Section: Module 9

Causal Machine Learning / Mediation

Kentaro Nakamura

GOV 2002

November 21st, 2025

Logistics

- Important Dates
  - Problem Set 8: Due Next Monday (Nov 24th)
  - Problem Set 9: Due December 1st
  - Problem Set 10: Due December 8th
  - Review Session: December 8th (CGIS K354)
  - Final Exam: December 11th

- Today's agenda (so many topics!)
  - Flexible Weighting
    - Covariate Balancing Propensity Score
    - Calibration Method
  - Causal Machine Learning
  - Causal Mediation Analysis
    - Controlled Direct Effect
    - Natural Direct / Indirect Effect

Toward Better Estimation of Propensity Score

- Recall from the last week that both HT and Hajek estimators require estimation of propensity score

- However, if propensity score is misspecified, we have the bias

- Three different approaches (next week)
  1. Covariate Balancing Propensity Score (CBPS)
     - Estimate propensity score s.t. we achieve balance
     - But still assume parametric assumption on propensity score function
  2. Calibration: Entropy Balancing / Stable Weights
     - Estimate weight so that we achieve balance
     - We no longer estimate propensity score
  3. Causal Machine Learning / Semiparametric Estimation
     - Flexibly estimate propensity score / outcome models
     - Relax parametric assumption as much as possible

Covariate Balancing Propensity Score (1)

- Think about the estimation of propensity score model
- Popular choice is logistic regression with parameter θ:

$$\pi_\theta(X_i) = \frac{\exp(\boldsymbol{X}_i^\top \theta)}{1 + \exp(\boldsymbol{X}_i^\top \theta)}$$

- Recall that the log likelihood of logistic regression model is

$$\ell(\theta) = \sum_{i=1}^{n} \Big(T_i \log \pi_\theta(X_i) + (1 - T_i) \log(1 - \pi_\theta(X_i)) \Big)$$

- To obtain MLE, we want to maximize the log likelihood. The first order condition is written as

$$\frac{\partial}{\partial \theta} \ell(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{T_i}{\pi_\theta(X_i)} - \frac{1 - T_i}{1 - \pi_\theta(X_i)} \right) \underbrace{\frac{\partial}{\partial \theta} \pi_\theta(X_i)}_{:= \pi'_\theta(X_i)} = 0$$

Covariate Balancing Propensity Score (2)

- The first order condition can be re-written as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_\theta(X_i)}\pi'_\theta(X_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{1-T_i}{1-\pi_\theta(X_i)}\pi'_\theta(X_i)$$

  which can be interpreted as "balancing $\pi'_\theta(X_i)$"

- We can instead directly balance covariates rather than $\pi'_\theta(X_i)$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_\theta(X_i)}X_i = \frac{1}{n}\sum_{i=1}^{n}\frac{1-T_i}{1-\pi_\theta(X_i)}X_i$$

- Then, even if propensity score is misspecified, as long as the estimated model balances the covariates, we are fine
  - The condition above only balances the first moment (expectation)
  - We can balance the higher moments too.
  - Think as modeling balancing score using a parametric model
- BUT still we use parametric model $\pi_\theta(X_i)$!

Calibration Method: Entropy Balancing

- **Idea**: Without estimating propensity score model, we just want to learn the *weight* that balances the covariates
- **Entropy balancing**: Find the weight that matches the moment exactly

$$\begin{aligned} \min_{w_i} \quad & \sum_{i:T_i=0} w_i \log \frac{w_i}{q_i} \\ \text{s.t.} \quad & \sum_{i:T_i=0} w_i f(X_i) = \frac{1}{n_1} \sum_{i:T_i=1} f(X_i) \\ & \sum_{i:T_i=0} w_i = 1, \quad w_i \geq 0 \end{aligned}$$

- Each unit has different weight
  - More flexible, no direct modeling of propensity score
  - However, in reality we can balance only the finite dimensional moment (i.e., cannot directly balance two distributions without assumptions)
  - If there is imbalances in higher moments which are not in optimization problem, we suffer from bias
  - Thus, calibration method is in some sense still parametric

Doubly Robust Estimation

- We learn two approaches to estimate causal effect: outcome model and weighting

$$\mathbb{E}[Y_i(1) - Y_i(0)]$$
$$= \begin{cases} \mathbb{E}[\mathbb{E}[Y_i \mid T_i = 1, X_i] - \mathbb{E}[Y_i \mid T_i = 0, X_i]] \text{ (Outcome)} \\ \mathbb{E}\left[\frac{T_i Y_i}{\pi(X_i)} - \frac{(1-T_i)Y_i}{1-\pi(X_i)}\right] \quad \text{(weighting)} \end{cases}$$

