M., and Ba, J. Mastering atari with discrete world models. In *ICLR*, 2021.

[8] Janner, M., Li, Q., and Levine, S. Offline reinforcement learning as one big sequence modeling problem. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1273–1286. Curran Associates, Inc., 2021.

[9] Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[10] Lange, S., Gabel, T., and Riedmiller, M. A. Batch reinforcement learning. In *Reinforcement Learning*, volume 12 of *Adaptation, Learning, and Optimization*, pp. 45–73. Springer, 2012.

[11] Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems, 2020.

[12] Ortega, P. A., Kunesch, M., Delétang, G., Genewein, T., Grau-Moya, J., Veness, J., Buchli, J., Degrave, J., Piot, B., Perolat, J., Everitt, T., Tallec, C., Parisotto, E., Erez, T., Chen, Y., Reed, S., Hutter, M., de Freitas, N., and Legg, S. Shaking the foundations: delusions in sequence models for interaction and control, 2021.

[13] Pearl, J. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009.

[14] Pearl, J. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96 – 146, 2009.

[15] Pearl, J. The do-calculus revisited. In *UAI*, 2012.

[16] Schrittwieser, J., Hubert, T., Mandhane, A., Barekatain, M., Antonoglou, I., and Silver, D. Online and offline reinforcement learning by planning with a learned model. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27580–27591. Curran Associates, Inc., 2021.

[17] Striebel, C. Sufficient statistics in the optimum control of stochastic systems. *Journal of Mathematical Analysis and Applications*, 12(3):576–592, 1965. ISSN 0022-247X. doi: 10.1016/0022-247X(65)90027-2.

[18] Zheng, Q., Zhang, A., and Grover, A. Online decision transformer. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27042–27059. PMLR, 17–23 Jul 2022.

A. Appendix

A.1. Experimental details

The code for reproducing our experiments is made available online[4].

We perform experiments on three synthetic toy problems: the *door* problem described earlier (Figure 20), the classical *tiger* problem from the literature [3], and a 5x5 *gridworld* problem inspired from Alt et al. [1].

Data To assess the performance of our method, we consider a large observational dataset $\mathcal{D}_{obs}$ of fixed size (512 samples for *door*, 8192 samples for *tiger* and *gridworld*), and an interventional dataset $\mathcal{D}_{int}$ of varying size, ranging on an exponential scale from 4 to $|\mathcal{D}_{obs}|$.

Baselines We compare the performance of the transition model $\hat{q}$ recovered in three different settings: *no obs*, when only interventional data ($\mathcal{D} = \mathcal{D}_{int}$) is used for training; *naive*, when observational data is naively combined with interventional data as if there was no confounding ($\mathcal{D} = \mathcal{D}_{int} \cup \{(\tau, 1) | (\tau, i) \in \mathcal{D}_{obs})\}$); and *augmented*, our proposed method ($\mathcal{D} = \mathcal{D}_{int} \cup \mathcal{D}_{obs}$). Note that the only difference between each of those settings is the training dataset, all other aspects (learning procedure, model architecture, loss function) begin the same.

Training In all our experiments we use a tabular model for $\hat{q}$, that is, we use discrete probability tables for each building blocs of the transition model, $q(z_0)$, $q(o_t|z_t)$, $q(z_{t+1}|z_t, a_t)$, and $q(a_t|h_t, z_t, i = 0)$. We use a latent space $|\mathcal{Z}|$ of size 32, 32 and 128 respectively for each toy problem, while the true latent space $|\mathcal{S}|$ is of size 3, 6 and 42. We train $\hat{q}$ by directly minimizing the negative log likelihood (2) via gradient descent. We use the Adam optimizer [9] with a learning rate of 10^{-2}, and train for 500 epochs consisting of 50 gradient descent steps with minibatches of size 32. We divide the learning rate by 10 after 10 epochs without loss improvement (reduce on plateau), and we stop training after 20 epochs without improvement (early stopping). In the *door* experiment we derive the optimal policy $\hat{\pi}^\star$ exactly, while in the *tiger* and *gridworld* experiments we train a "dreamer" RL agent on imaginary samples $\tau \sim \hat{q}(\tau|i = 1)$ obtained from the model, using the belief states $\hat{q}(s_t|h_t)$ as features. We use a simple Actor-Critic algorithm for training, and our agents consist of a simple MLP with one hidden layers for both the critic and the policy parts. RL agents are trained until convergence or with a maximum number of 1000 epochs, with a learning rate of 10^{-2}, a discount factor $\gamma = 0.9$ and a batch size of 8.

JS divergence To evaluate the general quality of the recovered transition models, we compute the expected Jensen-Shannon divergence between the learned $\hat{q}(o_{t+1}|h_t, i = 1)$ and the true $p(o_{t+1}|h_t, i = 1)$, over transitions generated using a uniformly random policy π_{rand},

$$\frac{1}{2}\mathbb{E}_{\tau\sim p_{init}, p_{trans}, p_{obs}, \pi_{rand}}\left[\log \frac{p(o_0)}{m(o_0)} + \sum_{t=1}^{|\tau|} \log \frac{p(o_{t+1}|h_t, i = 1)}{m(o_{t+1}|h_t, i = 1)}\right]$$
$$+\frac{1}{2}\mathbb{E}_{\tau\sim \hat{q}_{init}, \hat{q}_{trans}, \hat{q}_{obs}, \pi_{rand}}\left[\log \frac{\hat{q}(o_0)}{m(o_0)} + \sum_{t=1}^{|\tau|} \log \frac{\hat{q}(o_{t+1}|h_t, i = 1)}{m(o_{t+1}|h_t, i = 1)}\right],$$

where $m(.) = \frac{1}{2}\left(\hat{q}(.) + p(.)\right)$. In the first experiment we compute the JS exactly, while in the second experiment we use a stochastic approximation over 100 trajectories τ to estimate each of the expectation terms in the JS empirically.

Reward. To evaluate quality of the recovered transition models for solving the original RL task, that is, maximizing the expected long-term reward, we evaluate the policy $\hat{\pi}^\star$, obtained after planning with the recovered model $\hat{q}$, on the true environment p,

$$\mathbb{E}_{\tau\sim p_{init}, p_{trans}, p_{obs}, \hat{\pi}^\star}\left[\sum_{t=0}^{|\tau|} R(o_t)\right].$$

In the first experiment we compute this expectation exactly, while in the second experiment we use a stochastic approximation using 100 trajectories τ.

[4] `https://supplementary.materials/disclosed.after.acceptance`

A.2. Complete empirical results

A.2.1. Door experiment

The *door* experiment (Figure 20) corresponds to a simple binary bandit setting, that is, a specific POMDP with horizon $H = 1$. The observation space is of size $|\mathcal{O}| = 0$, since the learning agent receives no observation, and the hidden state space is of minimal size $|\mathcal{S}| = 3$ to encode both the initial light color and the reward obtained afterwards. The bandit dynamics are described in Table 1.

light	
red	green
0.6	0.4

$p(\textit{light})$

		door	
light	*button*	closed	open
red	A	0.0	1.0
	B	1.0	0.0
green	A	1.0	0.0
	B	0.0	1.0

$p(\textit{door}|\textit{light}, \textit{button})$

Table 1. Probability tables for our *door* bandit problem.

We repeat the *door* experiment in six different scenarios, corresponding to different privileged policies π_{prv} ranging from a totally random agent to a perfectly good and a perfectly bad agent. Each time, we evaluate the performance of the *no obs*, *naive* and *augmented* approaches under different data regimes, by varying the sample size for both the observational data $\mathcal{D}_{obs}$ and the interventional data $\mathcal{D}_{int}$ in the range $(4, 8, 16, 32, 64, 128, 256, 512)$.

