Introduction to Dynamic Treatment Regimes

Jae-kwang Kim

Outline

1. Introduction
2. Decision theory for finding optimal decision
3. Q-learning
4. A-learning and G-estimation
5. Connection with reinforcement learning
6. Conclusion and advertisement

Personalized medicine

- Heterogeneity in the treatment effect: patients often respond differently to a particular treatment, both in terms of the primary outcome and side-effects.
- Personalized medicine is a medical paradigm that emphasizes systematic use of individual patient information to optimize that patient's health care.
- Idea: Wish to combine decision theoretic approach with the evidence-based (data-driven) approach
  1. Define the utility function $\mathcal{U}(x, a)$ as a function of covariate x and treatment option a. The loss function (regret function) is defined as

$$L(x, a) = \sup_{a} \mathcal{U}(x, a) - \mathcal{U}(x, a)$$

  2. We need to estimate $\mathcal{U}(x, a)$ from the training sample.

Prescriptive inference

- Evolution of data science
  1. Descriptive inference
  2. Predictive inference
  3. Prescriptive inference
- In the prescriptive inference, we should answer "what if?" questions.
- Potential outcome framework is used to estimate the causal parameter (= the parameter that is associated with the potential outcome variables).
- To estimate the causal parameter, we introduce a model and estimate the parameters in the model. This will be a nuisance parameter.

Potential outcomes framework (Neyman-Rubin)

Data: Observe $(X_i, A_i, Y_i), i = 1, \ldots, n$

- X_i: features of patient i
- A_i: treatment option chosen for patient i
- Y_i: realized outcome variable for patient i

$$Y_i = Y_i^*(1)A_i + Y_i^*(0)(1 - A_i) \tag{1}$$

where $Y_i^*(a)$ is the potential outcome under $A_i = a$. Condition (1) is often called the SUTVA (stable unit treatment value assumption) by Rubin (1980).

Goal: We are interested in finding the optimal decision rule $d^{\text{opt}}(x)$ such that $\mathcal{V}(d) = E\{Y^*(d)\}$ is maximized, where

$$Y^*(d) = Y^*(1)\mathbb{I}\{d(X) = 1\} + Y^*(0)\mathbb{I}\{d(X) = 0\}.$$

and $d : \mathcal{X} \mapsto \mathcal{A} = \{0, 1\}$.

