Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models: Supplementary Material

Michael Oberst [1] David Sontag [1]

A. Omitted Proofs

Lemma 1 (Counterfactual Decomposition of Expected Reward)**.** *Let trajectories τ be drawn from $p^{\pi_{obs}}$. Let $\tau_{\hat{\pi}}$ be a counterfactual trajectory, drawn from our posterior distribution over the exogenous U variables under the new policy $\hat{\pi}$. Note that under the SCM, $\tau_{\hat{\pi}}$ is a deterministic function of the exogenous U variables, so we can write $\tau_{\hat{\pi}}(u)$ to be explicit:*

$$\begin{aligned}
&\mathbb{E}_{\hat{\pi}}[R(\tau)|O_1 = o_1] \\
&= \int_\tau p^{\pi_{obs}}(\tau|O_1 = o_1)\mathbb{E}_{u\sim p^{\pi_{obs}}(u|\tau)}[R(\tau_{\hat{\pi}}(u))]d\tau
\end{aligned}$$

Proof. This proof is similar to the proof of Lemma 1 from (Buesing et al., 2019), but is spelled out here for the sake of clarity. Recall that the distribution of noise variables U is the same for every intervention / policy. Thus, $p^{\pi_{obs}}(u) = p^{\hat{\pi}}(u) = p(u)$. We will write p' and $\hat{p}$ for $p^{\pi_{obs}}$ and $p^{\hat{\pi}}$ respectively to simplify notation.

Furthermore, recall that all variables are a deterministic function of their parents in the causal DAG implied by the SCM. Most importantly, this means that the trajectory τ is a deterministic function of the policy π and the exogenous variables U. With that in mind, let $\tau_{\hat{\pi}}(u)$ indicate the trajectory τ as a deterministic function of $\hat{\pi}$ and u. We will occasionally use indicator functions to indicate whether or not a deterministic value is compatible with the variables that determine it, e.g., $\mathbb{1}\left[\tau|u,\pi\right]$ is equivalent to the indicator for $\mathbb{1}\left[\tau = \tau_\pi(u)\right]$. Note that the first observation is independent of the policy, and is just a function of the exogenous U, so we will write $\mathbb{1}\left[o_1|u\right]$ in that case. For simplicity, we will remove the conditioning on O_1 to start with:

$$\begin{aligned}
&\mathbb{E}_{\hat{p}}[R(\tau)] \\
&= \int R(\tau_{\hat{\pi}}(u)) \cdot \hat{p}(u)du && (1)\\
&= \int R(\tau_{\hat{\pi}}(u)) \cdot p'(u)du && (2)\\
&= \int R(\tau_{\hat{\pi}}(u)) \cdot \left(\int p'(\tau, u)d\tau\right) du && (3)\\
&= \int\int R(\tau_{\hat{\pi}}(u)) \cdot p'(u|\tau) \cdot p'(\tau)dud\tau && (4)\\
&= \mathbb{E}_{\tau\sim p'}\left[\int R(\tau_{\hat{\pi}}(u)) \cdot p'(u|\tau)du\right] && (5)\\
&= \mathbb{E}_{\tau\sim p'}\mathbb{E}_{u\sim p'(u|\tau)}\left[R(\tau_{\hat{\pi}}(u))\right] && (6)\\
&= \int_\tau p^{\pi_{obs}}(\tau)\mathbb{E}_{u\sim p'(u|\tau)}\left[R(\tau_{\hat{\pi}}(u))\right]d\tau && (7)
\end{aligned}$$

In step (1) we are just using the definition of the expectation under $\hat{p}$, along with the notation $\tau_{\hat{\pi}}(u)$ to indicate that the trajectory is a deterministic function of the exogenous u and the policy $\hat{\pi}$. In step (2) we replace $\hat{p}(u)$ with $p'(u)$ because they are equivalent, as noted earlier. In step (3) we expand $p'(u)$ over possible trajectories τ arising from the observed policy. In step (4) we rearrange terms and swap the order of the integral, and in step (5) we rewrite the outer integral as an expectation. In step (6) we further condense notation, and then expand in step (7) to match the notation in the Lemma. If we introduce the conditioning on O_1, we see that it is substantively the same.