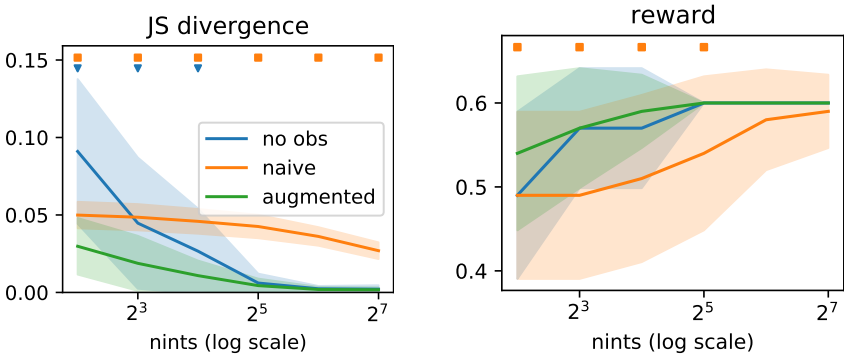
In each scenario, we report both the expected reward and the JS as heatmaps with $|\mathcal{D}_{int}|$ and $|\mathcal{D}_{obs}|$ in the x-axis and y-axis respectively, to highlight the combined effect of the sample sizes on each approach. We also provide as a heatmap the difference between our approach, *augmented*, and the two other approaches *no obs* and *naive*. We always plot the expected reward in the first row, and JS in the second row. As a remark, shades of green show gains in reward (the higher the better), while shades of purple show gains in JS (the lower the better).

Finally, we also present two plots which provide a focus on the data regime that corresponds to the largest number of observational data ($|\mathcal{D}_{obs}| = 512$), as in the main paper.

Noisy Good Expert In the noisy good expert setting, the expert plays halfway between a perfect and a random policy. The diversity of its action leads to a good start for the *naive* model but the bias it contains is hard to overcome. In contrast, our method makes good use of the observational data from the start and is also able to correct the bias as interventional data come in, eventually converging towards the true transition model.

$\pi_{prv}(button|light)$

	button	
light	A	B
red	0.9	0.1
green	0.4	0.6



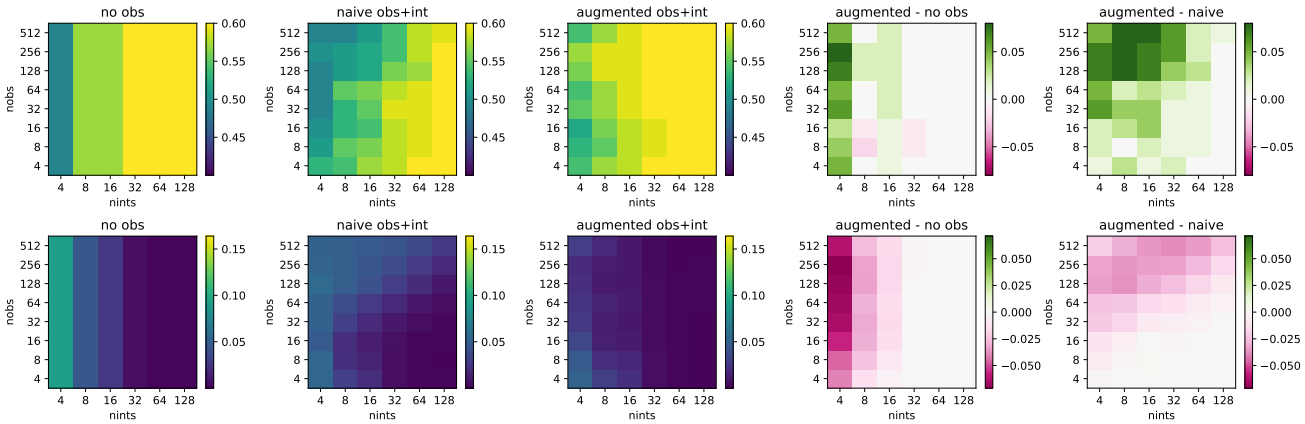


Figure 6. Noisy good expert setting. Heatmaps correspond respectively to the expected reward (top row, higher is better) and the JS divergence (bottom row, lower is better).

Random Expert A random policy naturally results in unconfounded observational data, since it does not exploits the privileged information. Hence, the *naive* approach is unbiased in this case, and actually makes the best use of the observational data. Our approach, *augmented*, exhibits an overall comparable performance, only slightly worse at times. We believe this can be explained by the additional complexity of our method which tries to disentangle a confounded regime in the data, and is not best suited to unconfounded data.

$\pi_{prv}(button|light)$

	button	
light	A	B
red	0.5	0.5
green	0.5	0.5

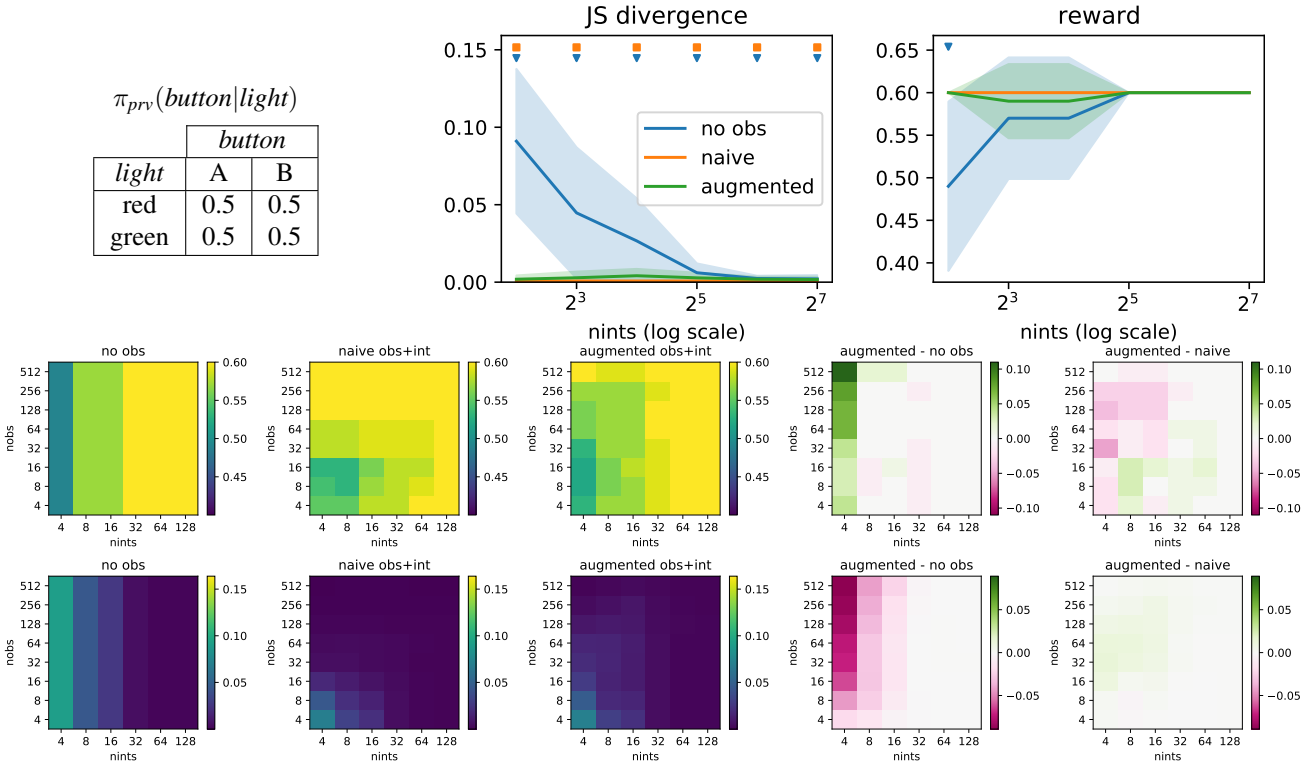


Figure 7. Random expert setting. Heatmaps correspond respectively to the expected reward (top row, higher is better) and the JS divergence (bottom row, lower is better).

Perfectly Good Expert Observing a perfectly good expert playing in the *door* problem induces a strong bias, because every observed action always results in a positive reward. As such, the *naive* approach struggles to learn a good transition model. The bias however is quickly corrected by our *augmented* approach, which eventually converges to the true transition model faster than the *no obs* approach.