- Decision Theory: The optimal decision d^{opt} can be obtained by

$$d^{\text{opt}}(x) = \arg\max_{a} E\{Y^*(a) \mid X = x\} \tag{2}$$

satisfies

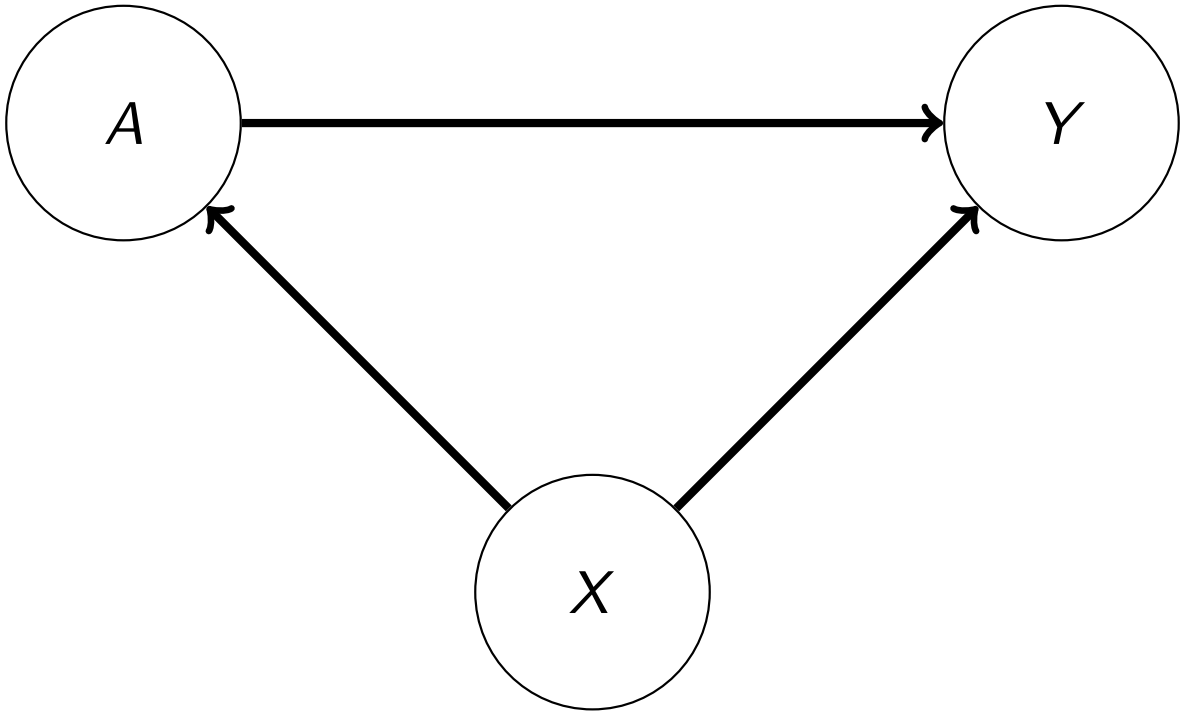
$$\mathcal{V}(d^{\text{opt}}) \geqslant \mathcal{V}(d).$$

- Identification: We wish to estimate the conditional expectation from the sample using

$$E\{Y^*(a) \mid X = x\} = E\{Y \mid X = x, A = a\}. \tag{3}$$

- Result (3) holds under some identification conditions (SUTVA, NUC, positivity).

No unmeasured confounder (NUC) assumption



Multi-stage Decisions and Dynamic Treatment Regimes

- In many cases, decisions are made in multiple stages.
- In the context of multi-stage decisions, a dynamic treatment regime (DTR) is a sequence of decision rules, one per stage of intervention, for adapting a treatment plan to the time-varying state of an individual subject.
- Each decision rule takes a subject's individual characteristics and treatment history observed up to that stage as inputs, and outputs a recommended treatment at that stage; recommendations can include treatment type, dosage, and timing.
- The decision rule d_j at the j-th stage is a mapping from $\mathcal{H}_j$ to $\mathcal{A}_j$, where $\mathcal{H}_j$ is the history space and $\mathcal{A}_j$ is the action space at the j-th decision, for $j = 1, \ldots, T$. We only consider deterministic policy in the sense that the decision mapping is not stochastic.

