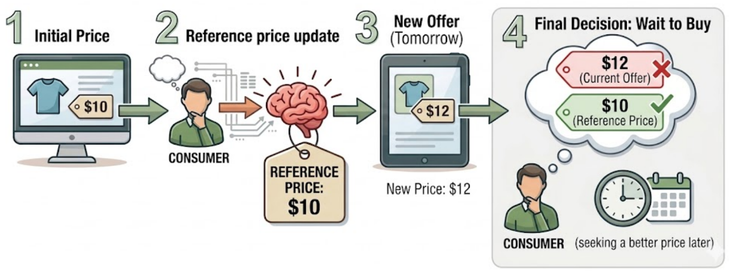
Estimating Dynamic Marginal Policy Effects under Sequential Unconfoundedness

I-han Lai[1] Stefan Wager[2]

[1]MS&E, Stanford University. [2]Stanford Graduate School of Business.

Motivation



Abstract

We develop methods for estimating how infinitesimal policy changes affect long-term outcomes in dynamic systems. We show that dynamic marginal policy effects (MPEs) can be identified via tractable reduced-form expressions, and can be estimated under a general sequential unconfoundedness assumption.

We also propose a **doubly robust** estimator for dynamic MPEs. Our approach does not require observing full dynamic state information (as is typically assumed for off-policy evaluation in Markov decision processes), and does not incur an exponential curse of horizon (as is typical in non-Markovian off-policy evaluation).

We demonstrate practicality and robustness of our approach in a number of simulations, including one motivated by a dynamic pricing application where people use past prices to form a reference level for current prices.

Identification

Notation:

- $t \in \{1, \dots, T\}$: Time period in a finite-horizon sequential decision problem.
- $X_t \in \mathcal{X}_t$: Covariates observed at the beginning of period t.
- $A_t \in \mathcal{A}_t$: Action taken in period t.
- $R_t \in \mathbb{R}$: Reward accrued in period t.
- $S_t := (X_1, A_1, R_1, \dots, X_t)$: The accumulated information observed before A_t is chosen.
- $\pi_t(da_t \mid s_t)$: The status-quo policy given the observed history s_t.

Assumptions & Factorization:

- (A1) Consistency:
$$(X_{t+1}, R_t) = (X_{t+1}(A_{1:t}), R_t(A_{1:t}))$$
- (A2) Sequential unconfoundedness:
$$A_t \perp\!\!\!\perp \left\{X_{k+1}(A_{1:t-1}, a_t, \dots, a_k),\ R_k(A_{1:t-1}, a_t, \dots, a_k)\right\} \mid S_t.$$

Then, the status quo data-generating distribution factors as:

$$P(dx_{1:T+1}, da_{1:T}, dr_{1:T}) = P(dx_1) \prod_{t=1}^{T} \pi_t(da_t \mid s_t)\, P(dx_{t+1}, dr_t \mid s_t, a_t),$$

Identification

Theorem 1 (Generalized policy-gradient representation). For any supported perturbation path $\{\pi_\varepsilon\}$ around baseline π, define the stage-t policy direction

$$(\mathcal{L}_t \pi_t)(a \mid s) := \frac{d}{d\varepsilon} \pi_{t,\varepsilon}(a \mid s)\Big|_{\varepsilon=0}.$$

Under A1–A2 and regularity,

$$\Theta = \sum_{t=1}^{T} G_t(q_t), \qquad G_t(q) = \gamma^{t-1}\, \mathbb{E}\left[\int_{\mathcal{A}_t} q(S_t, a)\, (\mathcal{L}_t \pi_t)(a \mid S_t)\, d\lambda(a)\right].$$

If the direction score

$$H_t(s, a) = \frac{(\mathcal{L}_t \pi_t)(a \mid s)}{\pi_t(a \mid s)}$$

exists on the baseline support, then equivalently

$$G_t(q) = \gamma^{t-1}\, \mathbb{E}[H_t(S_t, A_t)\, q(S_t, A_t)].$$

Key point: $G_t(q_t)$ is the *stagewise welfare sensitivity*: it measures how a local change in the policy at time t changes total discounted welfare, with all downstream effects summarized by $q_t(s, a)$.

Estimation

Convolution perturbation path. At stage t, draw a baseline action $A_{t,0} \sim \pi_t(\cdot \mid s)$ and perturb it through a *known* Markov kernel

$$M_{t,\varepsilon}(da \mid a_0, s), \qquad M_{t,0}(da \mid a_0, s) = \delta_{a_0}(da).$$

The perturbed policy is the pushforward

$$\pi_{t,\varepsilon}(da \mid s) = \int M_{t,\varepsilon}(da \mid a_0, s)\, \pi_t(da_0 \mid s).$$

Its adjoint acts on the value side:

$$(\mathcal{L}_t^\dagger q)(s, a_0) := \frac{d}{d\varepsilon} \int q(s, a)\, M_{t,\varepsilon}(da \mid a_0, s)\Big|_{\varepsilon=0}, \qquad G_t(q) = \gamma^{t-1} \mathbb{E}\left[(\mathcal{L}_t^\dagger q)(S_t, A_t)\right].$$

Three estimators.

$$\widehat{\Theta}_{\mathrm{Direct}} = \sum_{t=1}^{T} \gamma^{t-1}\, \mathbb{P}_n\left[(\mathcal{L}_t^\dagger \widehat{q}_t)(S_t, A_t)\right], \qquad \widehat{\Theta}_{\mathrm{SRW}} = \sum_{t=1}^{T} \gamma^{t-1}\, \mathbb{P}_n\left[\widehat{H}_t(S_t, A_t) \Gamma_t\right].$$

$$\psi_t(Z; q, H) = \gamma^{t-1}(\mathcal{L}_t^\dagger q)(S_t, A_t) + \gamma^{t-1} H(S_t, A_t)\{\Gamma_t - q(S_t, A_t)\},$$

$$\widehat{\Theta}_{\mathrm{ASRW}} = \sum_{t=1}^{T} \mathbb{P}_n\left[\psi_t\big(Z; \widehat{q}_t, \widehat{H}_t\big)\right].$$