$\pi_{prv}(button|light)$

	button	
light	A	B
red	1.0	0.0
green	0.0	1.0

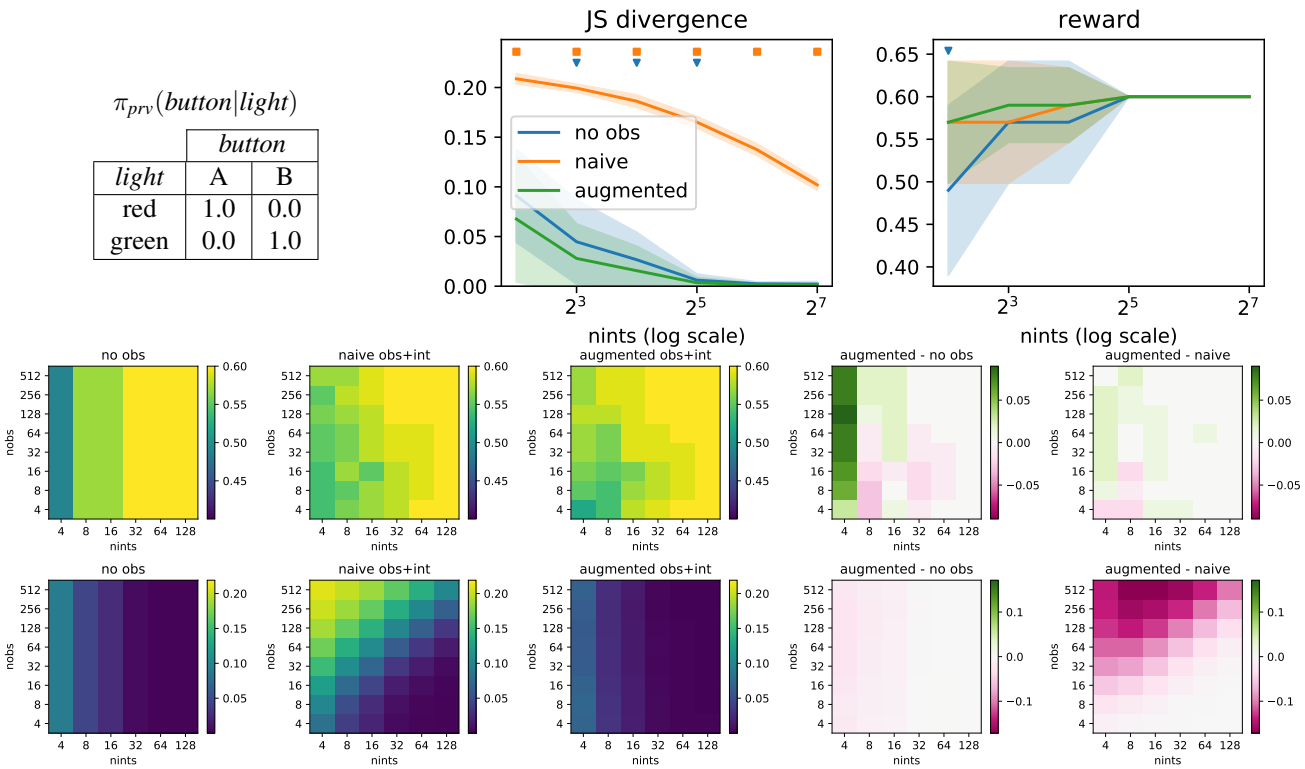


Figure 8. Perfectly good expert setting. Heatmaps correspond respectively to the expected reward (top row, higher is better) and the JS divergence (bottom row, lower is better).

Perfectly Bad Expert Similarly to the previous setting, observing an expert that always chooses a bad action leads to a strong bias, as every action is associated to a low reward. The behaviour in terms of JS and reward is similar as well.

$\pi_{prv}(button|light)$

	button	
light	A	B
red	0.0	1.0
green	1.0	0.0

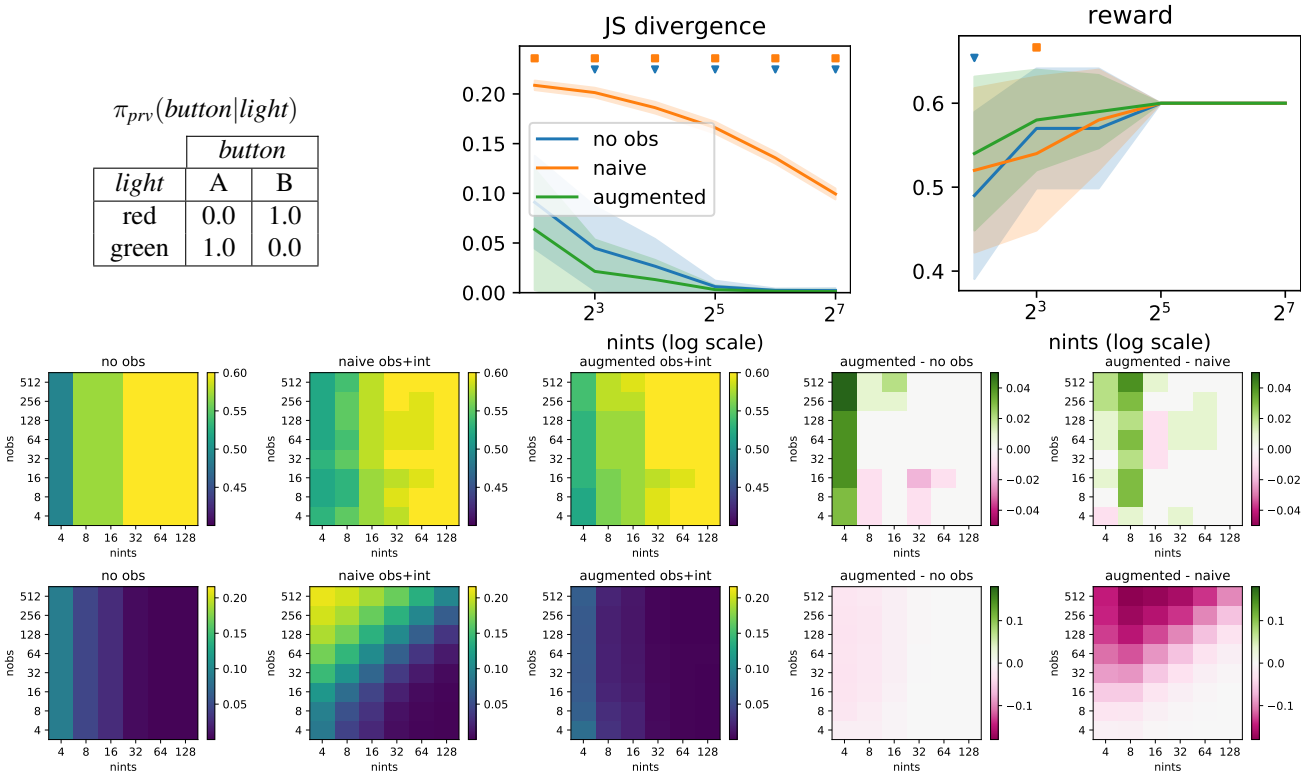


Figure 9. Perfectly bad expert setting. Heatmaps correspond respectively to the expected reward (top row, higher is better) and the JS divergence (bottom row, lower is better).

Positively Biased Expert Here the expert's policy is considered as *positively biased* in the sense that the agent will only obtain a positive reward when playing button A (with 55% chances) and never by playing button B (0% chances). Because playing button A is actually the optimal policy, this strong bias has a positive effect on the reward for the *naive* approach. Hence, although worse in terms of JS than our approach, the *naive* approach always results in a very good policy in terms of reward. Our *augmented* approach, however, seems more conservative.

