# Policy Evaluation and Structural Identification

Statistics · Module 11 — Causality

Node 11.6

## Position in Knowledge Graph

System: Statistics. Structural Domain: Causality (Module 11). Node 11.6.

This capstone node extends causal inference from estimating treatment effects to evaluating decision rules (policies). It develops inverse propensity weighting for policy evaluation, the doubly robust AIPW estimator, and the semiparametric efficiency bound for the ATE. The connection to structural equation models links causal inference to mechanism design and reinforcement learning. It synthesizes the entire causality module: potential outcomes (Node 11.1), DAGs (Node 11.2), identification (Node 11.3), experiments (Node 11.4), and observational methods (Node 11.5) into a unified framework for evaluating interventions in complex systems.

---

## Formal Definition

A **policy** (or treatment rule) is a mapping $\pi : \mathcal{X} \to \{0, 1\}$ (deterministic) or $\pi : \mathcal{X} \to [0, 1]$ (stochastic) that assigns treatment based on observed covariates.

The **value** of policy $\pi$ is

$$V(\pi) = \mathbb{E}[Y(\pi(X))] = \mathbb{E}[\pi(X)\, Y(1) + (1 - \pi(X))\, Y(0)].$$

The **optimal policy** maximizes the value:

$$\pi^* = \arg\max_{\pi} V(\pi).$$

Under ignorability and overlap, the value is identified:

$$V(\pi) = \mathbb{E}[\mathbb{E}[Y \mid T = 1, X]\, \pi(X) + \mathbb{E}[Y \mid T = 0, X]\, (1 - \pi(X))]\,.$$

---

## Intuition

Treatment effect estimation asks: what is the average effect of treatment? Policy evaluation asks a more refined question: who should be treated? The optimal policy treats unit $i$ if and only if $\tau(X_i) = \mathbb{E}[Y(1) - Y(0) \mid X = X_i] > 0$. This connects causal inference to statistical decision theory: the policy is a decision rule, and the value function is the expected utility. The challenge is estimating the value of a proposed policy from data collected under a different (possibly non-randomized) treatment assignment.

---

## Structural Role

**Dependencies:** Builds on observational treatment estimation (Node 11.5) for the estimators, regression (Node 8.3) for outcome modeling, and decision theory (Node 6.3) for the policy optimization framework. **Enables:** This is the terminal node of Module 11 and the statistics course, connecting to reinforcement learning in AI and mechanism design in economics.

Policy evaluation is the bridge between causal inference and decision-making. It transforms estimates of causal effects into actionable rules, closing the loop from data to intervention.

---

## Mathematical Development

### Inverse Propensity Weighted Policy Evaluation

Given data $\{(X_i, T_i, Y_i)\}_{i=1}^n$ collected under a **behavior policy** $b(x) = \mathbb{P}(T = 1 \mid X = x)$, the value of a target policy $\pi$ is estimated by reweighting:

$$\hat{V}^{\mathrm{IPW}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}(T_i = \pi(X_i))}{T_i\, b(X_i) + (1 - T_i)(1 - b(X_i))}\, Y_i.$$

For a stochastic policy:

$$\hat{V}^{\mathrm{IPW}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(X_i)}{b(X_i)}\, T_i\, Y_i + \frac{1 - \pi(X_i)}{1 - b(X_i)}\, (1 - T_i)\, Y_i.$$

This is unbiased for $V(\pi)$ when $b(x)$ is known, and consistent when $b(x)$ is estimated consistently. The variance is high when $\pi$ and $b$ differ substantially (large importance weights).

### Outcome Regression Policy Evaluation

Directly model the outcome:

$$\hat{V}^{\mathrm{OR}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \left[\pi(X_i)\, \hat{\mu}_1(X_i) + (1 - \pi(X_i))\, \hat{\mu}_0(X_i)\right].$$

Consistent under correct specification of $\mu_t(x)$. No importance weights, so lower variance, but sensitive to model misspecification.