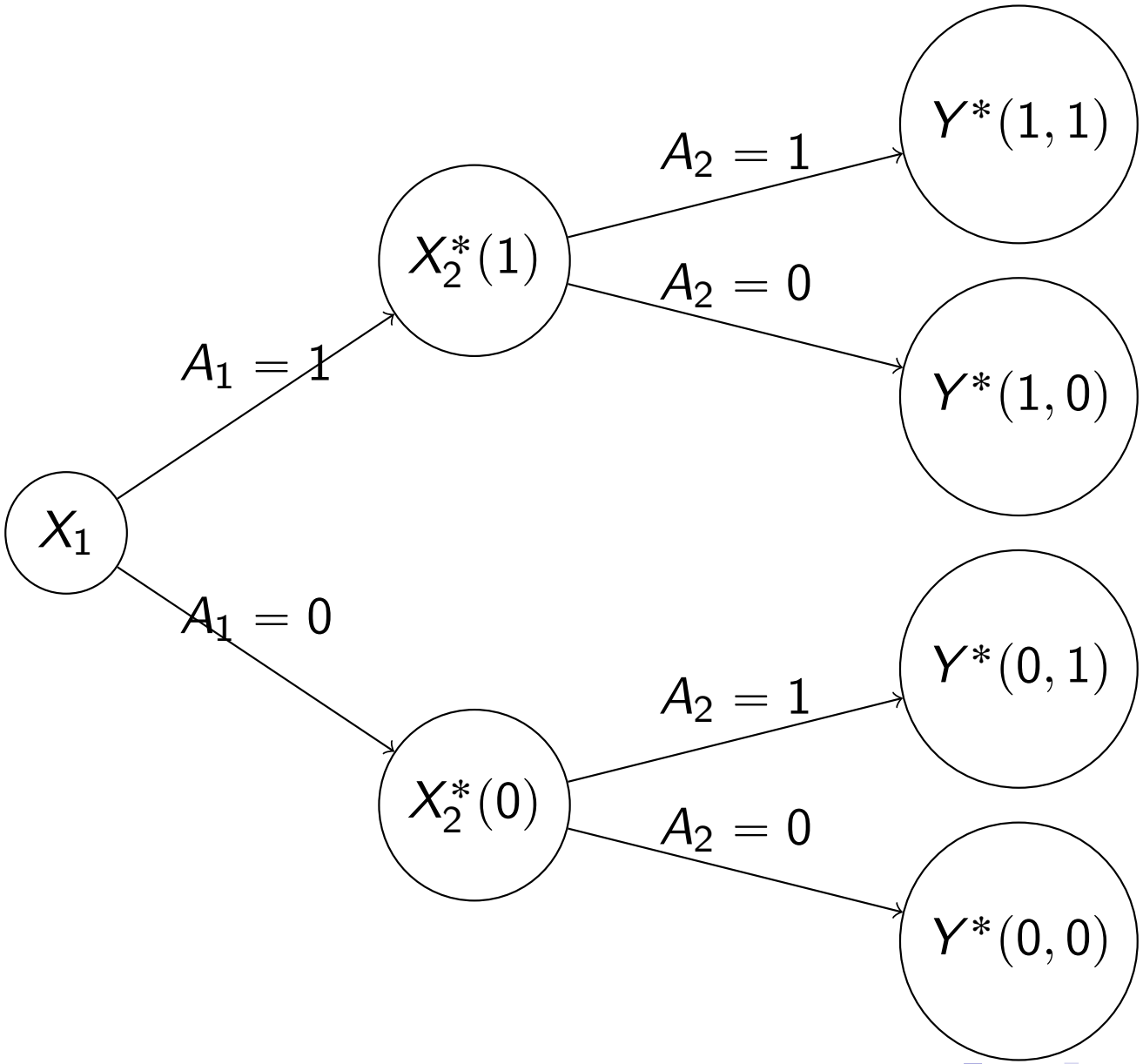
Toy Example: $T = 2$

- Suppose that there are two decision point with action space $\mathcal{A}_1$ and $\mathcal{A}_2$.
- Baseline auxiliary variable X_1
- $X_2^*(a_1)$: potential intermediate outcome (state) under $A_1 = a$.
- $Y^*(a_1, a_2)$: potential final outcome (reward) under $A_1 = a_1$ and $A_2 = a_2$.
- Let $A_1 \in \mathcal{A}_1 = \{0, 1\}$ and $A_2 \in \mathcal{A}_2 = \{0, 1\}$
- Potential outcomes:

$$X_2^*(1), X_2^*(0), Y^*(1,1), Y^*(1,0), Y^*(0,1), Y^*(0,0)$$

- Observed data: $(X_{1i}, X_{2i}, A_{1i}, A_{2i}, Y_i), i = 1, \ldots, n$
- Goal: We are interested in finding an optimal decision rule (d_1, d_2) such that $E\{Y^*(d_1, d_2)\}$ is maximized.