$$
\begin{aligned}
&\mathbb{E}_{\hat{p}}[R(\tau)|o_1] \\
&= \int R(\tau_{\hat{\pi}}(u)) \cdot \mathbb{1}\left[o_1|u\right] \cdot \hat{p}(u)du && (8) \\
&= \int R(\tau_{\hat{\pi}}(u)) \cdot \mathbb{1}\left[o_1|u\right] \cdot p'(u)du && (9) \\
&= \int R(\tau_{\hat{\pi}}(u)) \cdot p'(u|o_1)du && (10) \\
&= \int R(\tau_{\hat{\pi}}(u)) \cdot \left(\int p'(\tau, u|o_1)d\tau\right) du && (11) \\
&= \int \int R(\tau_{\hat{\pi}}(u)) \cdot p'(u|\tau) \cdot p'(\tau|o_1)dud\tau && (12) \\
&= \int p'(\tau|o_1) \left[\int R(\tau_{\hat{\pi}}(u)) \cdot p'(u|\tau)du\right] d\tau && (13) \\
&= \int_\tau p'(\tau|o_1)\mathbb{E}_{u\sim p'(u|\tau)}[R(\tau_{\hat{\pi}}(u))]d\tau && (14)
\end{aligned}
$$

The main difference in this case is that is just that we carry the indicator into the prior on U at step (10), which we can do because O_1 does not depend on the policy that is applied. Note that Equation (14) matches the statement of the Lemma. □

Corollary 1 (Counterfactual Decomposition of δ_o).

$$
\begin{aligned}
&\delta_o := \mathbb{E}_{\hat{\pi}}[R(\tau)|O_1 = o_1] - \mathbb{E}_{obs}[R(\tau)|O_1 = o_1] \\
&= \int_\tau p^{\pi_{obs}}(\tau|O_1 = o_1)\mathbb{E}_{u\sim p^{\pi_{obs}}(u|\tau)}[R(\tau_{\hat{\pi}}(u)) - R(\tau)]d\tau
\end{aligned}
$$

Proof. By Lemma 1, we have it that

$$
\begin{aligned}
&\delta_o := \mathbb{E}_{\hat{\pi}}[R(\tau)|O_1 = o] - \mathbb{E}_{obs}[R(\tau)|O_1 = o] \\
&= \int_\tau p'(\tau|o_1)\mathbb{E}_{u\sim p'(u|\tau)}[R(\tau_{\hat{\pi}}(u))]d\tau \\
&\quad - \int_\tau p'(\tau|o_1)\mathbb{E}_{u\sim p'(u|\tau)}[R(\tau_{\pi_{obs}}(u))]d\tau \\
&= \int_\tau p^{\pi_{obs}}(\tau|O_1 = o_1)\mathbb{E}_{u\sim p^{\pi_{obs}}(u|\tau)}[R(\tau_{\hat{\pi}}(u)) - R(\tau)]d\tau
\end{aligned}
$$

Note that in the last step, we recognize that $\mathbb{P}_{u\sim p'(u|\tau)}[\tau_{\pi_{obs}}(u) = \tau] = 1$, because the posterior density over u is zero for all u such that $\tau_{\pi_{obs}}(u) \neq \tau$. □

Theorem 1. *Let $Y = f_y(t, u)$ be the SCM for a binary variable Y, where T is also a binary variable. If this SCM satisfies the counterfactual stability property, then it also satisfies the monotonicity property with respect to T.*