### Doubly Robust Policy Evaluation (AIPW)

The **augmented IPW (AIPW)** estimator for policy value is

$$\hat{V}^{\mathrm{DR}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \left[\pi(X_i) \left(\hat{\mu}_1(X_i) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{b}(X_i)}\right) + (1 - \pi(X_i)) \left(\hat{\mu}_0(X_i) + \frac{(1 - T_i)(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{b}(X_i)}\right)\right].$$

**Properties:** 1. Doubly robust: consistent if either $\hat{b}$ or $\hat{\mu}_t$ is correctly specified. 2. Achieves the semiparametric efficiency bound when both are correctly specified. 3. Allows the use of flexible/nonparametric estimators for nuisance functions while maintaining $\sqrt{n}$-convergence for the policy value (via cross-fitting).

### Semiparametric Efficiency Bound for the ATE

The ATE $\tau = V(\pi \equiv 1) - V(\pi \equiv 0) = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$ is a special case of policy evaluation. The **efficient influence function** for $\tau$ under the nonparametric model is

$$\phi(Y, T, X; \tau) = \frac{T(Y - \mu_1(X))}{e(X)} - \frac{(1-T)(Y - \mu_0(X))}{1 - e(X)} + \mu_1(X) - \mu_0(X) - \tau.$$

The **semiparametric efficiency bound** is

$$\mathrm{Var}(\phi) = \mathbb{E}\left[\frac{\mathrm{Var}(Y \mid T = 1, X)}{e(X)} + \frac{\mathrm{Var}(Y \mid T = 0, X)}{1 - e(X)} + (\tau(X) - \tau)^2\right].$$

**Derivation.** The influence function is derived from the tangent space of the nonparametric model. The nuisance tangent space is spanned by functions of $(T, X)$; the efficient influence function is the projection of the pathwise derivative of $\tau$ onto the orthocomplement of this space in $L^2(P)$.

The three terms in the efficiency bound have clear interpretations: - $\mathrm{Var}(Y \mid T = 1, X)/e(X)$: noise in treated outcomes, amplified when few units are treated. - $\mathrm{Var}(Y \mid T = 0, X)/(1 - e(X))$: noise in control outcomes. - $(\tau(X) - \tau)^2$: treatment effect heterogeneity.

### Cross-Fitting and Machine Learning

When $\hat{\mu}_t$ and $\hat{b}$ are estimated by flexible methods (random forests, neural networks), the AIPW estimator requires **cross-fitting** (sample splitting) to avoid Donsker conditions:

1. Split data into $K$ folds.
2. For each fold $k$, estimate nuisance functions $\hat{\mu}_t^{(-k)}, \hat{b}^{(-k)}$ on the complement.
3. Evaluate the AIPW estimand on fold $k$ using the cross-fitted nuisance estimates.
4. Average across folds.

This **DML (Double/Debiased Machine Learning)** procedure (Chernozhukov et al., 2018) achieves $\sqrt{n}$-consistent, asymptotically normal estimation of $\tau$ even when nuisance functions converge at slower rates (e.g., $n^{-1/4}$).

### Optimal Policy Learning

The optimal deterministic policy is

$$\pi^*(x) = \mathbf{1}(\tau(x) > 0).$$

**Estimation approaches:**

1. **Indirect method:** Estimate $\hat{\tau}(x)$ (e.g., via CATE estimation using causal forests or metalearners), then set $\hat{\pi}(x) = \mathbf{1}(\hat{\tau}(x) > 0)$.
2. **Direct method (policy search):** Maximize the estimated policy value over a class of policies $\Pi$:

$$\hat{\pi} = \arg\max_{\pi \in \Pi} \hat{V}^{\mathrm{DR}}(\pi).$$

The policy regret $V(\pi^*) - V(\hat{\pi})$ converges at rate $O(n^{-1/2})$ for finite-dimensional policy classes and $O(n^{-1/(2+d)})$ for $d$-dimensional policy classes under smoothness.