$\pi_{prv}(button|light)$

	button	
light	A	B
red	0.8	0.2
green	1.0	0.0

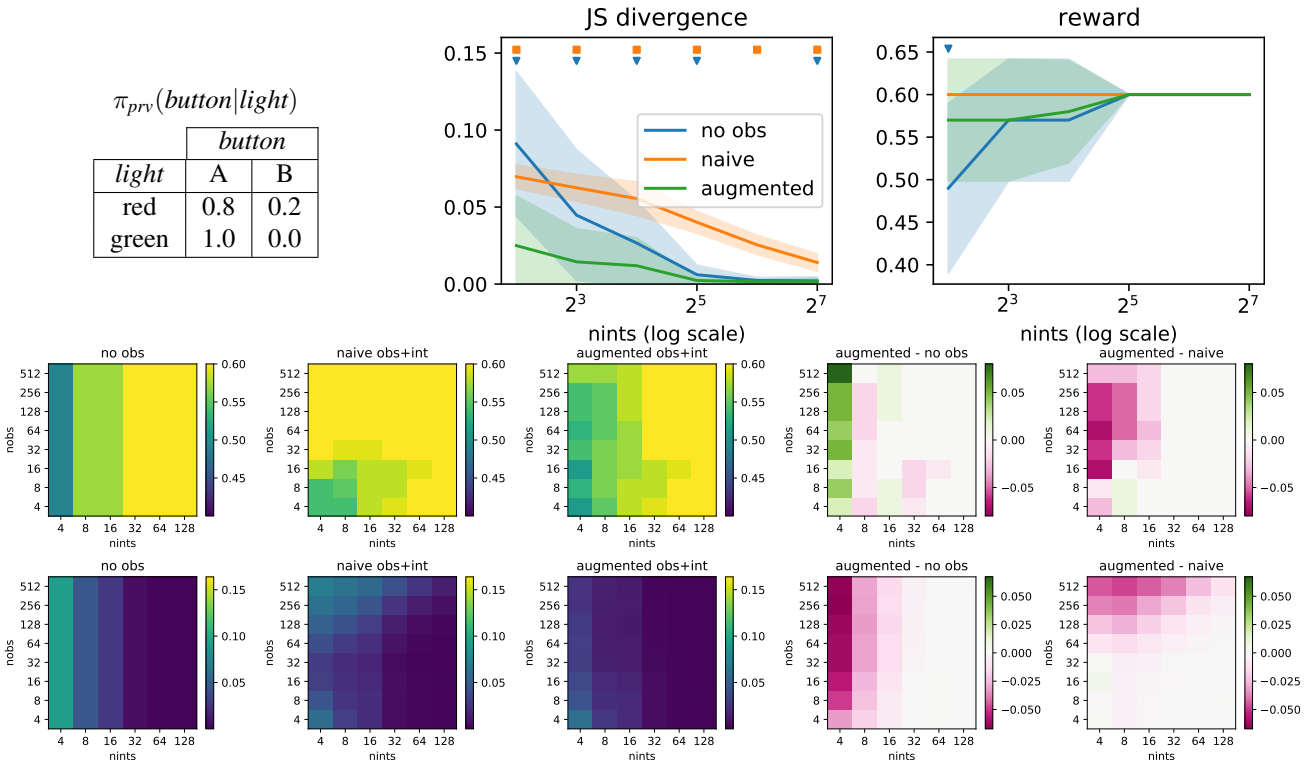


Figure 10. Positively biased expert setting. Heatmaps correspond respectively to the expected reward (top row, higher is better) and the JS divergence (bottom row, lower is better).

Negatively Biased Expert In an analogous way, a negatively biased expert will overuse button A, leading to mixed feelings regarding this button, whereas it will always get a positive reward each time it uses button B. This leads to the opposite behavior as we had in the previous setting, with the *naive* approach always favoring the use of button B, and obtaining a bad performance in terms of reward. The *naive* approach only gets better when a lot of interventional data is combined with the biased observational data, while our *augmented* approach is able to overcome this pessimistic bias very early on, and converges faster than both *no obs* and *naive*.

$\pi_{prv}(button|light)$

	button	
light	A	B
red	1.0	0.0
green	0.8	0.2

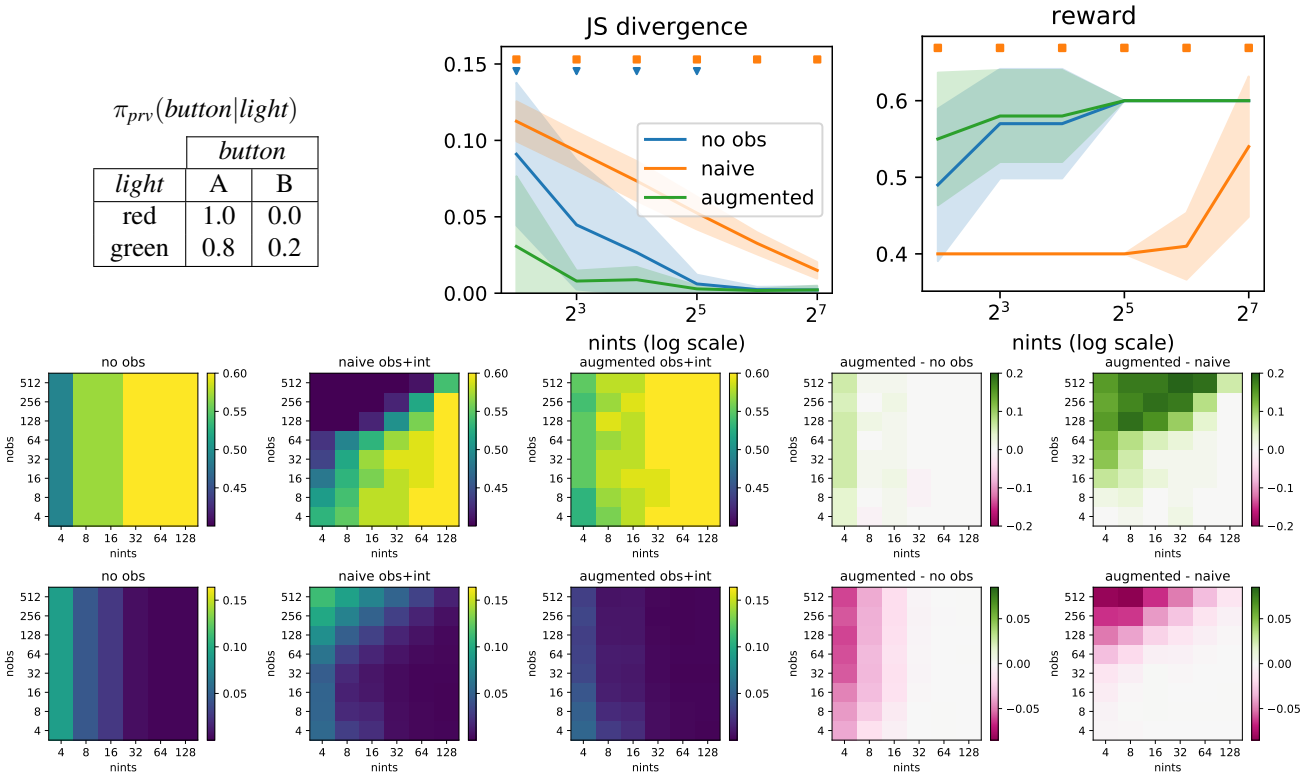


Figure 11. Pessimistic bias expert setting. Heatmaps correspond respectively to the expected reward (top row, higher is better) and the JS divergence (bottom row, lower is better).

A.2.2. TIGER EXPERIMENT

The *tiger* experiment corresponds a synthetic POMDP toy problem proposed by Cassandra et al. [3]. In short, in this problem the agent stands in front of two doors to open, one of them having a tiger behind it (-100 reward), and the other one a treasure (+10 reward). The agent also gets a noisy observation of the system in the form of the roar from the tiger, which seems to originate from the correct door most of the time (85% chances) and the wrong door sometimes (15% chances). In order to reduce uncertainty the agent can listen to the tiger's roar again, at the cost of a small penalty (-1). We present the simplified POMDP dynamics in Table 2, and in our experiments we impose a fixed horizon of size $H = 50$. The observation space is of size $|\mathcal{O}| = 6$, to encode the roar location perceived by the agent and the obtained reward, $o_t = (roar_t, reward_t)$, and the hidden state space is of minimal size $|\mathcal{S}| = 6$ to encode both the tiger position and the reward obtained at each time step, $s_t = (tiger_t, reward_t)$.