Proof. We collect Definitions (4) and (5) here for ease of reference

Monotonicity: An SCM of a binary variable Y is monotonic relative to a binary variable T if and only if it has the following property: $\mathbb{E}[Y|do(T = t)] \geq \mathbb{E}[Y|do(T = t')] \implies f_y(t, u) \geq f_y(t', u), \ \forall u$. We can write equivalently that the following event never occurs, in the case where $\mathbb{E}[Y|do(T = 1)] \geq \mathbb{E}[Y|do(T = 0)]$: $Y_{do(T=1)} = 0 \wedge Y_{do(T=0)} = 1$. Conversely for $\mathbb{E}[Y|do(T = 1)] \leq \mathbb{E}[Y|do(T = 0)]$, the following event never occurs: $Y_{do(T=1)} = 1 \wedge Y_{do(T=0)} = 0$

Counterfactual Stability: An SCM of a categorical variable Y satisfies *counterfactual stability* if it has the following property: If we observe $Y_I = i$, then for all $j \neq i$, the condition $\frac{p'_i}{p_i} \geq \frac{p'_j}{p_j}$ implies that $P^{\mathcal{M}|Y_I=i;I'}(Y = j) = 0$. That is, if we observed $Y = i$ under intervention I, then the counterfactual outcome under I' cannot be equal to $Y = j$ unless the multiplicative change in p_i is less than the multiplicative change in p_j

To simplify notation further, let $p^{t=1} := P(Y = 1|do(T = 1))$, $p^{t=0} := P(Y = 1|do(T = 0))$, and let $Y_t := Y_{do(T=t)}$. Without loss of generality, assume that $p^{t=1} \geq p^{t=0}$.

To show that counterfactual stability implies monotonicity, we want to show that the probability of the event $(Y_1 = 0 \wedge Y_0 = 1)$ is equal to zero. We will do so by proving both cases: First that $P^{\mathcal{M}|Y_0=1;do(T=1)}(Y = 0) = 0$ and second that $P^{\mathcal{M}|Y_1=0;do(T=0)}(Y = 1) = 0$. We can start with the assumption that $p^{t=1} \geq p^{t=0}$ and write:

$$
\begin{aligned}
&p^{t=1} \geq p^{t=0} \\
\implies &p^{t=1}(1 - p^{t=0}) \geq p^{t=0}(1 - p^{t=1}) \\
\implies &\frac{p^{t=1}}{p^{t=0}} \geq \frac{(1 - p^{t=1})}{(1 - p^{t=0})}
\end{aligned}
$$

Using the counterfactual stability condition, the last inequality implies that if we observe $Y_0 = 1$, then the counterfactual probability of $Y_1 = 0$ is equal to $P^{\mathcal{M}|Y_0=1;do(T=1)}(Y = 0) = 0$, as desired. For the second case, where we observe $Y_1 = 0$, we can simply manipulate the inequality to see that

$$
\frac{(1 - p^{t=0})}{(1 - p^{t=1})} \geq \frac{p^{t=0}}{p^{t=1}}
$$

Which yields the conclusion that $P^{\mathcal{M}|Y_1=0;do(T=0)}(Y = 1) = 0$, as desired, completing the proof. □

Theorem 2. *The Gumbel-Max SCM satisfies the counterfactual stability condition.*

Proof. Recall that we write the shorthand $p_i := P^{\mathcal{M};I}(Y = i)$, and $p'_i := P^{\mathcal{M};I'}(Y = i)$. Suppose that Y is generated from a Gumbel-Max SCM $\mathcal{M}$ under intervention I, and we observe that $Y_I = i$. The Gumbel-Max SCM implies that almost surely:

$$\log p_i + g^{(i)} > \log p_j + g^{(j)} \;\; \forall j \neq i \tag{15}$$

To demonstrate that the Gumbel-Max SCM satisfies the counterfactual stability condition, we need to demonstrate that $\frac{p'_i}{p_i} \geq \frac{p'_j}{p_j} \implies P^{\mathcal{M}|Y_I=i;I'}(Y = j) = 0$ for all $j \neq i$.

We will proceed by proving the contrapositive, that for all $j \neq i$, $P^{\mathcal{M}|Y_I=i;I'}(Y = j) \neq 0 \implies \frac{p'_i}{p_i} < \frac{p'_j}{p_j}$.