### Structural Equations and Policy Counterfactuals

In the **structural equation model (SEM)**:

$$Y = f(T, X, U), \quad T = g(Z, X, V),$$

where $U, V$ are unobserved. A policy $\pi$ replaces the treatment equation with $T = \pi(X)$. The policy counterfactual is

$$Y^{\pi} = f(\pi(X), X, U),$$

and the policy value is $V(\pi) = \mathbb{E}[f(\pi(X), X, U)]$.

When the SEM is identified (via instruments, panel structure, or functional form), policy evaluation reduces to computing expectations under the structural functions. This connects to mechanism design: the structural equations describe how the system responds to interventions, and the policy is the designed intervention.

### Connection to Reinforcement Learning

In sequential settings, the policy maps state histories to actions: $\pi : \mathcal{H}_t \to \mathcal{A}$. The value function is

$$V(\pi) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{T} \gamma^t R_t\right],$$

where $R_t$ is the reward at time $t$ and $\gamma$ is a discount factor. **Off-policy evaluation** estimates $V(\pi)$ from data collected under a behavior policy $b$, using:

- **Importance sampling:** $\hat{V}^{\mathrm{IS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \prod_{t=0}^{T} \frac{\pi(A_{it}|H_{it})}{b(A_{it}|H_{it})} \sum_t \gamma^t R_{it}$. Variance grows exponentially with horizon $T$.
- **Doubly robust (sequential):** Analogous to the AIPW estimator, using fitted Q-functions as the outcome model and importance ratios as the propensity weights.

This is the sequential generalization of the single-stage causal inference of Nodes 11.1–11.5.

---

## Worked Examples

### Example 1: Policy Evaluation via AIPW

Data: $n = 500$, binary treatment. Behavior policy: $b(x) = \mathrm{logit}^{-1}(0.5 + x_1)$. Estimated outcome models: $\hat{\mu}_1(x) = 3 + 2x_1$, $\hat{\mu}_0(x) = 3 + 0.5x_1$.

Target policy: treat if $x_1 > 0$, i.e., $\pi(x) = \mathbf{1}(x_1 > 0)$.

For a specific observation with $X_1 = 0.8$, $T = 1$, $Y = 5.2$:

$\hat{b}(X) = \mathrm{logit}^{-1}(0.5 + 0.8) = \mathrm{logit}^{-1}(1.3) = 0.786$. $\hat{\mu}_1(X) = 3 + 1.6 = 4.6$, $\hat{\mu}_0(X) = 3 + 0.4 = 3.4$. $\pi(X) = 1$ (since $x_1 > 0$).

AIPW contribution: $1 \cdot (4.6 + \frac{1 \cdot (5.2 - 4.6)}{0.786}) + 0 \cdot (\cdots) = 4.6 + 0.763 = 5.363$.

Averaging such terms over all 500 observations gives $\hat{V}^{\mathrm{DR}}(\pi)$.

**Example 2: Optimal Policy via CATE Estimation**

Estimated CATE: $\hat{\tau}(x) = 1.5 - 3x_1 + 2x_2$. The optimal policy treats when $\hat{\tau}(x) > 0$, i.e., $1.5 - 3x_1 + 2x_2 > 0$, or $x_2 > 1.5x_1 - 0.75$.

For patient profiles: - $(x_1, x_2) = (0.2, 0.5)$: $\hat{\tau} = 1.5 - 0.6 + 1.0 = 1.9 > 0$. Treat. - $(x_1, x_2) = (1.0, 0.0)$: $\hat{\tau} = 1.5 - 3.0 + 0.0 = -1.5 < 0$. Do not treat. - $(x_1, x_2) = (0.5, 0.3)$: $\hat{\tau} = 1.5 - 1.5 + 0.6 = 0.6 > 0$. Treat.

The policy $\hat{\pi}$ personalizes treatment based on predicted benefit, targeting treatment to those who gain most.

