Q- and A- Learning Methods for Estimating Optimal Dynamic Treatment Regimes

TJ Weaver

University of Rochester

October 23, 2024

Overview

What is a Dynamic Treatment Regime?

- In many contexts, the effect of a drug depends both on a patient's covariates (e.g. blood pressure, age, disease stage, etcetera) and the history of previous drugs used in treatment to treat the patient
- A Dynamic Treatment Regime is a sequence of decisions that dictate how a person should be treated given a set of that person's covariates and the history of the person's treatment
- Treatment can be thought of in terms of "rounds" in which the first drug the person is treated with for the condition is the first round, and after some amount of time a decision is made by the physician to either keep the patient on the same treatment or switch to another treatment (often there is only one other treatment), this is the second "round"

Notation

- The rounds are notated $k \in 1, 2, ..., K$, k is an arbitrary round in the set of rounds, and corresponds to the physician's decision points
- The possible actions (i.e. treatments to give the patient) at a given decision point k is notated $\mathbb{A}_k$, where $a_k \in \mathbb{A}_k$ is the action that is taken at decision point k, and the allowed actions at decision point k given state action-history $\bar{s}_k, \bar{a}_{k-1}$ is notated $\Psi(\bar{s}_k, \bar{a}_{k-1})$
- The previous actions taken before decision point k, the action history (or treatment history), are notated $\bar{a}_{k-1} = (a_1, ..., a_{k-1})$
- s_k represents the "state" (i.e. the covariates) at decision point k, and $Y(\bar{a}_K)$ is the final outcome of interest after all actions have been taken, and $\bar{s}_k$ is the state history

Definition of a Dynamic Treatment Regime

- The potential values of the covariates for patient $\omega \in \Omega$ (the set of patients) under a hypothetical set of actions $\bar{a_K}$ in the set of all possible sets of actions $\mathbb{A}$ is written $W^* = \{S_2^*(a_1), S_2^*(\bar{a}_2), ..., S_K^*(\bar{a}_{K-1}), Y^*(\bar{a}_K) \; \forall \; \bar{a}_K \in \bar{\mathbb{A}}_K\}$
- A dynamic treatment regime is a set of rules (i.e. functions) $d = (d_1(s_1), d_2(s_2, a_1), d_3(s_3, \bar{a}_2), ...d_K(s_K, \bar{a}_{K-1}))$ that determine how a patient should be treated given any possible current covariates and past treatment history
- The set of all possible state action-history pairs at decision point k is defined as $\Gamma_k = \{\bar{s}_k, \bar{a}_{k-1} \in \bar{\mathbb{S}}_k \times \bar{\mathbb{A}}_{k-1} \text{ st } \forall j \in \{1, ..., k\} \text{ we have } (j \neq 1)\; a_{j-1} \in \Psi_j(\bar{s}_j, \bar{a}_{j-2}) \text{ and } P(\bar{S}_j^*(\bar{a}_{j-1}) = \bar{s}_j) > 0\}$

Potential Outcomes of a regime d

- The set $\mathbb{D}$ of Ψ-specific dynamic treatment regimes is the set of all d such that for all $k \in \{1, 2, ..., K\}$ the rule d_k is a mapping from Γ_k the set of all possible state action-history pairs at decision point k to $\mathbb{A}_k$ the set of actions at k such that $d_k(\bar{s}_k, \bar{a}_{k-1}) \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})$
- The potential outcomes associated with d are defined as $\{S_2^*(d_1), ..., S_k^*(d_{k-1}), ..., S_K^*(d_{K-1}), Y^*(d))\}$, and for all $\omega \in \Omega$ with $S_1(\omega) = s_1$ we notate

$$d_1(s_1) = u_1$$

$$S_2^*(d_1)(\omega) = S_2^*(u_1)(\omega) = s_2$$

$$...$$

$$d_K(\bar{s}_K, \bar{u}_{K-1}) = u_K$$

$$Y^*(d)(\omega) = Y^*(\bar{u}_k)(\omega) = y$$

Defining an Optimal Regime

- The Optimal Regime is the one that, for a patient entering with covariates s_1, maximizes the expected value of their potential outcome $Y^*(d)$, thus we define $d^{\text{opt}} \in \mathbb{D}$ as the regime such that

$$\mathbb{E}[Y^*(d)|S_1 = s_1] \leq \mathbb{E}[Y^*(d^{\text{opt}})|S_1 = s_1] \; \forall d \in \mathbb{D} \text{ and } \forall s_1 \in \mathbb{S}_1$$

- Because for each patient we can only observe one potential outcome, our goal is to estimate d^{opt} using the data actually observed from a sample of patients from Ω for which baseline evolving covariate information and treatment history received is available

Observational vs SMART data

- Data used for estimating a dynamic treatment regime are typically observational or from a Sequential Multiple Assignment Randomized Trial
- In the observational setting, no intervention is necessary, treatment assignment is determined by routine clinical practice
- Many suitable records of treatments from routine practice are randomly selected and their results are combined for use in training a model to determine d^{opt}
- Because the data are observational it is necessary for some learning methods to adjust for the propensity of treatment

SMART

- A SMART is a trial design in which patients are randomly started on one treatment, and then at one or several fixed decision points re-randomized, generating data that can be used to train the model d^{opt} without having to worry about mis-specifying the propensity of treatment
- In a SMART the randomization probabilities at point k may still depend on $\bar{s}_k, \bar{a}_{k-1}$ to improve the expected health outcomes for the patients in the trial

Example SMART Data

patient_id	round	sbp	previous_treatment	current_treatment	outcome
246	1	159	*NA*	A	0
246	2	159	A	B	0
247	1	143	*NA*	A	1
247	2	133	A	B	1
248	1	133	*NA*	A	1
248	2	129	A	A	1
249	1	139	*NA*	A	1
249	2	129	A	A	1
250	1	148	*NA*	A	1
250	2	142	A	B	1
251	1	141	*NA*	B	1
251	2	146	B	A	1
252	1	140	*NA*	B	0
252	2	141	B	B	0
253	1	141	*NA*	B	1
253	2	144	B	A	1

Assumptions

- Regardless of whether we are using observational or SMART data, certain assumptions must be made for the models used to estimate d^{opt}
- We assume the outcomes and covariates observed are the potential outcomes and covariates under the treatments actually administered
- We must also make the Stable Unit Treatment Value Assumption (the patient's covariates and outcomes are unaffected by how other patients are treated)
- Assume that there are no unmeasured confounders (this is satisfied by design in the SMART)
- In this talk, all assumptions are assumed to hold, but the unmeasured confounder assumption in the observational setting can sometimes be addressed with Instrumental Variables