Potential outcomes under two stage decision process



- For $j = 2$, the optimal decision can be obtained similarly to the single decision case:

$$\begin{aligned} d_2^{\text{opt}} &= \arg\max_{a_2} E\{Y^*(a_1, a_2) \mid X_1 = x_1, X_2 = x_2, A_1 = a_1, A_2 = a_2\} \\ &= \arg\max_{a_2} E\{Y \mid X_1 = x_1, X_2 = x_2, A_1 = a_1, A_2 = a_2\} \end{aligned}$$

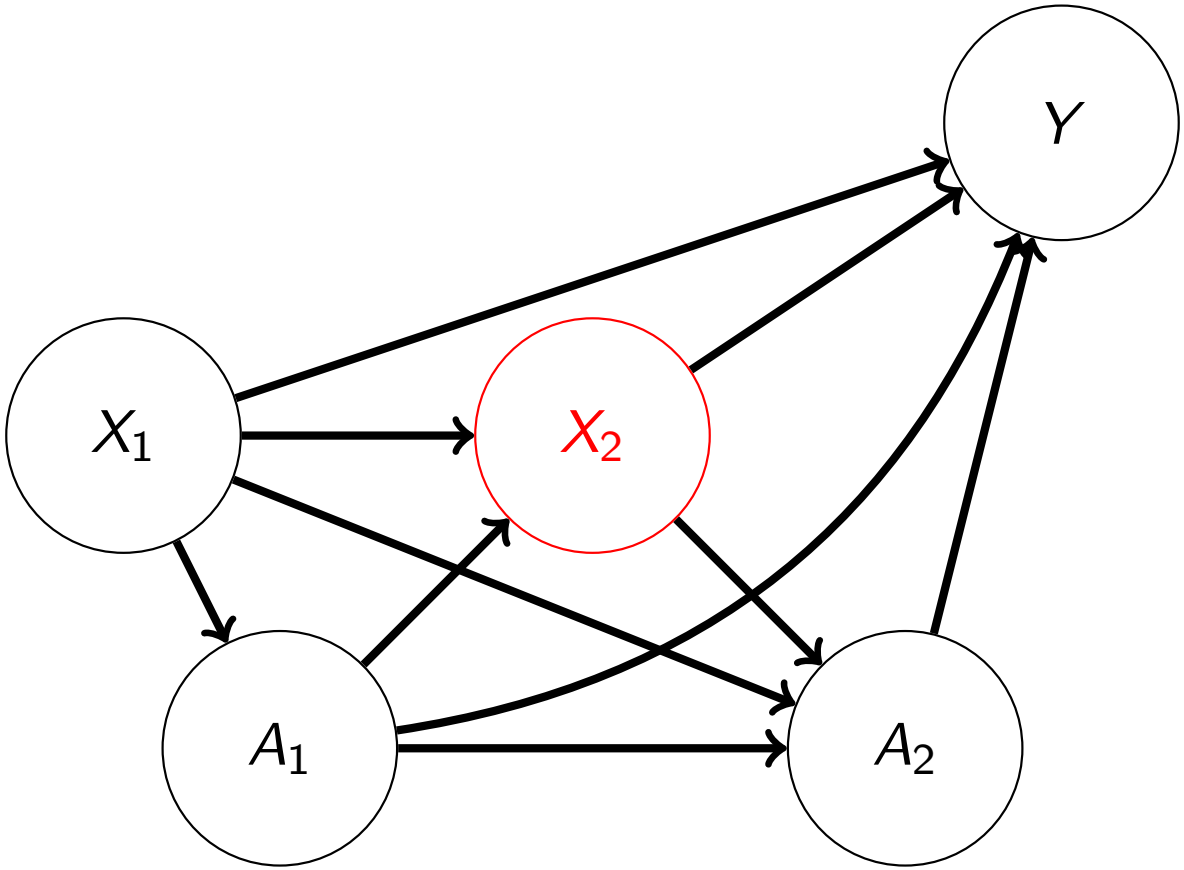
  where the second equality is justified under some identification conditions (SUTVA, sequential ignorability, and positivity).

- However, for $j = 1$, the optimal decision is more tricky:

$$d_1^{\text{opt}} \neq \arg\max_{a_1} E\{Y^*(a_1, a_2) \mid X_1 = x_1, A_1 = a_1, X_2^*(a_1), A_2 = d_2^{\text{opt}}\}$$

- Thus, X_2 is a confounder in the causal path from A_2 to Y, but it is a mediator in the causal path from A_1 to Y.