Fix some index $j \neq i$. The condition $P^{\mathcal{M}|Y_I=i;I'}(Y = j) \neq 0$ implies that there exist values $g^{(i)}, g^{(j)}$ such that

$$\log p'_i + g^{(i)} < \log p'_j + g^{(j)} \tag{16}$$

Because the Gumbel variables $g^{(i)}, g^{(j)}$ are fixed across interventions, this implies there exist values for these variables which satisfy both inequalities (15) and (16). Thus, we proceed by subtracting inequality (15) from inequality (16), maintaining the direction of the inequality and cancelling out the Gumbel terms. The rest is straightforward manipulation using the monotonicity of the logarithm.

$$\begin{aligned} \log p'_i - \log p_i &< \log p'_j - \log p_j \\ \log(p'_i/p_i) &< \log(p'_j/p_j) \\ (p'_i/p_i) &< (p'_j/p_j) \end{aligned}$$

This demonstrates that $P^{\mathcal{M}|Y_I=i;I'}(Y = j) \neq 0 \implies (p'_i/p_i) < (p'_j/p_j)$ as desired, and taking the contrapositive completes the proof. □

B. Non-Identifiability Example

Figure 1 gives a visual depiction of the unidentifiability example given in Section 3.1.

C. Experimental Details

C.1. Sepsis Simulator

All the code required to reproduce our experiments (including the figures in this appendix) is available online at https://www.github.com/clinicalml/gumbel-max-scm, and we refer to that for more in-depth information about our simulator setup.

C.2. Impact of hidden state

In the experiments given in the paper, we hide the glucose and diabetes state from the model of dynamics used for the RL policy. In this section we explore the impact of that choice on the off-policy evaluations used in the paper, as well as on the quality of the RL policy.

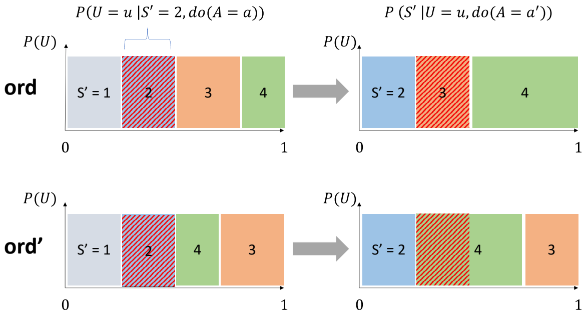
$P(U = u \,|S' = 2, do(A = a))$
$P\,(S' \,|U = u, do(A = a'))$
$P(U)$
ord
S' = 1
2
3
4
S' = 2
3
4
0
1
ord'
S' = 1
2
4
3
S' = 2
4
3

Figure 1. Example of non-identifiability of categorical counterfactual outcomes. The orderings **ord** and **ord'** both define a causal mechanism $S' = f(S, A, U)$ with $U \sim Unif(0, 1)$ that replicates the interventional probability distribution $P(S'|S, A)$. On the left-hand side, the red shading represents the posterior $P(U|S' = 2, A = a, S = s)$, and when this posterior is used on the right-hand side to sample from the counterfactual distribution, these ordering produce different counterfactual outcomes ($S' = 3$ in the case of **ord** and $S' = 4$ in the case of **ord'**)

To demonstrate, in Figure 2, we replicate Figure 3 from the main paper, but with some important differences. First, instead of using 100 bootstrapped samples of the original 1000 trajectories, we instead repeat the entire process 100 times, with an independent set of trajectories drawn from the simulator in each case. These uncertainty intervals are wider, reflecting the variation which is not captured by bootstrapping alone. Second, we compare the use of a WIS estimator used on the training data (i.e., the original 1000 episodes used to learn the model of dynamics), with a WIS estimator used on a held-out set of 1000 independent episodes. While the example given in the paper is meant to conceptually capture what might happen in a single analysis (where only a single set of trajectories is available), Figure 2 demonstrates the variability across analyses, including those with access to a large held-out set of trajectories.

