SPATIAL-TEMPORAL CAUSAL EFFECTS EVALUATION: A MULTI-AGENT REINFORCEMENT LEARNING FRAMEWORK

A PREPRINT

Suppose the observed data set are summarized by $\{S_{i,t}, A_{i,t}, R_{i,t}\}_{1\le i\le N, 0\le t<T}$ where $S_{i,t}$ denotes the state variable for Region i observed at time t, $A_{i,t}$ denotes the action that Region i takes at time t and $R_{i,t}$ the corresponding immediate reward. For the central agent, the joint state and action space is given by product space of each individual state and action space. Let $\boldsymbol{S}_t = (S_{1,t}, S_{2,t}, \cdots, S_{N,t})^\top$, $\boldsymbol{A}_t = (A_{1,t}, A_{2,t}, \cdots, A_{N,t})^\top$ and $\boldsymbol{R}_t = (R_{1,t}, R_{2,t}, \cdots, R_{N,t})^\top$ be the joint state, action and reward vectors at time t, respectively. The global reward at time t is defined as a simple average of individual rewards, i.e.,

$$\bar{R}_t = \frac{1}{N}\sum_{i=1}^{N} R_{i,t}.$$

We assume the system is Markovian and focus on binary treatments so that $A_{i,t} \in \{0,1\}$ for any i,t. Policies that we would like to evaluate are nondynamic in time. They can be summarized into N-dimensional vectors, consisting of N binary elements. Let $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_N)^\top$ denote such a policy.

The major challenge of policy evaluation under this setup lies in handling high-dimensional state-action space and characterizing the spatial-temporal dependence of the spillover effects. To address the first challenge, we impose the following conditions on the transition and reward function:

$$\begin{aligned}
\mathbb{P}(S_{i,t+1} \in \mathcal{S} | \boldsymbol{S}_t, \boldsymbol{A}_t, \boldsymbol{R}_{t-1}, \cdots, \boldsymbol{R}_0, \boldsymbol{S}_0, \boldsymbol{A}_0) &= \mathcal{P}_i(\mathcal{S}; S_{i,t}, A_{i,t}, T_s(\{S_{j,t}\}_{j\in\mathcal{N}(i)}), T_a(\{A_{j,t}\}_{j\in\mathcal{N}(i)})), \\
\mathbb{P}(R_{i,t} \le z | \boldsymbol{S}_t, \boldsymbol{A}_t, \boldsymbol{R}_{t-1}, \cdots, \boldsymbol{R}_0, \boldsymbol{S}_0, \boldsymbol{A}_0) &= r_i(z; S_{i,t}, A_{i,t}, T_s(\{S_{j,t}\}_{j\in\mathcal{N}(i)}), T_a(\{A_{j,t}\}_{j\in\mathcal{N}(i)})),
\end{aligned}$$

for any $z \in \mathbb{R}$, $\mathcal{S}_i$ that is a subset of the state space almost surely, where $\mathcal{N}(i)$ denotes the neighborhood of Region i and T_s, T_a are some known functions. Under this model setup, the temporal dependence is characterized by the system transition functions. The spatial dependence is characterized by the function T.

These conditions are reasonable in our applications. Treatment at one region can affect another only through its impact on the distribution of drivers. Within each time unit, each driver can travel at most from one region to one of its neighbors. Hence, the spillover effect might not exist for non-neighborhood regions.

Suppose now we want to evaluate the value of a given policy $\boldsymbol{\pi}$. We define value as the average cumulative reward that we will obtain

$$V_{\boldsymbol{\pi}} = \lim_t \frac{1}{Nt} \sum_{i=1}^{N} \sum_{j=0}^{t-1} \mathbb{E}^{\boldsymbol{\pi}} R_{i,j}.$$

Similarly define the region-specific average cumulative reward as

$$V_{i,\boldsymbol{\pi}} = \lim_t \frac{1}{t} \sum_{j=0}^{t-1} \mathbb{E}^{\boldsymbol{\pi}} R_{i,j}.$$

In the following, we propose an inverse propensity score weighted estimator for $V(\boldsymbol{\pi})$. Suppose the system is stationary in our observed data. Let $p_{b,i}$ and $p_{\pi,i}$ be the stationary density functions of $(S_{i,t}, A_{i,t}, T_{s,i,t}, T_{a,i,t})$ where $T_{s,i,t} = T_s(\{S_{j,t}\}_{j \in \mathcal{N}(i)})$ and $T_{a,i,t} = T_a(\{A_{j,t}\}_{j \in \mathcal{N}(i)})$ under the behavior and target policy, respectively. Denote by $\omega_i = p_{\pi,i}/p_{b,i}$ be its ratio, we have

$$V_{\boldsymbol{\pi}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \mathbb{E} \omega_i R_{i,t}.$$