- **Doubly Robust Estimator / Augmented IPW (AIPW)**: Combine weighting (IPW) with outcome model so that if either works, we can estimate causal effect

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\Big)$$
$$+ \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \frac{(1-T_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{\pi}(X_i)}\right)$$

  - It turns out that AIPW can be used for machine learning
  - Try Problem Set 8 Question 2 for Stat286 before taking final

Proof of Double Robustness

- We only prove that the AIPW of $\mathbb{E}[Y_i(1)]$ part is unbiased if either propensity score model or outcome model is correctly specified.

$$
\begin{aligned}
\text{Bias} &:= \mathbb{E}\left[\hat{\mu}_1(X_i) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)}\right] - \mathbb{E}[Y_i(1)] \\
&= \mathbb{E}\left[\frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \Big(Y_i(1) - \hat{\mu}_1(X_i)\Big)\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[T_i Y_i \mid X_i] - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \Big(\mathbb{E}[Y_i(1) \mid X_i] - \hat{\mu}_1(X_i)\Big)\right] \quad \text{(L.I.E)} \\
&= \mathbb{E}\left[\frac{\mathbb{E}[T_i Y_i(1) \mid X_i] - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \Big(\mathbb{E}[Y_i(1) \mid X_i] - \hat{\mu}_1(X_i)\Big)\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[T_i \mid X_i]\mathbb{E}[Y_i(1) \mid X_i] - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \Big(\mathbb{E}[Y_i(1) \mid X_i] - \hat{\mu}_1(X_i)\Big)\right] \\
&= \mathbb{E}\left[\left(\frac{\mathbb{E}[T_i \mid X_i]}{\hat{\pi}(X_i)} - 1\right)\Big(\mathbb{E}[Y_i(1) \mid X_i] - \hat{\mu}_1(X_i)\Big)\right] \\
&= \mathbb{E}\left[\left(\frac{\mathbb{E}[T_i \mid X_i]}{\hat{\pi}(X_i)} - 1\right)\Big(\mathbb{E}[Y_i \mid T_i = 1, X_i] - \hat{\mu}_1(X_i)\Big)\right]
\end{aligned}
$$

Machine Learning: Quick Overview (1)

- Linear regression has many problems
  - Restrictive parametric assumptions
  - Often does not work in the case of high-dimensional covariates
    - To obtain $\hat{\beta} = (X^\top X)^{-1} X^\top Y$, we need to obtain the inverse of $X^\top X$.
    - It is not possible under perfect collinenarity

- We want to use **machine learning** to flexible model the high-dimensional confounding variables

- Why Machine Learning?
  - Flexible (little parametric assumptions)
  - Handle many confounding variables effectively

Machine Learning: Quick Overview (2)

- In many cases, the algorithm is designed to better predict the new points
  - It is not designed for statistical inference of the parameter

- Rather than bias, we often care about MSE (mean squared error)
  - **Bias-Variance Trade-off**

$$\underbrace{\mathbb{E}\left[(y-\hat{f}(x))^2\right]}_{\text{MSE}} = \underbrace{\left(\mathbb{E}[\hat{f}(x)]-f(x)\right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{V}\left[\hat{f}(x)\right]}_{\text{Variance}} + \underbrace{\mathbb{V}[\varepsilon]}_{\text{Irreducible noise}}$$

  - Intiutively, bias refers to underfitting whereas variance refers to overfitting

- We want to use **regularization** to achieve this to some extent
  - Regularization: introduce bias but minimize variance

Example: Ridge Regression

- The optimization problem is written as

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta}\{||Y_i - X\beta||_2^2 + \lambda||\beta||_2^2\}$$

- The closed form solution is given by

$$\hat{\beta}_{\text{ridge}} = (X^\top X + \lambda I_n)^{-1} X^\top Y$$

  - This makes $(X^\top X + \lambda I_n)^{-1}$ always invertible

- Ridge regression is **biased** unless $\lambda = 0$ since

$$\begin{aligned}\mathbb{E}[\hat{\beta}_{\text{ridge}} \mid X] &= (X^\top X + \lambda I_n)^{-1} X^\top \mathbb{E}[Y \mid X] \\ &= \{(X^\top X + \lambda I_n)^{-1} X^\top X\}\beta\end{aligned}$$

Causal Machine Learning

- But these MLs are designed to predict the outcome well, rather than estimating the parameter

- Thus, we might want to regard each ML model $\hat{Y} = \hat{f}(X)$, and using these ML model as a **nuisance parameter** (i.e., we do not interpret), we want to estimate low-dimensional interpretable parameter (e.g., ATE, ATT)