Towards understanding the impact of hiding variables from the RL policy, we performed the same experiment again, but giving the RL policy access to the entire state space. The results are shown in Figure 3, and the results from both figures are shown in Table 1

There are several reasons why weighted importance sampling, and other off-policy evaluation methods, could fail to capture the true performance of a target policy. These include issues like confounding and small sample sizes, as discussed in (Gottesman et al., 2019). In this particular synthetic example, all of the following factors may play a role in the above results, but it is difficult to say conclusively how

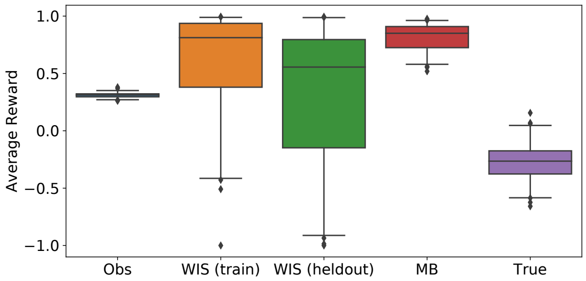

Figure 2. Boxplots show the median and intervals which capture 95% of the 100 evaluations, each time with a newly simulated set of 1000 episodes used for training and 1000 episodes used for the held-out WIS estimator; WIS (train) is used on the training episodes, as in the main paper, and WIS (held-out) is performed on the held-out set of 1000 episodes

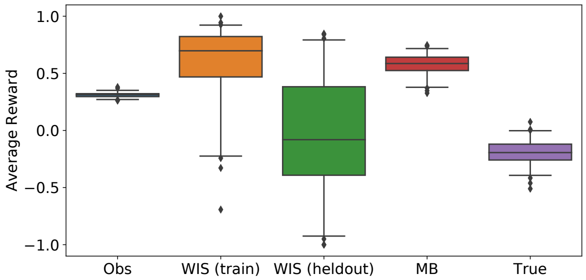

Figure 3. Same setup as Figure 2, but allowing the model of dynamics used by the MDP to see the full state

strong each factor is, and how they interact to produce the results: (i) Confounding due to unobserved states, (ii) sample complexity of learning the MDP, which is more pronounced when all state information is observed (144 states vs 1440 states), and (iii) small sample sizes in both the training and held-out datasets.

With that in mind, we believe that building a more comprehensive simulated environment, in which these various factors can be disentangled more precisely, would be a valuable direction for future work. In addition, we believe such an environment would be useful for evaluation of a variety of off-policy techniques beyond the limited set discussed in the paper e.g., more recently developed methods such as (Thomas & Brunskill, 2016; Liu et al., 2018).

Table 1. Performance given as Mean (95% CI) from Figures 2- 3

	Hidden state	No hidden state
Observed Reward	0.31 (0.27, 0.35)	0.31 (0.27, 0.35)
WIS (train)	0.61 (-0.42, 0.99)	0.58 (-0.23, 0.92)
WIS (heldout)	0.32 (-0.92, 0.99)	-0.04 (-0.94, 0.80)
MB Estimate	0.81 (0.57, 0.96)	0.58 (0.37, 0.73)
True RL Reward	-0.27 (-0.59, 0.05)	-0.19 (-0.41, 0.00)

References

Buesing, L., Weber, T., Zwols, Y., Heess, N., Racaniere, S., Guez, A., and Lespiau, J.-B. Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJG0voC9YQ.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18, 2019. ISSN 1078-8956. doi: 10.1038/s41591-018-0310-5. URL http://www.nature.com/articles/s41591-018-0310-5.

Liu, Y., Gottesman, O., Raghu, A., Komorowski, M., Faisal, A. A., Doshi-Velez, F., and Brunskill, E. Representation Balancing MDPs for Off-policy Policy Evaluation. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 2644–2653. Curran Associates, Inc., 2018.

Thomas, P. S. and Brunskill, E. Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning. In *33rd International Confernece on Machine Learning (ICML)*, volume 48, 2016. URL http://arxiv.org/abs/1604.00923.