Let $\widehat{\omega}_i$ denote some consistent estimator of ω_i, an inverse propensity-score weighted estimator is given by

$$\widehat{V}_{\boldsymbol{\pi}}^{\text{IS}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \widehat{\omega}_i R_{i,t}.$$

To derive an augmented version of $\widehat{V}_{\boldsymbol{\pi}}^{\text{IS}}$, we introduce region-specific Q-function below,

$$Q_i^{\pi}(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}^{\boldsymbol{\pi}} \left\{ \sum_{t=0}^{+\infty} (R_{i,t} - V_{i,\boldsymbol{\pi}}) \middle| \boldsymbol{S}_0 = \boldsymbol{s}, \boldsymbol{A}_0 = \boldsymbol{a} \right\}.$$

The Q-function satisfies the following Poisson equation:

$$\mathbb{E}^{\boldsymbol{\pi}} \{ R_{i,j} + Q_i^{\pi}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_{t+1}) - V_{\pi} - Q_i^{\boldsymbol{\pi}}(\boldsymbol{S}_t, \boldsymbol{A}_t) | \boldsymbol{S}_t, \boldsymbol{A}_t \} = 0.$$

The above Bellman equation can be used to estimate the Q-function from observed data. However, as commented before, it might suffer from the curse of dimensionality. Borrowing the idea from the mean field MARL (Yang et al., 2018), we propose to approximate $Q_i^{\pi}(\boldsymbol{S}_t, \boldsymbol{A}_t)$ by

$Q_i^\pi(S_{i,t}, A_{i,t}, T_{s,i,t}, T_{a,i,t})$. Let $\widehat{Q}_i^\pi$ denote the resulting estimator. Our test statistic is given by

$$\begin{aligned}\widehat{V}_{\boldsymbol{\pi}}^{\text{DR}} &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\widehat{\omega}_i\{R_{i,t} + \widehat{Q}_i^\pi(S_{i,t+1}, \pi(S_{i,t+1}), T_{s,i,t+1}, T_{a,i,t+1}(\pi)) - \widehat{Q}_i^\pi(S_{i,t}, A_{i,t}, T_{s,i,t}, T_{a,i,t}) - \widehat{V}_{i,\boldsymbol{\pi}}\} \\ &+ \frac{1}{N}\sum_{i=1}^{N}\widehat{V}_{i,\boldsymbol{\pi}},\end{aligned}$$

where $\widehat{V}_{i,\boldsymbol{\pi}}$ denotes any consistent estimator for $V_{i,\pi}$, $T_{a,i,t}(\pi) = T(\{\pi_j(S_{j,t})\}_{j\in\mathcal{N}(i)})$. In the following, we discuss some details on estimating the ratio and the Q-function.

Estimation of the weight: We impose the following assumptions in order to correctly estimate the weight function:

$$T_{s,i,t+1}, S_{i,t+1} \perp\!\!\!\perp \boldsymbol{S}_i, \boldsymbol{A}_i | S_{i,t}, A_{i,t}, T_{s,i,t}, T_{a,i,t}. \tag{1}$$

Let $p_{s,\pi,i}$ denote the stationary distribution of $(T_{s,i,t}, S_{i,t})$ under π and $p_{a,\pi,i|s}$ the marginal distribution of $T_{a,i,t}$ and $A_{i,t}$ conditional on $(T_{s,i,t}, S_{i,t})$. In a randomized study, the ratio $p_{a,\pi,i|s}/p_{a,b,i|s}$ is known. It suffices to estimate $\omega_{s,i} = p_{s,\pi,i}/p_{s,b,i}$. Under the assumptions in (1), we have

$$\mathbb{E}\left\{\omega_{s,i}(T_{s,i,t}, S_{i,t})\frac{p_{a,\pi,i|s}(A_{i,t}, T_{a,i,t}|S_{i,t}, T_{s,i,t})}{p_{a,b,i|s}(A_{i,t}, T_{a,i,t}|S_{i,t}, T_{s,i,t})}\middle| T_{s,i,t+1}, S_{i,t+1}\right\} = \omega_{s,i}(T_{s,i,t+1}, S_{i,t+1}).$$

As a result, for any function f, we obtain

$$\mathbb{E}\left\{\omega_{s,i}(T_{s,i,t}, S_{i,t})\frac{p_{a,\pi,i|s}(A_{i,t}, T_{a,i,t}|S_{i,t}, T_{s,i,t})}{p_{a,b,i|s}(A_{i,t}, T_{a,i,t}|S_{i,t}, T_{s,i,t})} - \omega_{s,i}(T_{s,i,t+1}, S_{i,t+1})\right\} f(T_{s,i,t+1}, S_{i,t+1}) = 0.$$