- **Challenge of using ML for Causal Inference**
  1. Regularization Bias: ML model is biased
  2. Overfitting: Too flexible

- **Goal**: You want the way to overcome these two challenges and get confidence interval under the realistic condition
  - We aim to derive asymptotic variance

Recap: Tools for Asymptotic Variance

- **Law of Large Numbers (LLN)**: If $X_1, \cdots, X_n$ are i.i.d.,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}[X_i]$$

- **Central Limit Theorem (CLT)**: If $X_1, \cdots, X_n$ are i.i.d.,

$$\sqrt{n}(\bar{X} - \mathbb{E}[X_i]) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}[X_i])$$

- **Slutsky's Lemma**: If $X_n \xrightarrow{d} X$ for some random variable X and $Y_n \xrightarrow{p} c$ for some constant c,

$$X_n Y_n \xrightarrow{d} cX$$

Taylor Expansion Viewpoint

- Think about the moment-based estimator that depends on both parameter of interest β and nuisance function η, denoted as $m(\beta, \eta)$
  - For regression: $m(\beta) = \mathbb{E}[X_i(Y_i - X\beta)] = 0$
  - You obtain the estimator $\hat{\beta}_n$ as a solution of sample moment $\hat{m}_n(\hat{\beta}_n) = \frac{1}{n}\sum_i X_i(Y_i - X\hat{\beta}_n) = 0$
- Recall that Taylor expansion gives you

$$f(x+h) = f(x) + \nabla f(x)^\top h + \frac{1}{2} h^\top H(x) h + \cdots$$

- With the (functional version of) Taylor expansion of sample moment $m_n(X_i; \hat{\beta}_n, \hat{\eta}_n)$ around true parameter (β_0, η_0),

$$m_n(\hat{\beta}_n,\hat{\eta}_n) = m_n(X_i; \beta_0, \eta_0) + \left.\frac{\partial m_n(X_i; \beta, \eta)}{\partial \beta}\right|_{\beta=\beta_0} (\hat{\beta}_n - \beta_0)$$
$$+ \left.\frac{\partial m_n(X_i; \beta, \eta)}{\partial \eta}\right|_{\eta=\eta_0} (\hat{\eta}_n - \eta_0) + \text{Reminder Term}$$

Neyman Orthogonality

- Thus, if we can ignore the reminder term, as $m_n(\hat{\beta}_n, \hat{\eta}_n) = 0$,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \sqrt{n}\left[\frac{\partial m_n(\beta,\eta)}{\partial \beta}\bigg|_{\beta=\beta_0}\right]^{-1} m_n(\beta_0, \eta_0)$$
$$+ \sqrt{n}\left[\frac{\partial m_n(\beta,\eta)}{\partial \beta}\bigg|_{\beta=\beta_0}\right]^{-1} \underbrace{\frac{\partial m_n(\beta,\eta)}{\partial \eta}\bigg|_{\eta=\eta_0}(\hat{\eta}_n - \eta_0)}_{\text{Estimation Error of Nuisance Functions}}$$

- From the case of ridge regression, we learn that $\hat{\eta}_n - \eta_0 \neq 0$ in the case of machine learning due to regularization
- If $\frac{\partial m_n(\beta,\eta)}{\partial \eta}\Big|_{\eta=\eta_0} = 0$, then the second term in the above equation becomes 0
  - In other words, our estimator $\hat{\beta}_n$ becomes not sensitive to the estimation error from nuisance functions $\hat{\eta}_n - \eta_0$
  - This is called **Neyman Orthogonality**
  - We thus need to use $m_n(X_i; \beta_0, \eta_0)$ that satisfies Neyman orthogonality

Asymptotic Normality (1)

- **Motivation**: We want to obtain the asymptotic normality
  - To derive asymptotic normality, we want to use CLT
  - To use central limit theorem, we need the average

- Suppose that $m_n(\beta_0, \eta_0) = \frac{1}{n}\sum_{i=1}^{n} \psi(X_i, \beta_0, \eta_0)$
  - In the case of OLS, $\psi(X_i, \beta_0) = X_i(Y_i - X\beta_0)$
  - Notice that $\mathbb{E}[\psi(X_i, \beta_0, \eta_0)] = 0$

- If our $m_n(\beta_0, \eta_0)$ satisfies Neyman orthogonality, then

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \left[\frac{\partial m_n(\beta, \eta)}{\partial \beta}\bigg|_{\beta=\beta_0}\right]^{-1} \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \psi(X_i, \beta_0, \eta_0)\right)$$