A two-stage DAG illustrating time-varying confounding and mediation.



Decision Theory ($T = 2$)

- For any $d = (d_1, d_2)$, define the expected reward:

$$\begin{aligned} V_1^d(h_1) &= E_d\{Y^*(A_1, A_2) \mid X_1 = x_1\} \\ V_2^d(h_2) &= E_d\{Y^*(a_1, A_2) \mid X_1 = x_1, X_2 = x_2, A_1 = a_1\} \end{aligned}$$

  This is called the value function wrt a policy d.

- The optimal (stage j) value function is

$$V_j^{\text{opt}}(h_j) = \max_{d \in \mathcal{D}} V_j^d(h_j), \quad j = 1, 2.$$

- Result: The optimal value functions satisfy

$$V_1^{\text{opt}}(h_1) = \max_{a_1} E\left\{ V_2^{\text{opt}}(x_1, X_2^*(a_1), a_1) \mid X_1 = x_1, A_1 = a_1 \right\}. \quad (4)$$

  Equation (4) is called the Bellman equation (Bellman, 1957).

Remark

- Instead of using the value function, it is more convenient to use the Q-function, where "Q" stands for the "quality of action".
- The Q-function for policy d is the expected reward starting from a history h_j at stage j, taking an action a_j, and following the policy d thereafter. Thus,

$$\begin{aligned} Q_1^d(h_1, a_1) &= E\{Y^*(a_1, d_2) \mid X_1 = x_1, A_1 = a_1\} \\ Q_2^d(h_2, a_2) &= E\{Y^*(a_1, a_2) \mid X_1 = x_1, A_1 = a_1, X_2 = x_2, A_2 = a_2\} \\ &= E\{Y \mid X_1 = x_1, A_1 = a_1, X_2 = x_2, A_2 = a_2\} \\ &= Q_2(h_2, a_2) \end{aligned} \tag{5}$$

- By (4), the optimal (stage 1) Q-function satisfies

$$Q_1^{\text{opt}}(h_1, a_1) = E\left\{\max_{a_2} Q_2(H_2, a_2) \mid X_1 = x_1, A_1 = a_1\right\}. \tag{6}$$

Backward induction for finding the optimal DTR

- Find $d_2^{\text{opt}}(h_2)$ first by

$$d_2^{\text{opt}}(h_2) = \arg\max_{a_2} Q_2(h_2, a_2)$$

where $Q_2(h_2, a_2)$ is defined in (5).

- Find $d_1^{\text{opt}}(h_1)$ by

$$d_1^{\text{opt}}(h_2) = \arg\max_{a_2} Q_1^{\text{opt}}(h_1, a_1)$$

where

$$Q_1^{\text{opt}}(h_1, a_1) = E\left\{ V_2^{\text{opt}}(H_2) \mid X_1 = x_1, A_1 = a_1 \right\}$$

and $V_2^{\text{opt}}(H_2) = Q_2(H_2, d_2^{\text{opt}})$ is the pseudo outcome at stage 1 decision.

- Note that $V_2^{\text{opt}}(H_2)$ is a predictor of the unobserved reward Y^* under $A_1 = a_1$ and $A_2 = d_2^{\text{opt}}$

Q-learning for estimating the optimal DTR (T=2)

Stage 2

- Learning: Estimate Q_2 by minimizing

$$\sum_{i=1}^{n} \{Y_i - Q_2(H_{2i}, A_{2i})\}^2 + P_{\lambda_2}(Q_2)$$

  where $P_{\lambda_2}(Q_2)$ is the penalty term on the complexity of Q_2

- Decision: Compute

$$\hat{d}_2^{\text{opt}}(h_2) = \arg\max_{a_2} \hat{Q}_2(h_2, a_2)$$

Stage 1

- Learning: Given $\hat{d}_2^{\text{opt}}$, estimate Q_1^{opt} by minimizing

$$\sum_{i=1}^{n} \left\{ \hat{V}_{2i}^{\text{opt}} - Q_1^{\text{opt}}(H_{1i}, A_{1i}) \right\}^2 + P_{\lambda_1}(Q_1^{\text{opt}})$$

  where $\hat{V}_{2i}^{\text{opt}} = \hat{Q}_2(H_{2i}, d_2^{\text{opt}})$ is the estimated pseudo outcome of unit i under $\hat{d}_2^{\text{opt}}$.