$tiger_0$	
left	right
0.5	0.5

$p(tiger_0)$

	$roar_t$	
$tiger_t$	left	right
left	0.85	0.15
right	0.15	0.85

$p(roar_t|tiger_t)$

		$tiger_{t+1}$	
$tiger_t$	$action_t$	left	right
left	listen	1.0	0.0
	open left	0.5	0.5
	open right	0.5	0.5
right	listen	0.0	1.0
	open left	0.5	0.5
	open right	0.5	0.5

$p(tiger_{t+1}|tiger_t, action_t)$

		$reward_{t+1}$		
$tiger_t$	$action_t$	-1	-100	+10
left	listen	1.0	0.0	0.0
	open left	0.0	1.0	0.0
	open right	0.0	0.0	1.0
right	listen	1.0	0.0	0.0
	open left	0.0	0.0	1.0
	open right	0.0	1.0	0.0

$p(reward_{t+1}|tiger_t, action_t)$

Table 2. Probability tables for the *tiger* problem.

For the tiger experiment we consider four different privileged policies π_{prv} for the observed agent. We then evaluate the performance of the *no obs*, *naive* and *augmented* approaches under different data regimes, by keeping the observational data fixed to $|\mathcal{D}_{obs}| = 8192$ while varying the varying the number of interventional data for $\mathcal{D}_{int}$ in the range $(4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192)$.

Noisy Good Expert In this scenario the privileged expert adopts a policy that plays the optimal action most of the time (open the treasure door), but also sometimes decides to just listen or to open the wrong door. As can be seen, in this scenario our *augmented* method makes the best use of the observational data, and is significantly better than both the *no obs* and *naive* approaches in the low-sample regime, both in terms of quality of the estimated transition model and obtained reward.

	action		
tiger	listen	open left	open right
Left	0.05	0.3	0.65
right	0.05	0.8	0.15

$\pi_{prv}(action|tiger)$

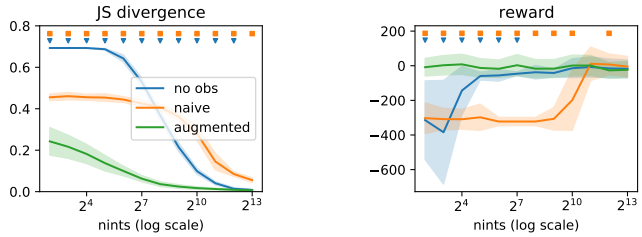


Figure 12. Noisy good agent.

Random Expert In the random scenario there is no confounding, and observational data can be safely mixed with interventional data. The *naive* approach thus does not suffer from any bias, and in fact is the one that converges the fastest to the optimal transition model and policy. Our method, while it manages to leverage the observational data to converge faster than *no obs*, suffers from a worse performance than *naive* in the low sample regime, most likely because it tries to recover a spurious confounding variable to distinguish the observational and interventional regimes, when none actually exists.

	action		
tiger	listen	open left	open right
left	0.33	0.33	0.33
right	0.33	0.33	0.33

$\pi_{prv}(action|tiger)$

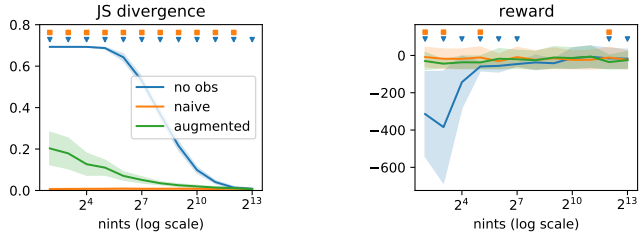


Figure 13. Random agent.

Very Good Expert Here the privileged expert never opens the wrong door, and thus never receives the very penalizing -100 reward. As a result the *naive* approach seems to be overly optimistic with respect to the action of opening a door, which strongly affects the expected reward it obtains in the true environment. While our *augmented* approach seems also to suffer from this bias in the very low sample regime (as can be seen on the reward plot), overall the quality of the recovered transition model is still superior to both other approaches, and converges faster to the true transition model.

	action		
tiger	listen	open left	open right
left	0.05	0.0	0.95
right	0.05	0.95	0.0

$\pi_{prv}(action|tiger)$

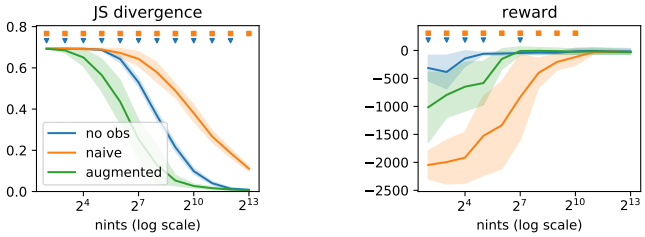


Figure 14. Very good agent.

Very Bad Expert Here the privileged expert never opens the correct door, and thus never receives a positive reward (+10). As a result, the *naive* approach seems to be very conservative, and prefers not to take any chances opening a door. It turns out that this strategy is not too bad in terms of reward (always listening yields a -51 total reward), and as such this causal bias seems to positively affect the performance of the *naive* approach in the low sample regime, but prevents it from obtaining a better policy in the high sample regime too. Our *augmented* method, on the other hand, is able to escape this overly conservative strategy earlier on, and converges to a good-performing policy faster than both other approaches.

	action		
tiger	listen	open left	open right
left	0.05	0.95	0.0
right	0.05	0.0	0.95

$\pi_{prv}(action|tiger)$

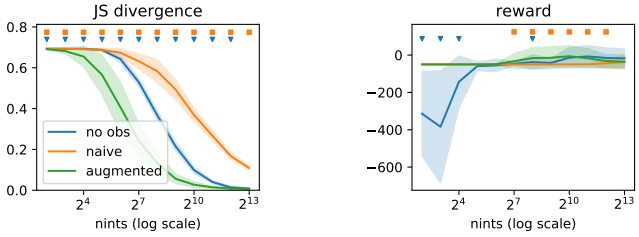


Figure 15. Very bad agent.

A.2.3. GRIDWORLD EXPERIMENT

The *gridworld* experiment, represented in Figure 16, is inspired from [1]. It consists in a small 5x5 grid where the agent starts on the top-left corner, and tries to get to a target placed on the bottom side behind a large wall. The agent can use five actions: *top*, *right*, *bottom*, *left* and *idle*, and only receives a noisy signal about its current position. At each time step, the agent's position is revealed with 20% chances, and remains completely hidden otherwise. In addition, the agent's actions only have a stochastic effect, i.e., the agent moves into the desired direction with 50% chances, and otherwise slips at random to one of the 5 adjacent tiles or current tile. In case the agent would bump into a wall, it simply remains at its current position. The observation space is of size $|\mathcal{O}| = 44$, to encode both the agent's location (or the indication that the location is hidden) and the reward, and the hidden state space is of size $|\mathcal{S}| = 21$ to encode the agent's location. In this experiment we impose a fixed horizon of size $H = 20$.

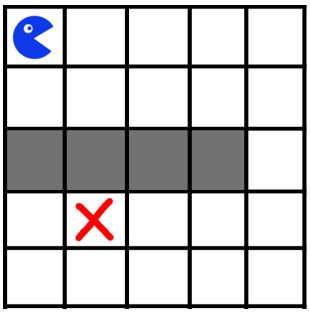


Figure 16. The gridworld problem

For the gridworld experiment we consider a single policy π_{prv} for the privileged agent, who acts optimally (shortest path from current location to target). We then evaluate the performance of the *no obs*, *naive* and *augmented* approaches under different data regimes, by keeping the observational data fixed to $|\mathcal{D}_{obs}| = 8192$ while varying the varying the number of interventional data for $\mathcal{D}_{int}$ in the range $(4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192)$.