Estimating the Optimal Regime

- It is only possible to estimate d^{opt} from the available data if we actually have data for all treatment options in $\Psi_k(\bar{s}_k, \bar{s}_{k-1})$, thus the class of Ψ regimes we can consider is limited by the data available
- Two approaches to estimating d^{opt} in the framework previously specified are Q- and A- Learning, which both involve backwards induction, relying on the fact that the optimal regime can be defined recursively
- First, note that: $d_K^{(1)\text{opt}}(\bar{s}_K, \bar{a}_{K-1}) = \arg\max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} \mathbb{E}\left[Y^*(\bar{a}_{K-1}, a_K) \mid \bar{S}_K^*(\bar{a}_{K-1}) = \bar{s}_K\right]$

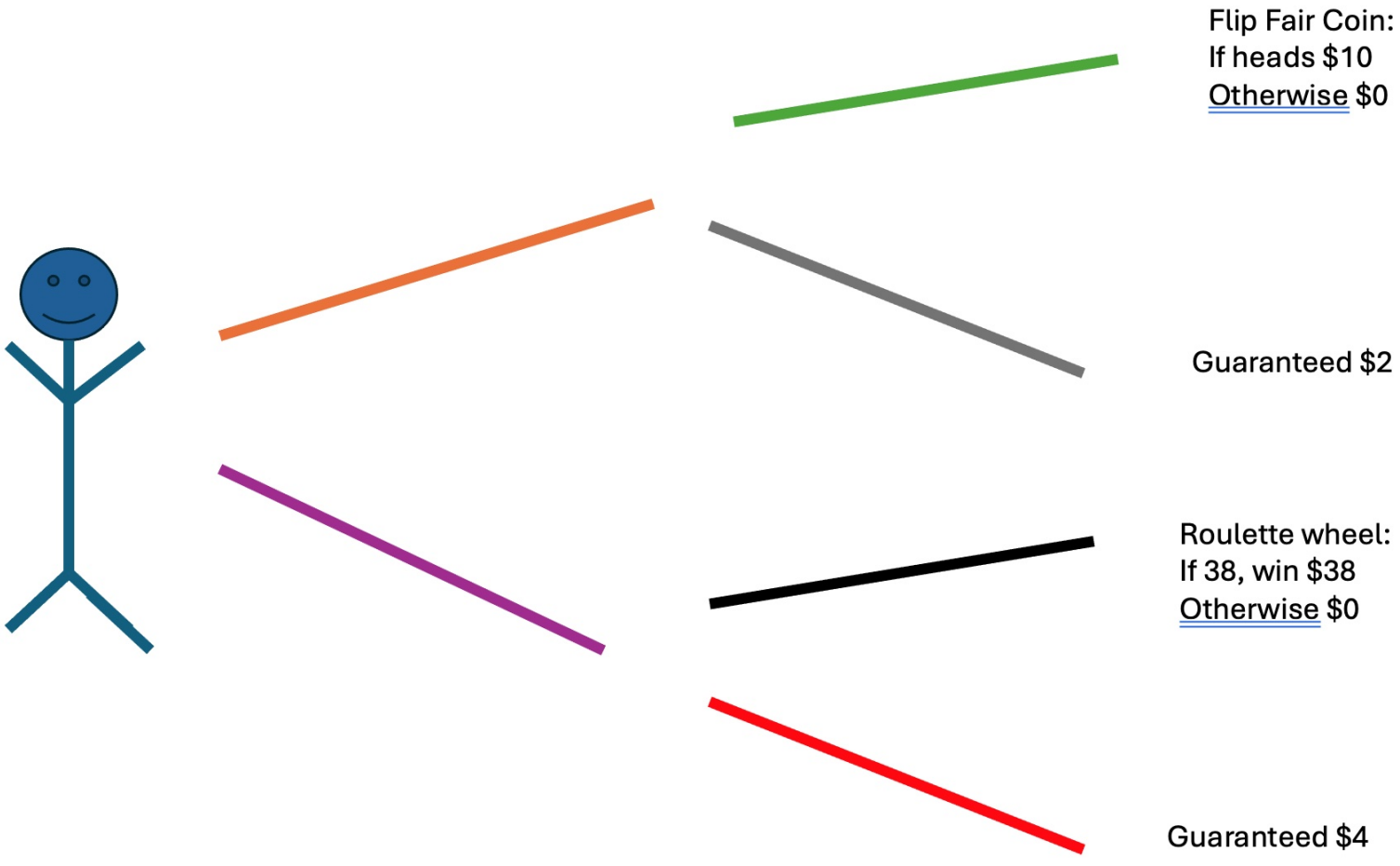
Backwards Induction

- $d_K^{(1)\text{opt}}(\bar{s}_K, \bar{a}_{K-1}) = \arg\max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} \mathbb{E}\left[Y^*(\bar{a}_{K-1}, a_K) \mid \bar{S}_K^*(\bar{a}_{K-1}) = \bar{s}_K\right]$
- Thus, the optimal regime at decision K is the one that maximizes the expected value of the outcome (given the covariates and treatment history)
- We write that max expected value under that optimal decision at K as $V_K^{(1)}(\bar{s}_K, \bar{a}_{K-1}) = \max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} \mathbb{E}\left[Y^*(\bar{a}_{K-1}, a_K) \mid \bar{S}_K^*(\bar{a}_{K-1}) = \bar{s}_K\right]$
- We then recurse $V_K^{(1)}$ to define the optimal decision at $K-1$: $d_{K-1}^{(1)\text{opt}}(\bar{s}_{K-1}, \bar{a}_{K-2}) = \arg\max_{a_{K-1} \in \Psi_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2})} \mathbb{E}\left[V_K^{(1)}(\bar{a}_{K-2}, a_{K-1}) \mid \bar{S}_{K-1}^*(\bar{a}_{K-2}) = \bar{s}_{K-1}\right]$