- Decision: Compute

$$\hat{d}_1^{\text{opt}}(h_1) = \arg\max_{a_1} \hat{Q}_1^{\text{opt}}(h_1, a_1)$$

Remark

- Model for Q_2 is for the observed data, but the model for Q_1^{opt} is for the potential outcome under d_2^{opt}.
- Instead of using $\hat{V}_{2i}^{\text{opt}} = \hat{Q}_2(H_{2i}, d_2^{\text{opt}})$, one may use

$$\hat{V}_{2i}^{\text{opt}} = Y_i + \hat{Q}_2(H_{2i}, d_2^{\text{opt}}) - \hat{Q}_2(H_{2i}, A_{2i}). \tag{7}$$

- Sequential ignorability assumption is critical.

A-learning ($T = 2$)

- Idea: Let's use a parametric model for the advantage (regret) functions

$$\mu_1(h_1, a_1) = E\left\{ Y^*(d_1^{\text{opt}}, d_2^{\text{opt}}) - Y^*(a_1, d_2^{\text{opt}}) \mid H_1 = h_1 \right\}$$

$$\mu_2(h_2, a_2) = E\left\{ Y^*(a_1, d_2^{\text{opt}}) - Y^*(a_1, a_2) \mid H_2 = h_2 \right\}$$

- A model specifying the form of $E[Y^*(a_1, a_2) - Y^*(a_1, a_2') \mid \text{covariate}]$ is a structural nested mean models (SNMM).
- A SNMM parameterizes the causal effect that is the difference between the conditional expectation of an outcome in the observed data and the conditional expectation of an outcome under some potential outcome scenario.

- Let $a_j = 0$ be the control group (no treatment) at stage j.
- Use the contrast function

$$C_2(h_2, a_2) = Q_2(h_2, a_2) - Q_2(h_2, 0)$$

to deduce $d_2^{\text{opt}}(h_2)$

$$d_2^{\text{opt}}(h_2) = \arg\max_{a_2} C_2(h_2, a_2)$$

- Also, for a given $d_2^{\text{opt}}(h_2)$,

$$C_1(h_1, a_1) = Q_1^{\text{opt}}(h_1, a_1) - Q_1^{\text{opt}}(h_1, 0)$$

to deduce $d_1^{\text{opt}}(h_1)$

$$d_1^{\text{opt}}(h_1) = \arg\max_{a_1} C_1(h_1, a_1)$$

- We will use a parametric model for contrast function but allow $\nu_j(h_k) = Q_j(h_j, 0)$ fully nonparametric.

Semiparametric model

- WLOG $\mathcal{A}_j = \{0, 1, \ldots, m_j - 1\}$
- Writing $\nu_j(h_j) = Q_j(h_j, 0)$, we can express

$$Q_j(h_j, a_j) = \nu_j(h_j) + \sum_{k=1}^{m_j-1} \mathbb{I}(a_j = k) C_j(h_j, a_j)$$

- Idea: Let's use a semiparametric model for $Q_j(h_j, a_j)$:

$$Q_j(h_j, a_j; \nu_j, \beta_j) = \nu_j(h_j) + \sum_{k=1}^{m_j-1} \mathbb{I}(a_j = k) C_j(h_j, k; \beta_{jk})$$

where ν_j plays the role of infinite-dimensional nuisance parameter and β_{jk} is finite dimensional.