Very Good Expert In this scenario the privileged agent adopts a perfect policy, and always chooses an action leading to the shortest path towards the target. As can be seen, here again our *augmented* method makes the best use of the observational data, and converges faster than both the *no obs* and the *naive* approaches for recovering the true transition model. This improvement in the transition model also translates into an improvement in terms of the learned policy, which starts converging towards high reward values with fewer samples (2^7) than both *no obs* and *naive* (2^9).

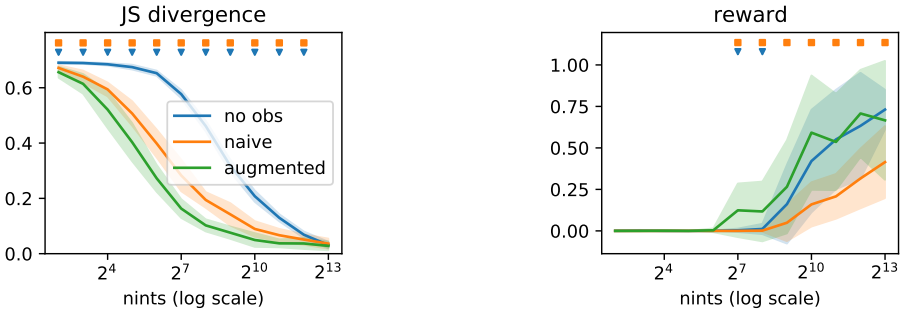


Figure 17. Perfect agent.

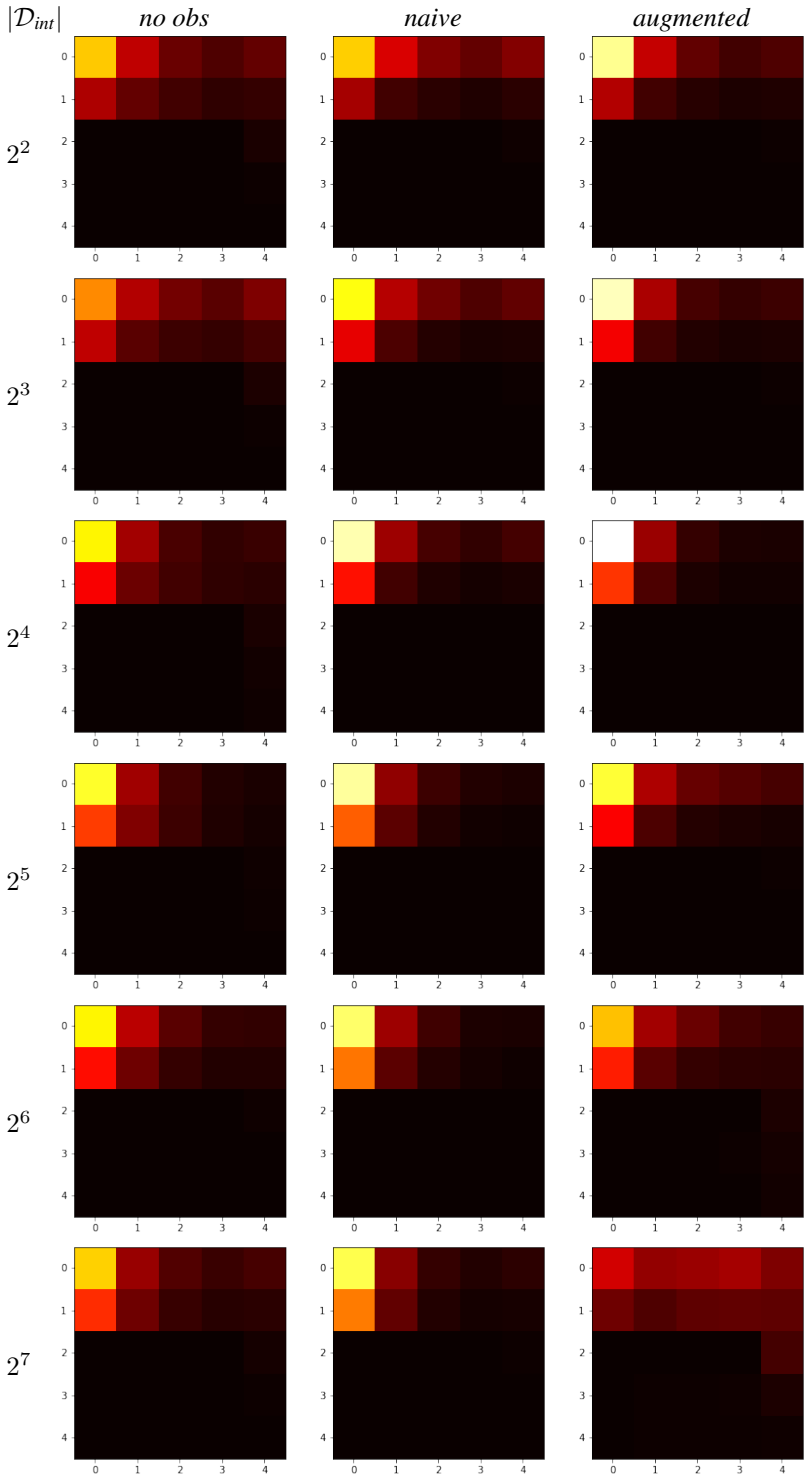


Figure 18. Average heat-maps over 100 episodes $\times$ 10 seeds, of the tiles visited by each trained agent (*no obs*, *naive*, *augmented*) for different interventional data sizes (2^2, 2^3, 2^4, 2^5, 2^6, 2^7). The *augmented* approach is the fastest (in terms of interventional data) to learn how to properly escape the top part of the maze through tile (4, 2), and then move towards the treasure on tile (1, 3).