It suffices to estimate $\omega_{s,i}$ that satisfies

$$\sup_f\left|\mathbb{E}\left\{\omega_{s,i}(T_{s,i,t}, S_{i,t})\frac{p_{a,\pi,i|s}(A_{i,t}, T_{a,i,t}|S_{i,t}, T_{s,i,t})}{p_{a,b,i|s}(A_{i,t}, T_{a,i,t}|S_{i,t}, T_{s,i,t})} - \omega_{s,i}(T_{s,i,t+1}, S_{i,t+1})\right\} f(T_{s,i,t+1}, S_{i,t+1})\right| = 0.$$

When f belongs to the class of RKHS, for a given kernel k, it suffices to solve $\omega_{s,i}$ that satisfies

$$\mathbb{E}\Delta(\omega, O_{i,t})\Delta(\omega, O_{i,t'})k(S_{i,t+1}, T_{s,i,t+1}, S_{i,t'+1}, T_{s,i,t'+1}),$$

where

$$\Delta(\omega, O_{i,t}) = \omega_{s,i}(T_{s,i,t}, S_{i,t})\frac{p_{a,\pi,i|s}(A_{i,t}, T_{a,i,t}|S_{i,t}, T_{s,i,t})}{p_{a,b,i|s}(A_{i,t}, T_{a,i,t}|S_{i,t}, T_{s,i,t})} - \omega_{s,i}(T_{s,i,t+1}, S_{i,t+1}).$$

The optimization can be similarly implemented as the algorithm mentioned in Liu (2018).

Estimation of the Q-function: We impose the following assumptions in order to correctly estimate the Q-function:

$$Q_i^\pi(\boldsymbol{S}_i, \boldsymbol{A}_i) = Q_i^\pi(S_{i,t}, A_{i,t}, T_{s,i,t}, T_{a,i,t}).$$

Based on this approximation, we can estimate the Q-function using the method described in Liao et al. (2019). Specifically, let $T_{i,t} = (T_{s,i,t}, T_{a,i,t})$ and $T_{i,t}(\pi) = (T_{s,i,t}, T_{a,i,t}(\pi))$, we define

$$\begin{aligned}\hat{g}_i(\cdot,\cdot,\cdot;\eta,Q) = \arg\min_{g\in\mathcal{G}} \frac{1}{T}\sum_{t=0}^{T-1}\{R_t + Q_i^\pi(S_{i,t+1},\pi(S_{i,t+1}),T_{i,t+1}(\pi)) - \eta - Q_i^\pi(S_{i,t},A_{i,t},T_{i,t})& \\ -g(S_{i,t},A_{i,t},T_{i,t})\}^2 + \mu J_2^2(g),& \\ (\hat{\eta}_i,\hat{Q}_i) = \arg\min_{(\eta,Q)\in\mathbb{R}\times\mathcal{Q}} \frac{1}{T}\sum_{t=0}^{T-1}\hat{g}_i^2(S_{i,t},A_{i,t},T_{i,t};\eta,Q) + \lambda J_1^2(Q),&\end{aligned}$$

where J_1 and J_2 denote some penalty functions. In the following, we derive the close-form of the estimator $\hat{Q}_i$ when we use RKHS to model Q and g. The derivation is based on the results in the paper Farahmand et al. (2016).

Define vectors $Z_{i,t} = (S_{i,t}, A_{i,t}, T_{i,t})^\top$ and $Z_{i,t}^* = (S_{i,t+1}, \pi(S_{i,t+1}), T_{i,t+1}(\pi))^\top$. Let K_g and K_Q denote the reproducing kernels used to model g and Q, respectively. In practice, we can use gaussian RBF kernels to model these two functions. For a given Q and η, the optimizer of $\hat{g}_i$ can be represented by $\sum_{t=0}^{T-1}\hat{\beta}_{i,t}K_g(Z_{i,t},\cdot)$. As such, we obtain

$$\begin{aligned}\hat{\beta}_i &= \arg\min_{\beta}\frac{1}{T}\sum_{t=0}^{T-1}\{R_t + Q_i^\pi(Z_{i,t}^*) - \eta - Q_i^\pi(Z_{i,t}) - \sum_{j=0}^{T-1}\beta_j K_g(Z_{i,j},Z_{i,t})\}^2 + \mu\beta^\top\boldsymbol{K}_g\beta \\ &= \frac{1}{T}\beta^\top(\boldsymbol{K}_g\boldsymbol{K}_g^\top + T\mu\boldsymbol{K}_g)\beta - \frac{2}{T}\beta^\top\boldsymbol{K}_g(\boldsymbol{R} + \boldsymbol{Q}_i^{\pi*} - \boldsymbol{Q}_i^\pi - \eta\boldsymbol{1}) + \text{some term},\end{aligned}$$