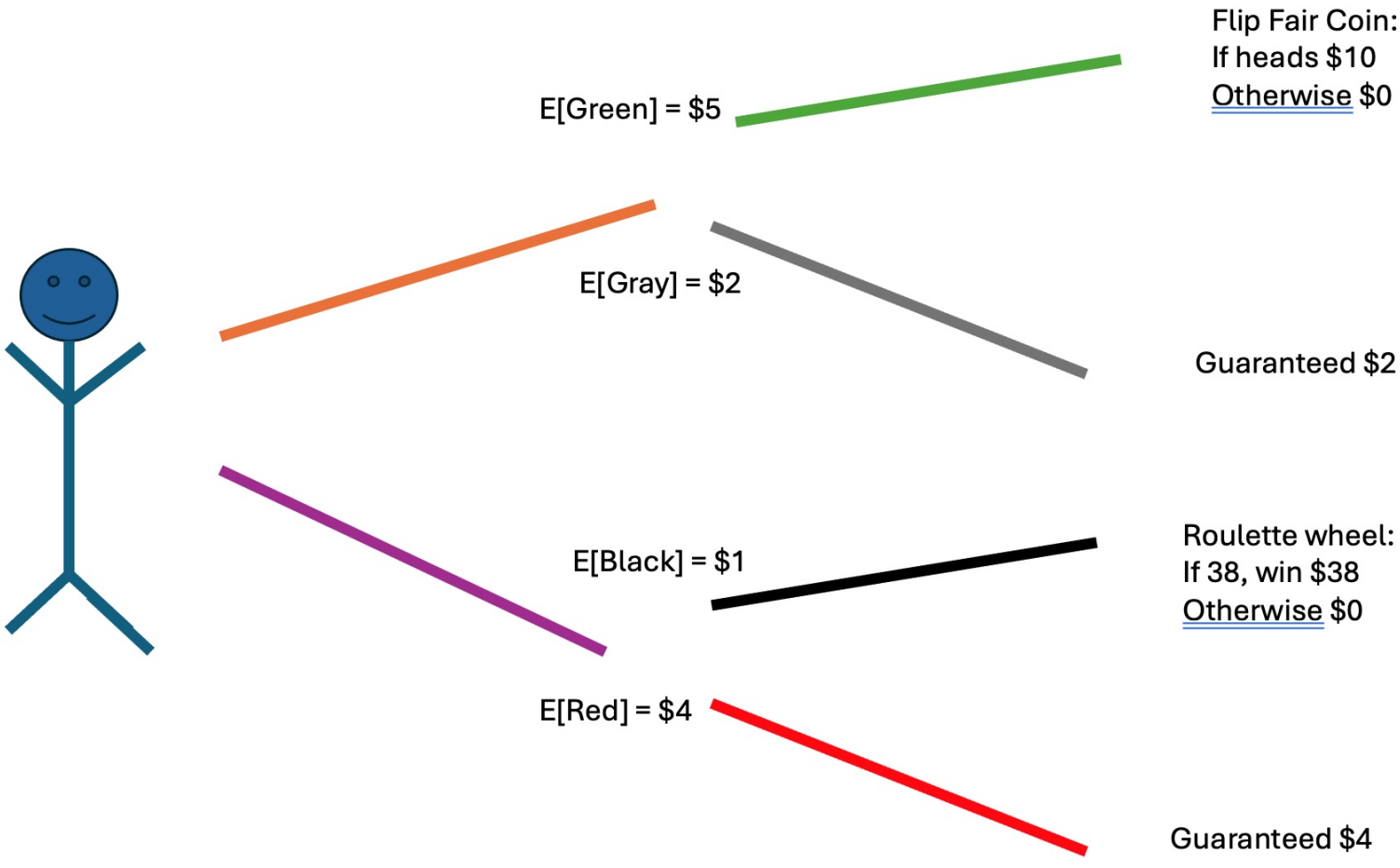
Backwards Induction

- We then use the optimal decision at point $K-1$ to define the value at $K-1$ as
- $V_{K-1}^{(1)}(\bar{s}_{K-1}, \bar{a}_{K-2}) = \max_{a_{K-1} \in \Psi_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2})} \mathbb{E}\left[V_K^{(1)}(\bar{a}_{K-2}, a_{K-1}) \mid \bar{S}_{K-1}^{*}(\bar{a}_{K-2}) = \bar{s}_{K-1}\right]$
- This procedure is then continued backwards until the first treatment is done
- In other words, you are always selecting the treatment that maximizes the expected value of the outcome assuming that you choose the optimal treatments in the future