Orthogonality. For oracle nuisances $(q_t^\star, H_t^\star)$,

$$\mathbb{E}[\psi_t(Z; q, H)] - G_t(q_t^\star) = \gamma^{t-1}\, \mathbb{E}[(H_t^\star - H)(q - q_t^\star)].$$

Theorem 2 (CLT for ASRW). Let $\widehat{\Theta}^\star_{\mathrm{ASRW}}$ denote the oracle ASRW estimator that uses the true nuisances. Under cross-fitting, bounded rewards, and mild regularity,

$$\sqrt{n}\big(\widehat{\Theta}_{\mathrm{ASRW}} - \widehat{\Theta}^\star_{\mathrm{ASRW}}\big) \to_p 0, \qquad \sqrt{n}\big(\widehat{\Theta}^\star_{\mathrm{ASRW}} - \Theta\big) \Rightarrow \mathcal{N}(0, \sigma_\star^2).$$

Therefore

$$\sqrt{n}\big(\widehat{\Theta}_{\mathrm{ASRW}} - \Theta\big) \Rightarrow \mathcal{N}(0, \sigma_\star^2),$$

provided, uniformly over t,

$$\|\widehat{q}_t - q_t^\star\|_2\, \|\widehat{H}_t - H_t^\star\|_2 = o_p(n^{-1/2}), \qquad \|(\mathcal{L}_t^\dagger \widehat{q}_t) - (\mathcal{L}_t^\dagger q_t^\star)\|_2 = o_p(1).$$

Estimation (continued)

Algorithm 1 Cross-fitted ASRW estimator

Input: i.i.d. trajectories $\{Z_i\}_{i=1}^n$, number of folds K, discount factor γ, and known stagewise operators $\{\mathcal{L}_t\}_{t=1}^T$

Output: $\widehat{\Theta}$

Partition $\{1, \dots, n\}$ into disjoint scoring folds $I_1, \dots, I_K$.

for $k = 1, \dots, K$ **do**

  Using the training sample $\{Z_i : i \notin I_k\}$, estimate nuisance functions $\{\widehat{q}_t^{(-k)}, \widehat{H}_t^{(-k)}\}_{t=1}^T$.

  for $i \in I_k$ **do**

    Compute $\Gamma_{i,t} := \sum_{u=t}^T \gamma^{u-t} R_{i,u}$ for $t = 1, \dots, T$.

    for $t = 1, \dots, T$ **do**

      Compute the stagewise contribution

$$\widehat{\psi}_{i,t}^{(-k)} := \gamma^{t-1}(\mathcal{L}_t \widehat{q}_t^{(-k)})(S_{i,t}, A_{i,t}) + \gamma^{t-1} \widehat{H}_t^{(-k)}(S_{i,t}, A_{i,t})\{\Gamma_{i,t} - \widehat{q}_t^{(-k)}(S_{i,t}, A_{i,t})\}.$$

$$\widehat{\Theta} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \sum_{t=1}^{T} \widehat{\psi}_{i,t}^{(-k)}.$$

Dynamic Pricing Experiment

Motivation. Customers compare today's price to an internal *reference price*. So a higher price can reduce current demand, but it can also raise the future reference level and change later demand.

Why this example is hard. The analyst observes recent prices, revenues, and seasonality, but *not* customer heterogeneity or the latent reference state. Thus the problem is dynamic and partially observed, while sequential unconfoundedness still holds because the behavior policy depends only on observed history.

Baseline policy and target. Prices are generated from a clipped Gaussian policy,

$$A_{i,t} = \mathrm{clip}\big(A_{i,t}^{\mathrm{raw}}, [p_{\min}, p_{\max}]\big), \qquad A_{i,t}^{\mathrm{raw}} \mid S_{i,t} \sim N\big(\mu_t(X_{i,t}), \sigma^2_{\mathrm{price}}\big).$$

We estimate the MPE of a small upward location shift,

$$A_{i,t,\varepsilon}^{\mathrm{raw}} \mid S_{i,t} \sim N\big(\mu_t(X_{i,t}) + \varepsilon, \sigma^2_{\mathrm{price}}\big), \qquad \mathcal{L}_t^\dagger q = \partial_a q.$$

Estimators compared. Direct plug-in, SRW, ASRW, and ASRW with oracle score (all with 5-fold cross-fitting).

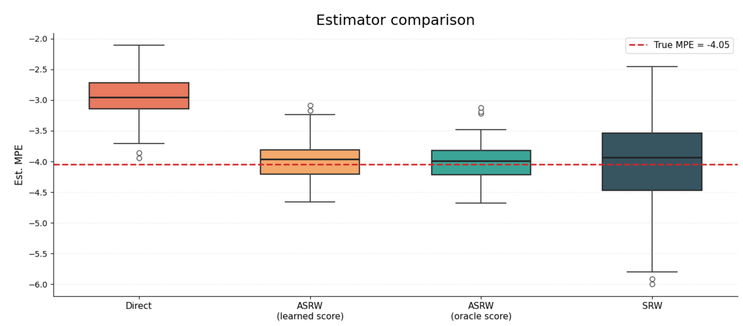


Figure 1. Sampling distributions of the four MPE estimators across replications in the dynamic pricing simulator with hidden customer heterogeneity and reference effects.