where $\boldsymbol{K}_g = \{K_g(Z_{i,j_1}, Z_{i,j_2})\}_{j_1,j_2}$ and $\boldsymbol{R}$, $\boldsymbol{Q}_i^{\pi*}$ and $\boldsymbol{Q}_i$ the column vectors formed by elements in R_t, $Q_i^\pi(Z_{i,t}^*)$ and $Q_i^\pi(Z_{i,t})$, respectively. Notice that $\boldsymbol{K}_g$ is symmetric, we obtain

$$\hat{\beta}_i = (\boldsymbol{K}_g\boldsymbol{K}_g^\top + T\mu\boldsymbol{K}_g)^{-1}\boldsymbol{K}_g(\boldsymbol{R} + \boldsymbol{Q}_i^{\pi*} - \boldsymbol{Q}_i^\pi - \eta\boldsymbol{1}) = (\boldsymbol{K}_g + T\mu\boldsymbol{I})^{-1}(\boldsymbol{R} + \boldsymbol{Q}_i^{\pi*} - \boldsymbol{Q}_i^\pi - \eta\boldsymbol{1}).$$

As a result, for a given Q and η, we have

$$\hat{g}_i(Z_{i,t};\eta,Q) = \hat{\beta}_i^\top\boldsymbol{K}_g e_t,$$

where e_t denotes the column vector with the t-th element equals to one and other elements equal to zero. As such,

$$\frac{1}{T}\sum_{t=0}^{T-1}\hat{g}_i^2(S_{i,t},A_{i,t},T_{i,t};\eta,Q) = \frac{1}{T}\hat{\beta}_i^\top\boldsymbol{K}_g\boldsymbol{K}_g^T\widehat{\beta}_i.$$

Similarly, we can represent Q as $\sum_{t=0}^{2T-1}\widehat{\alpha}_{i,t}K_Q(\widetilde{Z}_{i,t},\cdot)$ where $\widetilde{Z}_{i,t}$ denotes the t-th element in the vector $(Z_{i,0}^\top, Z_{i,1}^\top, \cdots, Z_{i,T-1}^\top, Z_{i,0}^{*\top}, \cdots, Z_{i,T-1}^{*\top})^\top$. Let $\boldsymbol{K}_Q$ denotes the corresponding $2T \times 2T$ matrix, we have

$$Q_i^\pi(Z_{i,t}) = \alpha_i^\top\boldsymbol{K}_Q e_t \ \text{ and } \ Q_i^\pi(Z_{i,t}^*) = \hat{\alpha}_i^\top\boldsymbol{K}_Q e_{t+T}.$$

It follow that

$$\boldsymbol{Q}_i^{\pi*} - \boldsymbol{Q}_i^{\pi} = \underbrace{[-\boldsymbol{I}_T, \boldsymbol{I}_T]}_{\boldsymbol{C}} \boldsymbol{K}_Q \hat{\alpha}_i,$$

noting that $\boldsymbol{K}_Q$ is symmetric. Let $\boldsymbol{E} = \boldsymbol{K}_g^\top (\boldsymbol{K}_g + T\mu \boldsymbol{I})^{-1}$, $\hat{\alpha}_i$ corresponds to the solution of the following optimization problem,

$$\hat{\alpha}_i = \arg\min_{\alpha} (\boldsymbol{R} + \boldsymbol{C}\boldsymbol{K}_Q \alpha_i - \eta \boldsymbol{1})^\top \boldsymbol{E}^\top \boldsymbol{E} (\boldsymbol{R} + \boldsymbol{C}\boldsymbol{K}_Q \alpha_i - \eta \boldsymbol{1}) + T\lambda \alpha_i^\top \boldsymbol{K}_Q \alpha_i.$$

Taking derivatives with respect to α_i and η, we obtain

$$(\hat{\alpha}_i, \hat{\eta})^\top = -([\boldsymbol{C}\boldsymbol{K}_Q, -\boldsymbol{1}]^\top \boldsymbol{E}^\top \boldsymbol{E} [\boldsymbol{C}\boldsymbol{K}_Q, -\boldsymbol{1}] + [T\lambda \boldsymbol{K}_Q, \boldsymbol{0}; \boldsymbol{0}^\top, 0])^{-1} [\boldsymbol{C}\boldsymbol{K}_Q, -\boldsymbol{1}] \boldsymbol{E}^\top \boldsymbol{E} \boldsymbol{R}.$$