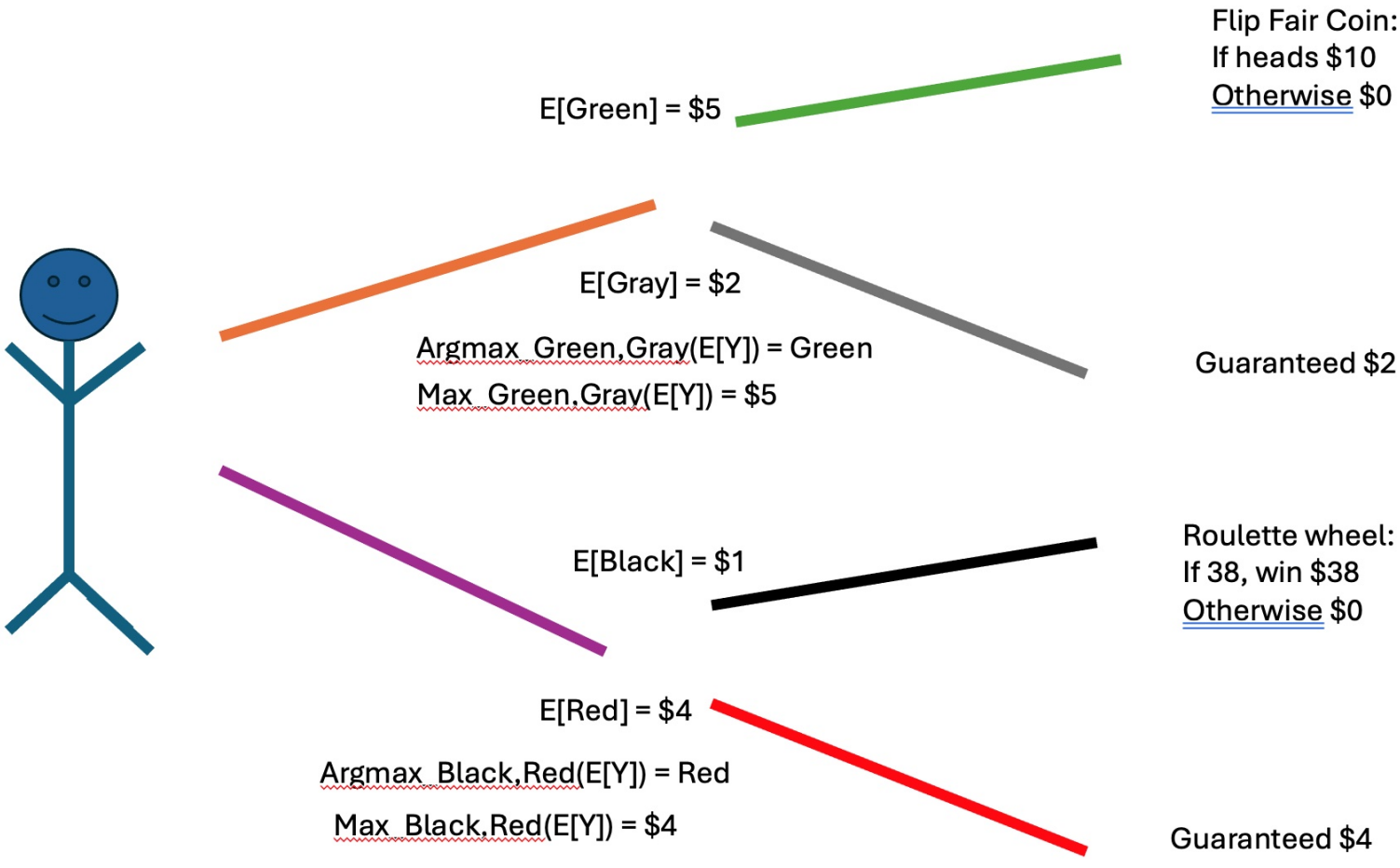
Simple Example



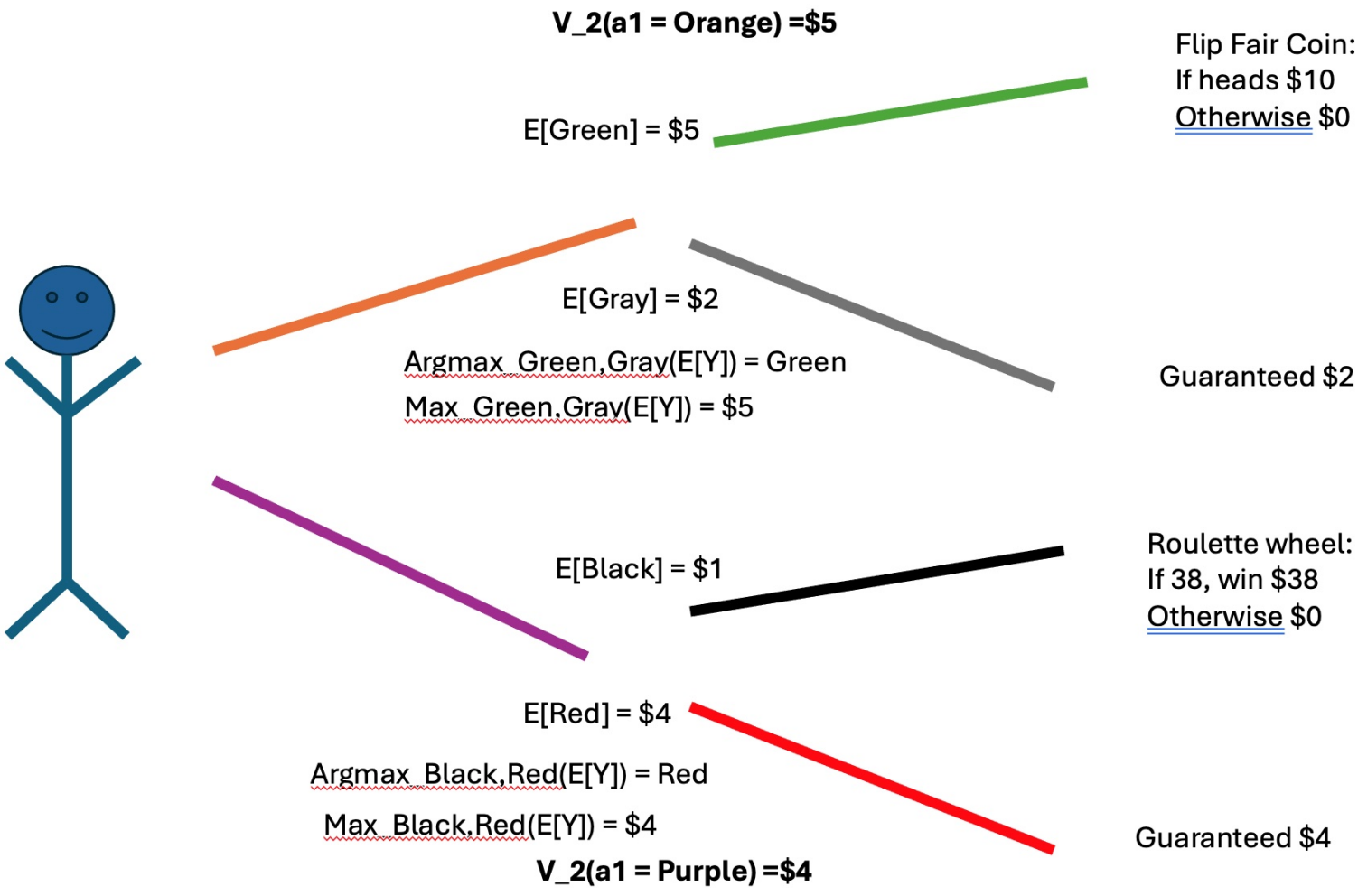
Simple Example



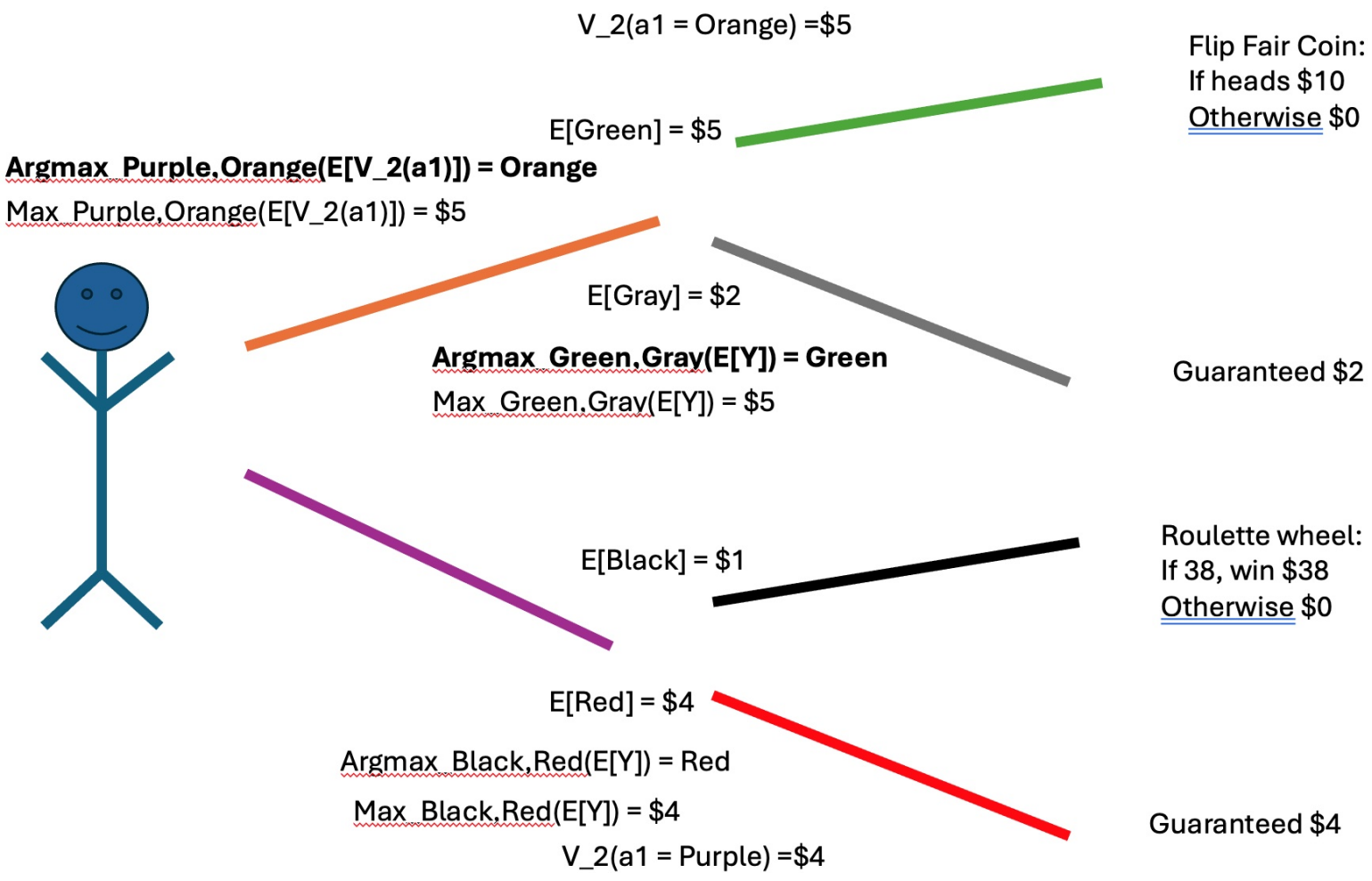
Simple Example



Simple Example



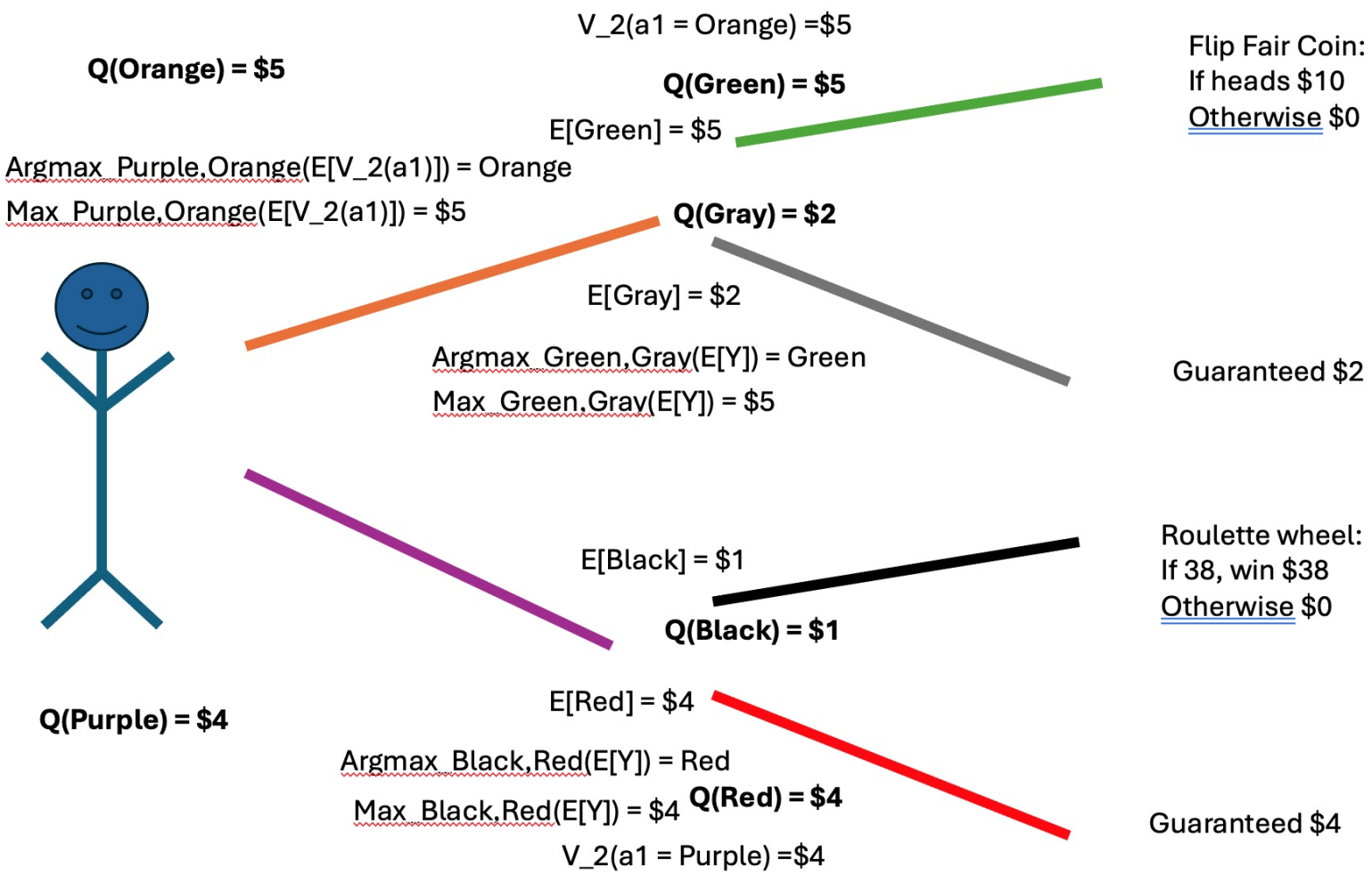
Simple Example



Quality Function

- The quality of each possible treatment at the final decision point K given action-history and covariates is defined as $Q_K(\bar{s}_K, \bar{a}_K) = E[Y|\bar{S}_K = \bar{s}_K, \bar{A}_K = \bar{a}_K]$
- Note that the Q is a function of $\bar{a}_k$ NOT $\bar{a}_{k-1}$, because it gives the quality of a current action given past actions and current covariates, so it is also a function of a current action
- The quality of previous treatments is again defined recursively:
  $Q_k(\bar{s}_k, \bar{a}_k) = \mathbb{E}\left[V_{k+1}(\bar{s}_k, S_{k+1}, \bar{a}_k) \mid \bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k\right]$
- Q is the quality of a treatment, given future treatments are decided optimally, at point k, while V is the value of the patient's state action-history at decision point $k+1$

Simple Example



So... how do we estimate the optimal regime?

- The regime is estimated by either estimating the Q functions at each decision point k, functions that give the qualities of the treatments at each point, starting closest to the end and then moving backwards, OR
- In the binary setting (2 treatments) the regime can be estimated by estimating the contrast function at each decision point k, which is estimating the difference between the two Q functions without explicitly modeling each Q- function, using that function and another function (the h function, which gives the part of the effect on the outcome that is not different for the two treatments) to estimate V and then moving backwards
- The first method above is Q-learning, the second is A-learning

Q-Learning

- In Q-Learning, estimation of d^{opt} is accomplished by directly modeling the Q functions and fitting them