Model parameter estimation (Robins, 2004)

- Estimate parameters in $T = 2$ first: For binary $A_2 \in \{0, 1\}$, use

$$\sum_{i=1}^{n} \lambda_2(H_{2i}) \{A_{2i} - \hat{\pi}_2(H_{2i}, 1)\} \{Y_i - \hat{\nu}_2(H_{2i}) - A_{2i} \cdot C_2(H_{2i}, 1; \beta_2)\} = 0 \quad (8)$$

  as an estimating equation for β_2 (This is essentially an application of Neyman orthogonalization), where $\lambda_2(H_{2i})$ is the same dimension as β_2.
- It enjoys the doubly robust property. May use sample-split ML/debiased estimation method (Chernozhukov et al., 2018).
- Once $\hat{\beta}_2$ is obtained, we can estimate the optimal decision by A-learning:

$$\hat{d}^{\text{opt}}_{A,2}(h_2) = \arg\max_{a_2} C_2(h_2, a_2; \hat{\beta}_2)$$

G-formula: Pseudo outcome construction

- To apply the backward induction (as in the Q-learning), we need to compute pseudo outcomes based on d_2^{opt}.
- Similarly to (7), we can compute pseudo outcomes by

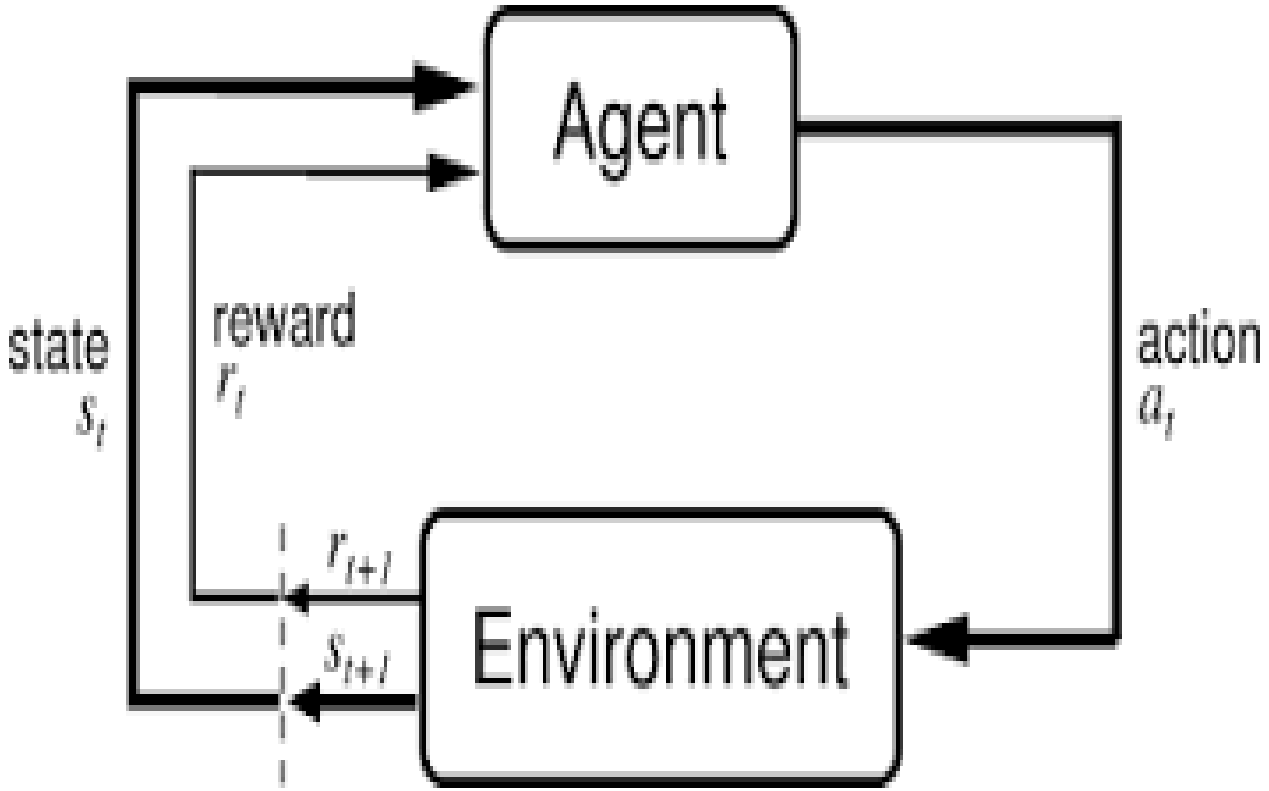
$$\hat{G}_{2i} = Y_i + \underbrace{C_2(H_{2i}, d_2^{\text{opt}}; \hat{\beta}_2) - C_2(H_{2i}, A_{2i}; \hat{\beta}_2)}_{\text{regret function}}$$