- By central limit theorem, we can obtain

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \psi(X_i, \beta_0, \eta_0)\right) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[\psi(X_i, \beta_0, \eta_0)^2])$$

Asymptotic Normality (2)

- Thus, as $m_n(\beta_0, \eta_0) \xrightarrow{p} m(\beta_0, \eta_0)$ (by LLN), we have

$$\left[\frac{\partial m_n(\beta, \eta)}{\partial \beta}\bigg|_{\beta=\beta_0}\right] \xrightarrow{p} \partial_\beta \mathbb{E}[m(\beta_0, \eta_0)]$$

- Therefore, by Slutsky's lemma, we finally obtain

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \left[\frac{\partial m_n(\beta, \eta)}{\partial \beta}\bigg|_{\beta=\beta_0}\right]^{-1} \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \psi(X_i, \beta_0, \eta_0)\right)$$
$$\xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}[\psi(X_i, \beta_0, \eta_0)^2]}{\partial_\beta \mathbb{E}[m(\beta_0, \eta_0)]}\right)$$

which corresponds to the lecture slide p.6

Influence Function

- The previous page derivation gives you the motivation of influence function
- In order to obtain the asymptotic distribution of $\hat{\beta}$, we want some function ψ s.t.

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i), \quad \mathbb{E}[\psi(X_i)] = 0$$

- If we obtain such $\psi(X_i)$, then we can apply CLT and obtain

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[\psi(X_i)^2])$$

- Such $\psi(X_i)$ is called **Influence function**
  - There are many influence function
  - The one that attains the minimum is called **efficient influence function**
- **Takeaway**: Influence function is a nice object for us to derive asymptotic normality

Causal Mediation Analysis: Overview

- **Estimand**

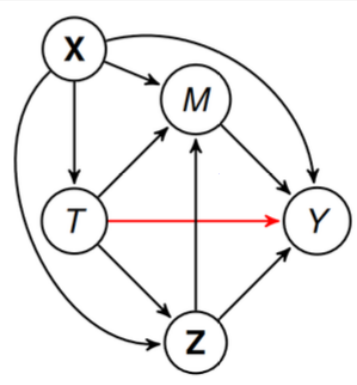
$$\begin{aligned}
\text{Controlled Direct Effect}: &\quad \bar{\xi}(m) = \mathbb{E}[Y_i(1,m) - Y_i(0,m)] \\
\text{Natural Indirect Effect}: &\quad \bar{\delta}(m) = \mathbb{E}[Y_i(t, M_i(1)) - Y_i(t, M_i(0))] \\
\text{Natural Direct Effect}: &\quad \bar{\zeta}(m) = \mathbb{E}[Y_i(1, M_i(t)) - Y_i(0, M_i(t))]
\end{aligned}$$

- NIE / NDE has effect decomposition

$$\overbrace{\mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))]}^{\text{Average Treatment Effect}}$$

$$= \begin{cases} \underbrace{\mathbb{E}[Y_i(1, M_i(1)) - Y_i(1, M_i(0))]}_{\text{Natural Indirect Effect}} + \underbrace{\mathbb{E}[Y_i(1, M_i(0)) - Y_i(0, M_i(0))]}_{\text{Natural Direct Effect}} \\ \underbrace{\mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(1))]}_{\text{Natural Direct Effect}} + \underbrace{\mathbb{E}[Y_i(0, M_i(1)) - Y_i(0, M_i(0))]}_{\text{Natural Indirect Effect}} \end{cases}$$

Controlled Direct Effect: Assumptions



- Identification Assumption: **Sequential Ignorability**

$$\{Y_i(t,m), M_i(t')\} \perp\!\!\!\perp T_i \mid X_i \quad \text{(Treatment Uncounfoundedness)}$$

$$Y_i(t,m) \perp\!\!\!\perp M_i \mid X_i = x, T_i, Z_i \quad \text{(Mediator Uncounfoundedness)}$$

  - Even under experiment, mediator unconfoundedness is difficult
  - As a result, people use them with sensitivity analysis

- However, if we simply run regression controlling M_i and Z_i, we suffer from **post-treatment bias**
  - Thus, under this assumption, new strategies are needed
  - They are called **Structural Nested Mean Model** / **Marginal Structural Model**
  - Note that they are also used for time-varying treatment / longitudinal studies (where you think $Y_i(T_1, T_2, T_3, \cdots)$)

Structural Nested Mean Model (SNMM) (1)

- For simplicity, let's assume **no intermediate interaction**

$$\begin{aligned}&\mathbb{E}[Y_i(t,m) - Y_i(t,m') \mid X_i, T_i, Z_i] \\ &= \mathbb{E}[Y_i(t,m) - Y_i(t,m') \mid X_i, T_i]\end{aligned}$$