- One may posit $Q_k(\bar{s}_k, \bar{a}_k; \xi_k)$ for $k = K, K-1, ..., 1$ for the Q-functions, with ξ_k as a finite dimensional parameter
- The models may be linear or nonlinear in ξ_k, can include main effects, interactions, etcetera
- The estimating equations are solved at K, the $V_{K,i}$ is estimated for each patient i, the values are projected backwards, and the estimating equations for $K-1$ are solved using $\hat{V}_{K,i}$ in place of Y, and so on

Estimating with WLS

- If a weighted least squares model is used for Q_k at every decision point, the estimating equation is:

$$\sum_{i=1}^{n} \frac{\partial Q_k(\bar{S}_{ki}, \bar{A}_{ki}; \xi_k)}{\partial \xi_k} \Sigma_k^{-1}(\bar{S}_{k,i}, \bar{A}_{k,i}) \times \{\hat{V}_{(k+1),i} - Q_k(\bar{S}_{k,i}, \bar{A}_{k,i}; \xi_k)\} = 0$$

- Once a model (such as the above) is fitted, at each decision point k we now can model the quality of each treatment for each possible state action-history at that decision point, taking the action that maximizes the quality at a given state action-history gives an estimate of the optimal treatment at point k, and taking this for all $k \in \{1, 2, ..., K\}$ gives an estimated optimal treatment regime $\hat{d}_Q^{\text{opt}}$

Hypothetical Model

- Suppose there are two treatment rounds and two treatments, a simple model for the Q-functions would be:
- $Q_1(s_1, a_1; \xi_1) = H_1^T \beta_1 + a_1(H_1^T \phi_1)$
- $Q_2(\bar{s}_2, \bar{a}_2; \xi_1) = H_2^T \beta_2 + a_2(H_2^T \phi_2)$
- Where $H_1 = (1, s_1^T)^T$ and $H_2 = (1, s_1^T, a_1, s_2^T)$
- It should be noted that even when the data is from a SMART, the estimated regime may be inconsistent unless all of the models for the Q-functions are correctly specified.

The Contrast