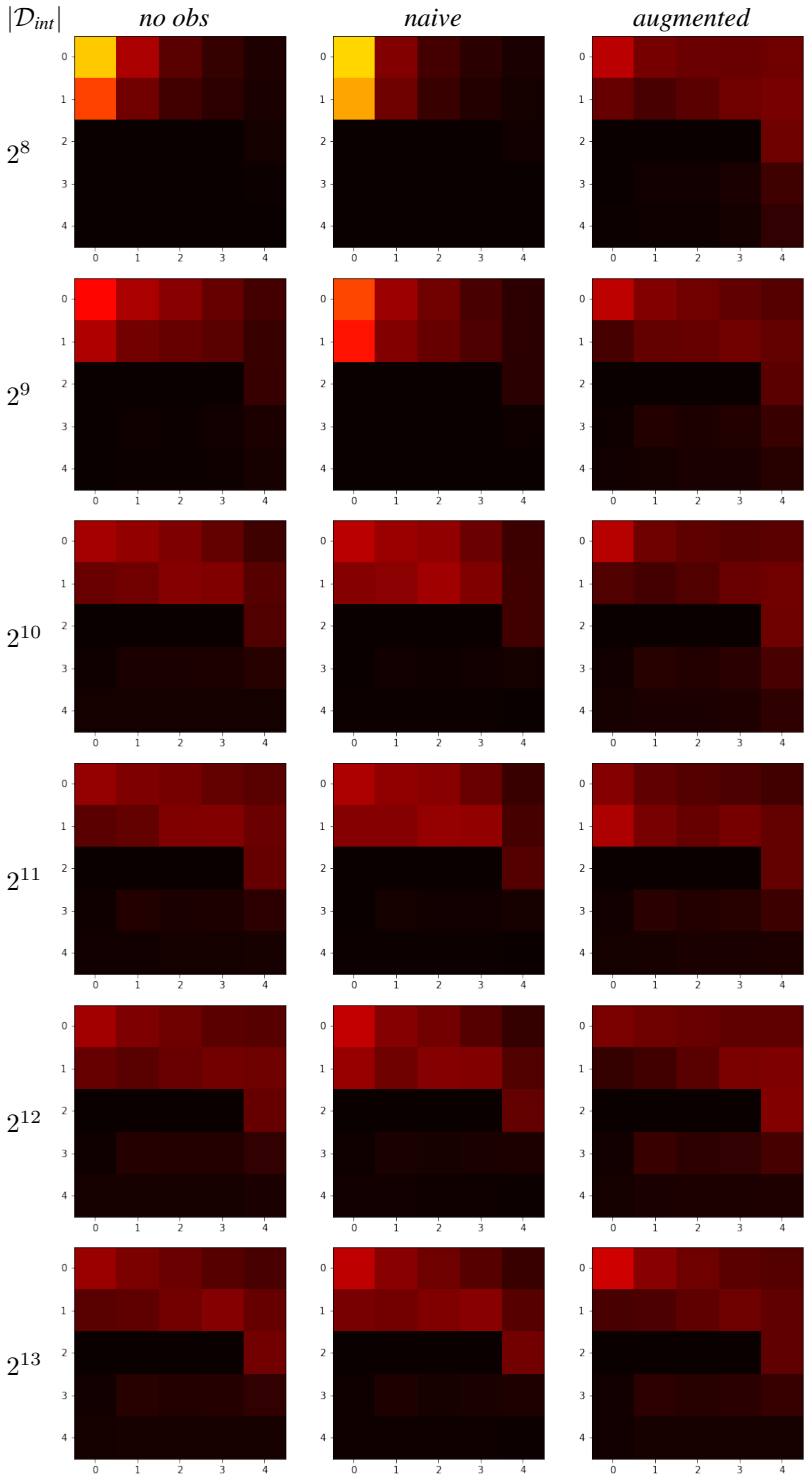


Figure 19. Average heat-maps over 100 episodes $\times$ 10 seeds, of the tiles visited by each trained agent (*no obs*, *naive*, *augmented*) for different interventional data sizes (2^8, 2^9, 2^{10}, 2^{11}, 2^{12}, 2^{13}). The *augmented* approach is the fastest (in terms of interventional data) to learn how to properly escape the top part of the maze through tile (4, 2), and then move towards the treasure on tile (1, 3).

A.3. Proofs.

Proposition 3.1. *Assuming $|\mathcal{Z}| \geq |\mathcal{S}|$, $\hat{q}(o_{t+1}|o_{0\to t}, do(a_{0\to t})$ is an unbiased estimator of $p(o_{t+1}|o_{0\to t}, do(a_{0\to t})$.*

Proof. The proof is straightforward. First, we have that $\mathcal{D} \sim p(\tau, i)$. Second, we have $p \in \mathcal{Q}$, because $\mathcal{Q}$ is only restricted to the augmented POMDP constraints, and because its latent space is sufficiently large ($|\mathcal{Z}| \geq |\mathcal{S}|$). Therefore, $\hat{q}(\tau, i)$ solution of (2) is an unbiased estimator of $p(\tau, i)$, and in particular $\hat{q}(o_{t+1}|h_t, a_t, i = 1)$ is an unbiased estimator of $p(o_{t+1}|h_t, a_t, i = 1)$. □

Lemma A.1. *Assuming $|\mathcal{D}_{obs}| \to \infty$, for any $\mathcal{D}_{int}$ the recovered causal model is bounded as follows:*

$$\prod_{t=0}^{T-1} \hat{q}(o_{t+1}|h_t, a_t, i = 1) \geq \prod_{t=0}^{T-1} p(a_t|h_t, i = 0)p(o_{t+1}|h_t, a_t, i = 0), \textit{ and}$$
$$\prod_{t=0}^{T-1} \hat{q}(o_{t+1}|h_t, a_t, i = 1) \leq \prod_{t=0}^{T-1} p(a_t|h_t, i = 0)p(o_{t+1}|h_t, a_t, i = 0) + 1 - \prod_{t=0}^{T-1} p(a_t|h_t, i = 0),$$

$\forall h_{T-1}, a_{T-1}, T \geq 1$ *where* $p(h_{T-1}, a_{T-1}, i = 0) > 0$.

Proof of Lemma A.1. Consider $q(\tau, i) \in \mathcal{Q}$ any distribution that follows our augmented POMDP constraints. Then, for every $T \geq 1$ we have

$$\begin{aligned}\prod_{t=0}^{T-1} q(a_t|h_t, i)q(o_{t+1}|h_t, a_t, i) &= \frac{q(\tau|i)}{q(h_0|i)} \\ &= \sum_{z_{0\to T}}^{\mathcal{Z}^{T+1}} q(z_0|h_0, i) \prod_{t=0}^{T-1} q(a_t, z_{t+1}, o_{t+1}|z_t, h_t, i),\end{aligned}$$

by using $A_t, Z_{t+1}, O_{t+1} \perp\!\!\!\perp Z_{0\to t-1} \mid Z_t, H_t, I$, which can be read via d-separation in the augmented POMDP DAG. Likewise, for every $t \geq 0$ we have

$$\begin{aligned}q(o_{t+1}|h_t, a_t, i = 1) &= \sum_{z_{t+1}}^{\mathcal{Z}} q(z_{t+1}, o_{t+1}|h_t, a_t, i = 1) \\ &= \sum_{z_t}^{\mathcal{Z}} q(z_t|h_t, i = 1) \sum_{z_{t+1}}^{\mathcal{Z}} q(z_{t+1}, o_{t+1}|z_t, h_t, a_t, i = 0),\end{aligned}$$

by using $Z_t \perp\!\!\!\perp A_t \mid H_t, I = 1$ and $Z_{t+1}, O_{t+1} \perp\!\!\!\perp I \mid Z_t, A_t, H_t$. Then for every $t \geq 1$ we can further write

$$q(o_{t+1}|h_t, a_t, i = 1) = \sum_{z_t}^{\mathcal{Z}} \frac{q(z_t, o_t|h_{t-1}, a_{t-1}, i = 1)}{q(o_t|h_{t-1}, a_{t-1}, i = 1)} \sum_{z_{t+1}}^{\mathcal{Z}} q(z_{t+1}, o_{t+1}|z_t, h_t, a_t, i = 0).$$

By recursively decomposing every $q(z_t, o_t|h_{t-1}, a_{t-1}, i = 1)$ until $t = 0$, and finally by using $Z_0 \perp\!\!\!\perp I \mid H_0$, we obtain that for any $T \geq 1$

$$\prod_{t=0}^{T-1} q(o_{t+1}|h_t, a_t, i = 1) = \sum_{z_{0\to T}}^{\mathcal{Z}^{T+1}} q(z_0|h_0, i = 0) \prod_{t=0}^{T-1} q(z_{t+1}, o_{t+1}|z_t, a_t, h_t, i = 0),$$

which can be re-expressed as

$$\prod_{t=0}^{T-1} q(o_{t+1}|h_t, a_t, i = 1) = \sum_{a'_{0\to T-1}}^{\mathcal{A}^T} \sum_{z_{0\to T}}^{\mathcal{Z}^{T+1}} q(z_0|h_0, i = 0) \prod_{t=0}^{T-1} q(a'_t|z_t, h_t, i = 0)q(z_{t+1}, o_{t+1}|z_t, h_t, a_t, i = 0).$$

By considering the case $a'_{0\to T-1} = a_{0\to T-1}$ in isolation, and by assuming probabilities are positive, we readily obtain our first bound,

$$\prod_{t=0}^{T-1} q(o_{t+1}|h_t, a_t, i=1) \geq \prod_{t=0}^{T-1} q(a_t|h_t, i=0) q(o_{t+1}|h_t, a_t, i=0).$$