- **Structural nested mean model** assumes that the conditional expectation of potential outcome is written as

$$\mathbb{E}[Y_i(t,m) \mid X_i] = \mathbb{E}[Y_i(t,0) \mid X_i] + \gamma(t,m,X_i)$$

  where γ is called **blip function**

$$\gamma(t,m,x) := \mathbb{E}[Y_i(t,m) - Y_i(t,0) \mid X_i = x]$$

- With sequential ignorability and no intermediate interaction,

$$\mathbb{E}[Y_i - \gamma(t,M_i,x) \mid T_i = t, X_i = x] = \mathbb{E}[Y_i(t,0) \mid X_i = x]$$

- Therefore,

$$\begin{aligned}\mathbb{E}[Y_i(t,0) - Y_i(0,0) \mid X_i = x] &= \mathbb{E}[Y_i - \gamma(t,M_i,x) \mid T_i = t, X_i = x] \\ &\quad - \mathbb{E}[Y_i - \gamma(0,M_i,x) \mid T_i = 0, X_i = x]\end{aligned}$$

Structural Nested Mean Model (SNMM) (2)

- Then, blip function is identified under sequential ignorability as

$$\gamma(t, m, x) = \mathbb{E}[Y_i \mid T_i = t, M_i = m, X_i = x, Z_i = z] - \mathbb{E}[Y_i \mid T_i = t, M_i = 0, X_i = x, Z_i = z]$$

- **Estimation Procedure (Sequential G-Estimation)**:
  - STEP1: Run first stage regression

$$\mathbb{E}[Y_i \mid T_i, M_i, X_i, Z_i] = \alpha_0 + \alpha_1 T_i + \alpha_2 M_i + \alpha_3 X_i + \alpha_4 Z_i$$

    and obtain blip function as $\gamma(t, m, x) = \alpha_2 m$
    - Note that you can make blip function more complex (e.g., w/ interaction)
  - STEP2: Using the blip function, estimate

$$\mathbb{E}[Y_i - \gamma(t, M_i, x) \mid T_i, X_i] = \beta_0 + \beta_1 T_i + \beta_2 X_i$$

    and interpret β_1 as CDE.

- **Intuition**: SNMM avoids collider-bias by not conditioning on M_i and Z_i directly
  - Blip function models the effect of switching mediator values

Marginal Structural Model (MSM)

- Problem of SNMM: need to correctly specify the blip function
  - Misspecification of blip function can lead to bias

- **Marginal Structural Model**[1]: Assume that the marginal mean of potential outcome is written as

$$\mathbb{E}[Y_i(t, m)] = g(t, m; \beta)$$

  - Importantly, it is marginal, meaning that we do not condition on M_i and $Z_i \rightarrow$ no collider bias

- We use weighting to estimate marginal structural model
  - Intuitively, at each point (T_i and M_i separately), we use weight so that we can recover potential outcome without directly conditioning on post-treatment at each time point

[1] For both MSM and SNMM, "Structural" means modeling the relationship between potential outcome and treatment directly. As a result, the parameter of structural model is not a nuisance parameter.

Natural Direct Effect: Assumptions

- Identification Assumption

$$\{Y_i(t, m), M_i(t')\} \perp\!\!\!\perp T_i \mid X_i$$
$$Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid X_i = x, T_i \quad \text{(Cross-world Counterfactual)}$$

- Note that this is stronger than CDE
  - Assume no intermediate variable Z_i
  - Cross-world Counterfactual

- Look at Question 2 of Review Question for module 9

Natural Direct Effect: Identification

$$
\begin{aligned}
&\mathbb{E}[Y(t, M(t'))|X] \\
&= \sum_m \mathbb{E}[Y(t,m) \mid X, M(t') = m]\mathbb{P}(M(t') = m \mid X) \quad (\because \text{L.I.E.}) \\
&= \sum_m \mathbb{E}[Y(t,m) \mid X, M(t') = m, T = t]\mathbb{P}(M(t') = m \mid X) \\
&= \sum_m \mathbb{E}[Y(t,m) \mid X, T = t]\mathbb{P}(M(t') = m \mid X, T = t') \\
&= \sum_m \mathbb{E}[Y(t,m) \mid X, T = t, M(t) = m]\mathbb{P}(M(t') = m \mid X, T = t') \\
&= \sum_m \mathbb{E}[Y \mid X, T = t, M = m]\mathbb{P}(M = m \mid X, T = t')
\end{aligned}
$$