- Advantage Learning (A- Learning) takes advantage of the fact that the optimal treatment only requires knowing the difference between $Q_k(\bar{s}_k, \bar{a}_{k-1}, a_k = 1) - Q_k(\bar{s}_k, \bar{a}_{k-1}, a_k = 0)$, or the advantage of choosing treatment 1 over treatment 0
- $C(\bar{s}_k, \bar{a}_{k-1}) = Q_k(\bar{s}_k, \bar{a}_{k-1}, a_k = 1) - Q_k(\bar{s}_k, \bar{a}_{k-1}, a_k = 0)$ is known as the contrast function
- The Quality function $Q(\bar{s}_k, \bar{a}_k)$ can be written as the sum of the contrast function and another function $h(\bar{s}_k, \bar{a}_{k-1})$ that is not a function of the current treatment

A- Learning

- First, posit models for the Contrast functions $C(\bar{s}_k, \bar{a}_{k-1}; \psi_k), k = 1, ..., K$, depending on parameters ψ_k
- Robins (2004) showed that all consistent and asymptotically normal estimators for the ψ_k's can be be represented in the following general form:

$$\sum_{i=1}^{n} \lambda_k \left(\overline{S}_{ki}, \overline{A}_{(k-1)i}; \psi_k \right) \left\{ A_{ki} - \pi_k \left(\overline{S}_{ki}, \overline{A}_{(k-1)i}; \phi_k \right) \right\}$$

$$\times \{ \tilde{V}_{(k+1)i} - A_{ki} C_k \left(\overline{S}_{ki}, \overline{A}_{(k-1)i}; \psi_k \right) - \theta_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \beta_k) \} = 0$$

Double Robustness

- If $Var(Y|\overline{A}_{(k-1)}, \overline{S}_k)$ is constant, we have:
- $\lambda_k(\overline{S}_k, \overline{A}_{(k-1)}) = \frac{\partial}{\partial \psi_k} C(\bar{s}_k, \bar{a}_{k-1}; \psi_k)$, and $\theta_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}) = h_k(\bar{S}_k, \bar{A}_{(k-1)})$
- Parametric models are typically adopted for these functions
- If the data is not from a SMART, the propensities of treatment $\pi_k(\bar{s}_k, \bar{a}_{k-1})$ must also be modeled
- A- Learning has the double-robustness property, if either $h_k(\bar{S}_k, \bar{A}_{(k-1)})$ or $\pi_k(\bar{s}_k, \bar{a}_{k-1})$ is correctly specified the estimator for ϕ_k is consistent, provided that $C(\bar{s}_k, \bar{a}_{k-1})$ is correctly specified