In order to obtain our second bound, we further isolate the cases $a'_0 \neq a_0$, then $a'_0 = a_0 \wedge a'_1 \neq a_1$, then $a'_0 = a_0 \wedge a'_1 = a_1 \wedge a'_2 \neq a_2$ and so on until $a'_{0\to T-2} = a_{0\to T-2} \wedge a'_{T-1} \neq a_{T-1}$, which yields

$$\begin{aligned}
&\prod_{t=0}^{T-1} q(o_{t+1}|h_t, a_t, i=1) = \prod_{t=0}^{T-1} q(a_t|h_t, i=0) q(o_{t+1}|h_t, a_t, i=0) \\
&+ \sum_{z_{0\to T}}^{\mathcal{Z}^{T+1}} q(z_0|h_0, i=0)\left(1 - q(a_0|z_0, h_0, i=0)\right) \prod_{t=0}^{T-1} q(z_{t+1}, o_{t+1}|z_t, h_t, a_t, i=0) \\
&+ \sum_{K=0}^{T-2} \sum_{z_{0\to T}}^{\mathcal{Z}^{T+1}} q(z_0|h_0, i=0) \prod_{t=0}^{K} q(a_t, z_{t+1}, o_{t+1}|z_t, h_t, i=0)\left(1 - q(a_K|z_K, h_K, i=0)\right) \\
&\prod_{t=K+1}^{T-1} q(z_{t+1}, o_{t+1}|z_t, h_t, a_t, i=0).
\end{aligned}$$

Then by assuming probabilities are upper bounded by 1, we obtain

$$\begin{aligned}
\prod_{t=0}^{T-1} q(o_{t+1}|h_t, a_t, i=1) &\leq \prod_{t=0}^{T-1} q(a_t|h_t, i=0) q(o_{t+1}|h_t, a_t, i=0) + 1 - q(a_0|h_0, i=0) \\
&+ \sum_{K=0}^{T-2} \prod_{t=0}^{K} q(o_{t+1}|h_t, a_t, i=0) \left(\prod_{t=0}^{K-1} q(a_t|h_t, i=0) - \prod_{t=0}^{K} q(a_t|h_t, i=0) \right) \\
&\leq \prod_{t=0}^{T-1} q(a_t|h_t, i=0) q(o_{t+1}|h_t, a_t, i=0) + 1 - \prod_{t=0}^{T-1} q(a_t|h_t, i=0).
\end{aligned}$$

Finally, with $\hat{q}$ solution of (2) and $|\mathcal{D}_{obs}| \to \infty$ we have that $D_{\mathrm{KL}}(p(\tau|i=0) \| \hat{q}(\tau|i=0)) = 0$, and thus $\hat{q}(a_t|h_t, i=0) = p(a_t|h_t, i=0)$ and $\hat{q}(o_{t+1}|h_t, a_t, i=0) = p(o_{t+1}|h_t, a_t, i=0)$, which shows the desired result. □

Proposition 3.2. *For any $\mathcal{D}_{int}$, the estimator $\hat{q}(o_{t+1}|o_{0\to t}, do(a_{0\to t})$ offers strictly better generalization guarantees when $|\mathcal{D}_{obs}| \to \infty$ than when $|\mathcal{D}_{obs}| = 0$.*

Proof. There exists at least one history-action couple (h_{T-1}, a_{T-1}), $T \geq 1$, that has non-zero probability in the observational regime. This ensures that there exists a value o_T for which $\prod_{t=0}^{T-1} p(a_t|h_t, i=0) p(o_{t+1}|h_t, a_t, i=0)$ is strictly positive, which in turn ensures $\hat{q}(o_{T+1}|h_T, a_T, i=1) > 0$ (Lemma A.1). As a result, the family of models $\{q(o_{t+1}|h_t, a_t, i=1) \mid q \in \mathcal{Q}, q(\tau|i=0) = p(\tau|i=0)\}$ is a strict subset of the unrestricted family $\{q(o_{t+1}|h_t, a_t, i=1) \mid q \in \mathcal{Q}\}$, and thus offers strictly better generalization guarantees. □

A.4. Companion example: the door problem

Consider the door problem described in Figure 20. The mechanism responsible for opening the door works as follows: when the light is red, button A opens the door, when the light is green, button B opens the door. The light is red 60% of the time, and green the rest of the time. I am told nothing about the door's mechanism, except that it depends on both the light color and the button pressed (*Light* → *Door* ← *Button*). Since I am colorblind I cannot use the light color to make decisions, and the question I am interested in is simply, which button is more likely to open the door? In the do-calculus framework, this question translates to

$$\underset{button \in \{A,B\}}{\arg\max}\ p(door{=}open|do(button)).$$

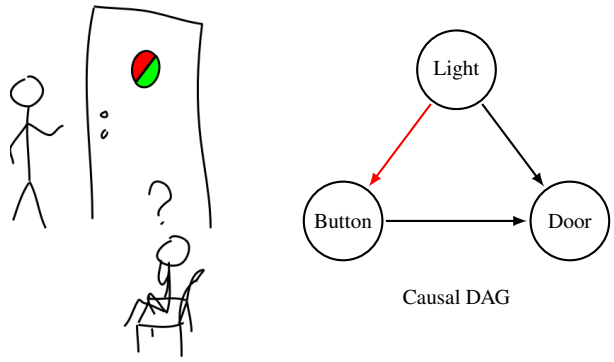


Figure 20. The door problem. You are sitting in a room with a door, a light that can be red or green, and two buttons that will open the door depending on the light color. You can collect data samples in two ways, either from interventions, i.e., you get up and press the buttons (expensive), or from observations, i.e., you watch someone else press the buttons (cheap). A key detail: you're colorblind and can't distinguish red from green. Your goal is to find which button is more likely to open the door.

I thus need to estimate two causal queries: $p(\textit{door=open}|do(\textit{button}=A))$ and $p(\textit{door=open}|do(\textit{button}=B))$.

Interventional setting If I observe my own (or another colorblind person's) interactions with the door, then I know that which button is pressed is unrelated to which color the light is ($\textit{Light} \not\rightarrow \textit{Button}$). Then, I can directly estimate the causal effect of the button on the door,

$$p(\textit{door=open}|do(\textit{button})) = p_{\textit{std}}(\textit{door=open}|\textit{button}).$$

In this regime, regardless of which policy is used to collect $(\textit{button}, \textit{door})$ samples, eventually I realize that button A has more chances of opening the door (60%) than button B (40%), and thus is the optimal action[5].

Observational setting Assume now that I observe another person interacting with the door. I do not know whether that person is colorblind or not ($\textit{Light} \rightarrow \textit{Button}$ is possible). In this regime, without additional knowledge, I cannot recover the causal queries $p(\textit{door=open}|do(\textit{button}))$ from the observed distribution $p(\textit{door}, \textit{button})$. In the do-calculus framework, the queries are said *non identifiable*. For example, with a privileged policy $\pi_{\textit{prv}}(\textit{button=A}|\textit{light=red}) = 0.9$ and $\pi_{\textit{prv}}(\textit{button=A}|\textit{light=green}) = 0.4$ the observed transition probabilities are $p_{\textit{prv}}(\textit{door=open}|\textit{button}=A) = 0.77$ and $p_{\textit{prv}}(\textit{door=open}|\textit{button}=B) = 0.8$, which are not causal due to confounding.

Augmented setting First, consider that I can observe many $(\textit{button}, \textit{door})$ interactions from a non-colorblind person, who's policy is $\pi(\textit{button=A}|\textit{light=red}) = 0.9$ and $\pi(\textit{button=A}|\textit{light=green}) = 0.4$. Then I can safely infer from Lemma A.1 that $p(\textit{door=open}|do(\textit{button=A})) \in [0.54, 0.84]$ and $p(\textit{door=open}|do(\textit{button=B})) \in [0.24, 0.94]$. Then, I get a chance to interact with the door, and I decide to press A 10 times and B 10 times. I am unlucky, and my interventional study results in the following frequencies: $q(\textit{door=open}|do(\textit{button=A})) = 0.5$ and $q(\textit{door=open}|do(\textit{button=B})) = 0.5$. This does not coincide with my (reliable) observational study, and therefore I adjust $q(\textit{door=open}|do(\textit{button=A}))$ to its lower bound 0.54. I now believe that pressing A is more likely to be my optimal strategy.

[5] One assumption though is strict positivity, $\pi(\textit{button}) > 0 \; \forall \textit{button}$, which ensures that both buttons are pressed.