We can treat $\hat{G}_{2i}$ as the pseudo outcome (predicted value of Y_i under d_2^{opt}) and apply the same semiparametric estimation method for obtaining $\hat{\beta}_1$. That is, solve

$$\sum_{i=1}^{n} \lambda_1(H_{1i}) \{A_{1i} - \hat{\pi}_1(H_{1i}, 1)\} \left\{ \hat{G}_{2i} - \hat{\nu}_1(H_{1i}) - A_{1i} \cdot C_1(H_{1i}, 1; \beta_1) \right\} = 0$$

for β_1.

Agent-environment interactions



- In RL literature, state S_t is used to represent X_t.
- $(R_0 = 0, S_0) \rightarrow A_0 \rightarrow (R_1, S_1) \rightarrow A_1 \rightarrow (R_2, S_2) \rightarrow \cdots$

Reinforcement Learning (Concepts)

- State, action and reward are the three basic elements of the RL framework. The most traditional (and perhaps the simplest) context in which RL is applied is called a Markov decision process (MDP).
- In an MDP setting, the probability of the environment making a transition to a new state, given the current state and action, does not depend on the distant past of the environment.
- In an MDP, the goal of RL is to learn how to map states to actions so as to maximize the total expected future reward. That is, using the DTR terminology, the goal of RL is to estimate a policy that maximizes the value over a specified class of policies.
- Thus, the optimal decision theory for DTR is quite relevant and the Q-learning and A-learning can be used in RL.

Major distinctions (Chakrabortym and Moodie, 2013)

- In RL, the system dynamics (state transition probabilities) are often known from the physical laws or other subject matter knowledge. This is not the case in the medical setting in DTR.
- In RL, the data are often very cheap and the computational complexity is the major issue. In the medical setting, data are extremely expensive in terms of both time and money.
- The MDP assumption is not realistic in DTR.

Conclusion

- Promising area of research
- Applications in marketing, econometrics, and other social sciences.
- Statistical properties underdeveloped
- I plan to teach these topics in Stat 621.

Advertisement

- Course sequence at ISU: Stat 521 → Stat 523 → Stat 621
- Stat 521 will cover two topics
  1. Basic theory for survey sampling
  2. Basic theory for causal inference
- Stat 523 will cover two topics
  1. Theory and methods for missing data analysis
  2. Introduction to projection technique in Hilbert space
- Stat 621 will cover two topics
  1. Theory and methods for DTR
  2. Statistical reinforcement learning

REFERENCES

Bellman, R. E. (1957), *Dynamic Programming*, Princeton: Princeton University Press.

Chakrabortym, Bibhas and Erica E. M. Moodie (2013), *Statistical Methods for Dynamic Treatment Regimes*, Springer.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and J. M. Robins (2018), 'Double/debiased machine learning for treatment and structural parameters', *The Econometrics Journal* **21**(1), C1–C68.

Robins, J. M. (2004), Optimal structural nested models for optimal sequential decisions, *in* D. Y.Lin and P.Heagerty, eds, 'Proceedings of the Second Seattle Symposium on Biostatistics', Springer, pp. 189–326.

Rubin, D. B. (1980), 'Randomization analysis of experimental data: The fisher randomization test comment', *Journal of the American Statistical Association* **75**, 591–593.