Estimating Equations

- The estimating equations for the A- Learning Model under these assumptions is:

$$\sum_{i=1}^{n} \frac{\partial}{\partial \psi_k} C(\bar{S}_k, \bar{A}_{k-1}; \psi_k) \left\{ A_{ki} - \pi_k \left(\overline{S}_{ki}, \overline{A}_{(k-1)i}; \phi_k \right) \right\}$$

$$\times \{ \tilde{V}_{(k+1)i} - A_{ki} C_k \left(\overline{S}_{ki}, \overline{A}_{(k-1)i}; \psi_k \right) - h_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \beta_k) \} = 0$$

$$\sum_{i=1}^{n} \frac{\partial h_k \left(\overline{S}_k, \overline{A}_{(k-1)}; \beta_k \right)}{\partial \beta_k} \{ \tilde{V}_{(k+1)i} - A_{ki} C_k(\overline{S}_{ki}, \overline{A}_{(k-1)i}; \psi_k) -$$

$$h_k(\overline{S}_{ki}, \overline{A}_{(k-1)i}; \beta_k) \} = 0$$

Backwards Iteration

- As before, these estimating equations are solved through backwards iteration, that is, solve jointly for $(\psi_K, \beta_K, \phi_K)$ (the parameters of the contrast, state action-history, and propensity functions), then use these parameters to get Q of each treatment for each person, and project the larger back as V
- If the contrast is positive, treatment 1 is optimal, otherwise treatment 0 is optimal
- How well our model estimates d^{opt} depends on how close the specification of C is to the truth, as well as the correct specification of either h or π
- There exist other kinds of A- Learning (e.g. regret based A- Learning) that work slightly differently and will not be discussed in this talk

Efficiency vs Robustness

- Q- Learning is more efficient than A-learning when models are correctly specified, but Q- Learning can be more sensitive to model misspecification
- From simulations this trade-off can be clearly seen, although Q- and A- learning often end up with similar results
- In this presentation we will show one such simulation, although others are contained in the paper

Simulation: One Decision Point

The simulation we will cover from the paper is the simplest one: only a single decision point, no backwards induction, and the 'regime' only consists of a single treatment. To avoid having to do extra typesetting, I just have an image from the paper:

In this and the next section, $n = 200$. Here, the observed data are (S_{1i}, A_{1i}, Y_i), $i = 1, \ldots, n$. With $\text{expit}(x) = e^x/(1 + e^x)$, we used the class of generative models

$$S_1 \sim \text{Normal}(0,1), A_1 \,|S_1 = s_1 \sim \text{Bernoulli}\{\text{expit}(\phi_{10}^0 + \phi_{11}^0 s_1 + \phi_{12}^0 s_1^2)\}, Y \,|\, S_1 = s_1,$$
$$A_1 = a_1 \sim \text{Normal}\{\beta_{10}^0 + \beta_{11}^0 s_1 + \beta_{12}^0 s_1^2 + a_1(\psi_{10}^0 + \psi_{11}^0 s_1), 9\}, \tag{34}$$

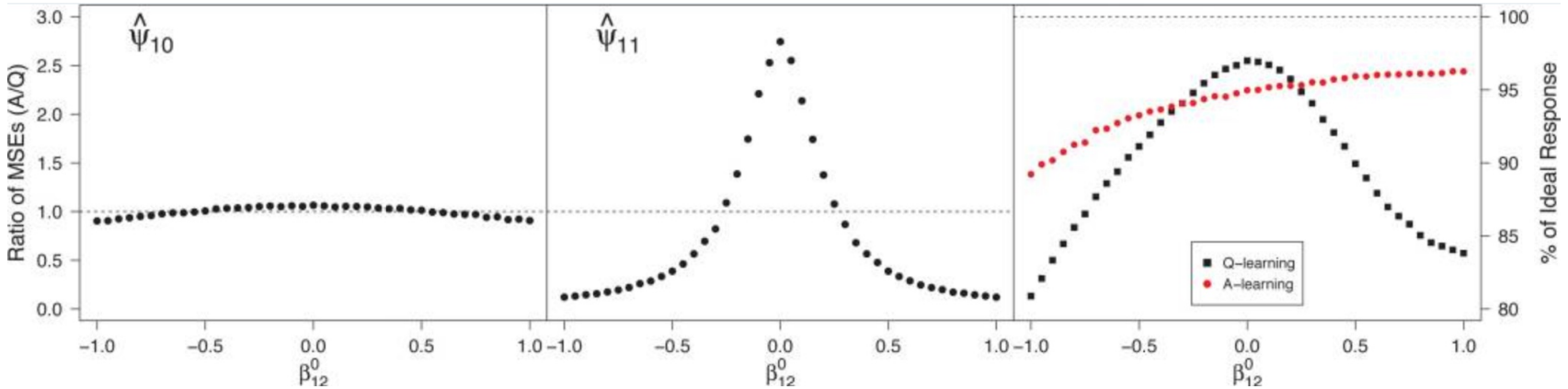
indexed by $\phi^0 = (\phi_{10}^0, \phi_{11}^0, \phi_{12}^0)^T, \beta^0 = (\beta_{10}^0, \beta_{11}^0, \beta_{12}^0)^T, \psi^0 = (\psi_{10}^0, \psi_{11}^0)^T$, so that $d^{\text{opt}} = d_1^{\text{opt}}, d_1^{\text{opt}}(s_1) = I(\psi_{10}^0 + \psi_{11}^0 s_1 > 0)$. For *A*-learning, we assumed models $h_1(s_1; \beta_1) = \beta_{10} + \beta_{11} s_1$, $C_1(s_1; \psi_1) = \psi_{10} + \psi_{11} s_1$, and $\pi_1(s_1; \phi_1) = \text{expit}(\phi_{10} + \phi_{11} s_1)$, and for *Q*-learning we used $Q_1(s_1, a_1; \xi_1) = h_1(s_1; \beta_1) + a_1 C_1(s_1; \psi_1)$. These models involve correctly specified contrast functions and are nested within the true models, with $h_1(s_1; \beta_1)$, and hence the *Q*-function, correctly specified when $\beta_{12}^0 = 0$. The propensity model $\pi_1(s_1; \phi_1)$ is correctly specified when $\phi_{12}^0 = 0$. To study the effects of misspecification, we varied $\beta_{12}^0 \text{and} \phi_{12}^0$ while keeping the others fixed, considering parameter settings of the form $\phi^0 = (0, -2, \phi_{12}^0)^T, \beta^0 = (1, 1, \beta_{12}^0)^T, \psi^0 = (1, 0.5)^T$.