---

## Cross-Links

**AI.** Offline reinforcement learning is the sequential generalization of policy evaluation. Fitted Q-evaluation, importance sampling, and doubly robust methods for sequential decisions all extend the single-stage methods of this node. Contextual bandits are the intermediate case between single-stage and full sequential problems.

**Economics.** Mechanism design asks: given agents' strategic responses (the structural equations), what policy maximizes social welfare? This is the game-theoretic generalization of optimal policy learning. The welfare analysis of tax policies, subsidies, and regulations uses structural models to evaluate counterfactual policies.

**Linear Algebra.** The semiparametric efficiency bound involves the inverse of the propensity score as a weighting matrix. In the linear model, the efficient estimator for the ATE reduces to a weighted least squares problem with weights $w_i = 1/[e(X_i)(1 - e(X_i))]$, and the efficiency bound is the trace of the inverse of the weighted Fisher information matrix.

**Quantum Mechanics.** Quantum control theory optimizes a policy (a sequence of pulses or measurements) to drive a quantum system to a target state. The value function is the fidelity of the final state, and the "treatment effect" is the change in fidelity from applying vs. not applying a control pulse. Doubly robust estimation has analogues in quantum process tomography with adaptive measurement designs.

---

## Structural Compression

**Invariant**

Each design (IPW, AIPW, structural models) provides a consistent estimator of the policy value $V(\pi)$ under distinct assumptions; the AIPW estimator achieves the semiparametric efficiency bound and is doubly robust.

**Mechanism**

IPW reweights observations by the ratio of target-to-behavior policy probabilities; the AIPW estimator adds an outcome-regression correction that reduces variance and provides double robustness. The efficient influence function characterizes the information-theoretic lower bound for estimating the policy value. Cross-fitting enables the use of flexible machine learning estimators for nuisance functions while preserving $\sqrt{n}$-inference for the policy value.

**Cross-System Echo**

Policy evaluation is the bridge between causal inference and decision theory. In AI, off-policy evaluation in RL is the sequential generalization, with importance sampling replacing IPW and fitted Q-functions replacing outcome regression. In economics, structural estimation of policy counterfactuals is the mechanism-based approach, where the "policy" is a change in the rules of a game and the "value" is social welfare. The doubly robust principle—combine a model of the world with a correction for model error—is universal across these domains.

---

## Exercises

1. Derive the IPW estimator for the policy value $V(\pi)$ starting from $V(\pi) = \mathbb{E}[\pi(X)Y(1) + (1-\pi(X))Y(0)]$ and using ignorability to replace potential outcomes with observables.
2. Show that the AIPW estimator for policy value is doubly robust: consistent if either the propensity score or the outcome model is correctly specified.
3. Derive the efficient influence function for the ATE $\tau = \mathbb{E}[Y(1) - Y(0)]$ by projecting the pathwise derivative onto the orthocomplement of the nuisance tangent space. Verify that the AIPW estimator is the sample analogue.
4. For a binary covariate $X \in \{0, 1\}$ with $\mathbb{P}(X = 1) = 0.6$, $\tau(0) = -1$, $\tau(1) = 3$, compute the value of the optimal policy $\pi^*(x) = \mathbf{1}(\tau(x) > 0)$ and compare it to the value of treating everyone and treating no one.
5. Explain the DML (cross-fitting) procedure for estimating the ATE with machine-learning nuisance estimators. Why is sample splitting necessary? What goes wrong without it?
6. Compare the variance of the IPW, OR, and AIPW estimators for the ATE in a setting where the propensity score is extreme ($e(x)$ near 0 or 1 for some $x$). Which estimator is most robust to this practical challenge?
7. Argue, structurally, why policy evaluation subsumes treatment effect estimation as a special case, and explain how the connection to reinforcement learning generalizes the single-decision framework of causal inference to sequential decision problems with delayed rewards.

---

*Statistics · Module 11 — Causality · Node 11.6*
**Concepts:** cost-function, conditional-probability, prior-posterior
**Prerequisites:** 11.5, 8.3, 6.3

*hbar.university — own.your.cognition*