Results under Correct Model

- When the models were both correctly specified, Q-learning was (measured by MSE ratio) 6% more efficient in estimating ϕ_{10}^0 and 174% more efficient in estimating ϕ_{11}^0
- However, the correct decision was still chosen by A-Learning 95% of the time, compared to 97 % for Q-Learning

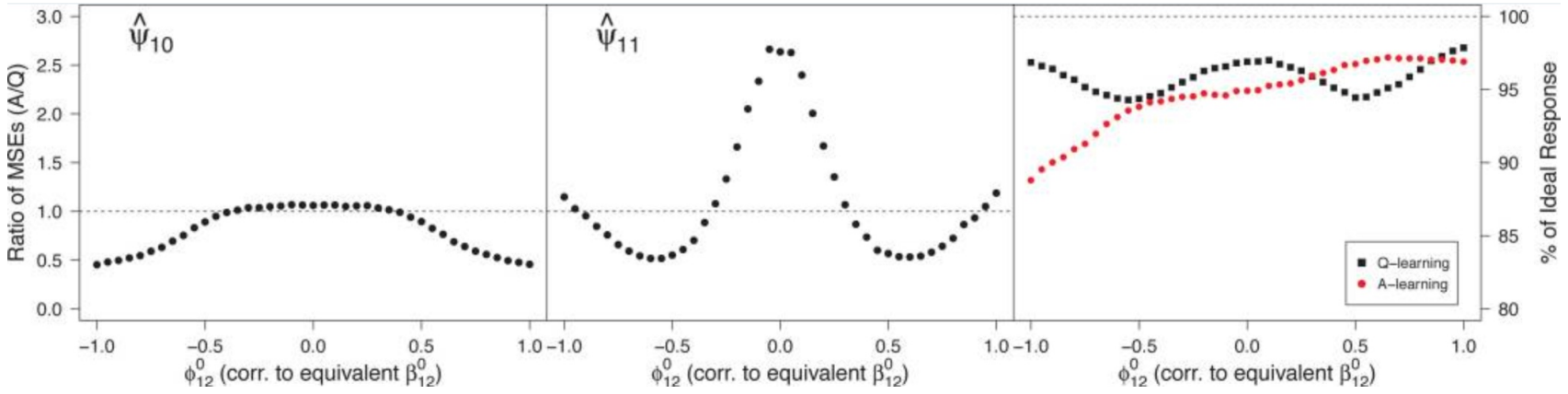
Results under Incorrect Q Model

- Suppose that β_{12}^0 is misspecified, but that the propensity model is correctly specified
- This will result in Q- Learning having bias, but A-Learning not having bias due to its double robustness, resulting in the bias-variance trade-off shown in the panels below:



Results under Incorrect Q Model and Propensity Model

- Suppose that β_{12}^0 is misspecified, but that the propensity model for A- Learning is also misspecified, and that they are both misspecified to similar degrees
- In the below visual, β_{12}^0 and ϕ_{12}^0 were moved together so that the t-tests of significance in their respective linear and logistic regressions were the same



My Own Experimentation

- Out of curiosity, I decided to try and create a simulation of my own in R and implement the methods from the paper, with a ”realistic”-ish but simplistic SMART
- In my simulation there is only 1 covariate, SBP, and two rounds of treatment at which a patient can receive either A or B, and based on their SBP after the round 2 treatment (which could be thought of as the ’round 3’ SBP) the patient has $Y = 0$ with probability $\frac{1}{1+exp(17-0.11*sbp)}$ and $Y = 1$ otherwise, which corresponds to about 0.5 survival probability with final sbp of 155, 0.75 survival probability for final sbp 145, and 0.25 survival probability for final sbp 165

Drug Effects

- Drug A reduces SBP by about 5 whenever it is applied if the person has somewhat high blood pressure, drug B does nothing in the first round, but decreases blood pressure by 15 in the second round if the previous treatment was also B, provided the person has a very high blood pressure
- I designed the simulation to see if the methods could correctly decide in round 1 to assign B when the person enters with high blood pressure and A if the person enters with not-so-high blood pressure, and then in round 2 to only give B if the person received B previously
- This simulation data is what I showed earlier in the presentation

Model One

- Two models were used, although they may give equivalent results, one model is better used for understanding how Q- Learning works, and the other is better used for the way A- Learning works
- The first model takes at each time point a separate regression for the two different current treatments:

$$\text{lm}(\text{outcome} \sim \text{sbp} + \text{prev A} + \text{prev A:sbp}, \text{data} = \text{round 2 A})$$
$$\text{lm}(\text{outcome} \sim \text{sbp} + \text{prev A} + \text{prev A:sbp}, \text{data} = \text{round 2 B})$$

- Then for all patients in round 2, use each regression to predict Y (which is binary, meaning this is the linear probability model), and whichever is higher,
  $\text{round 2\$V=pmax(predict(Q reg A, round 2),predict(Q reg B, rou}$
  is backwards inducted into round 1,
  $\text{round 1 <- left join(round 1, round 2, by = ”patient id”)}$

Model One

- Repeat this in round 1, regressing on the backwards inducted V:

$$\text{lm}(\text{V} \sim \text{sbp}, \text{data} = \text{round 1 A})$$

$$\text{lm}(\text{V} \sim \text{sbp}, \text{data} = \text{round 1 B})$$

Which ever current treatment model has the higher expected *V* under these models is the optimal treatment decision at decision point 1, and the models from the previous slide give the optimal treatment decision at decision point 2

Model Two

- Create one model at each decision point, the model at decision point 2 can is then broken up into an h and a Contrast function, when the contrast is positive A is optimal, when the contrast is negative B is optimal
- For each patient, again use the better treatment to get V, iterate backwards, repeat
- Details are in the code
- Note that in each of these formulations neither the contrast nor the full Q functions are specified fully correctly

Conclusion

- Both methods give the same results, and all results seem reasonable, you can check them out yourself in the code that I've provided
- In sum, both Q- and A- Learning provide good approaches to learning an Optimal Dynamic treatment regime, although both depend on certain relatively strong assumptions that may not be valid or verifiable in the real world
- Methods for learning the Optimal Dynamic Treatment Regime is an interesting open problem in statistics and causal inference

Thanks for your attention!