Semi-Parametric Efficient Policy Learning with Continuous Actions

Vasilis Syrgkanis

Microsoft Research, New England

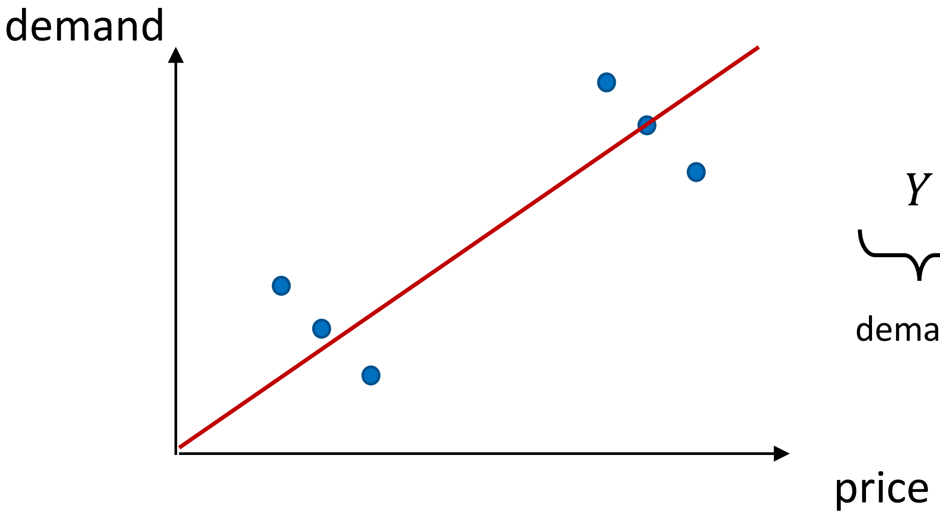
Joint with Mert Demirer (MIT), Victor Chernozhukov (MIT), Greg Lewis (MSR)

Optimal Pricing from Historical/Observational Data

We are given historical data of demand and price from a company

Goal: Find the optimal price point based on the data

Approach: Estimate demand function $Q(p)$ then optimize revenue $p \cdot Q(p)$



$$\underbrace{Y}_{\text{demand}} = \overbrace{\underbrace{\alpha_0}_{\text{elasticity}} \cdot \underbrace{P}_{\text{price}}}^{Q(p)} + \underbrace{\epsilon}_{\text{noise}}$$

Conclusion: Increasing price increases demand!

Problem: Demand increases in winter and price anticipates demand

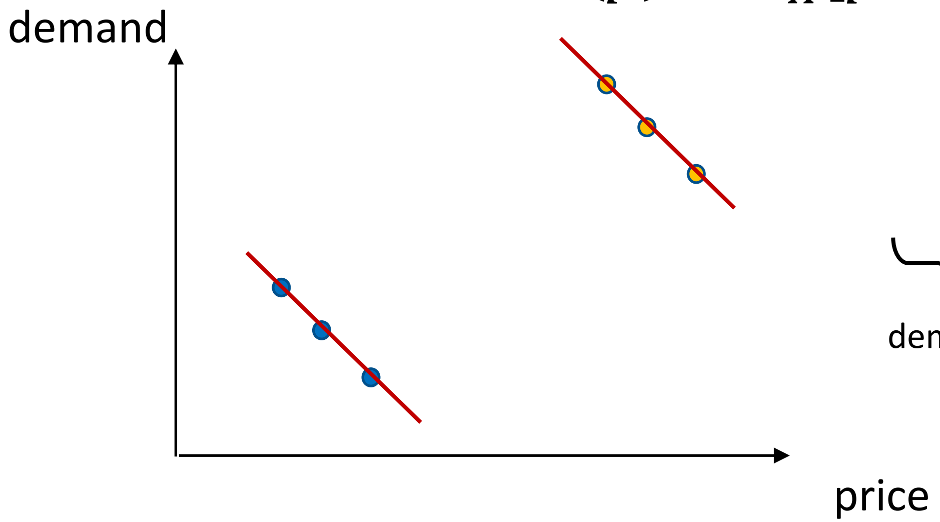
Optimal Pricing from Observational Data

We are given historical data of demand and price from a company

Goal: Find a new optimal price point based on the data

Approach: Estimate demand function $Q(p, x)$ then optimize expected revenue

$$V(p) = E_X[p \cdot Q(p, X)]$$



$$\underbrace{Y}_{\text{demand}} = \overbrace{\underbrace{\alpha_0 \cdot}_{\text{elasticity}} \underbrace{P}_{\text{price}} + \underbrace{\beta_0(X)}_{\text{confounders}}}^{Q(P,X)} + \underbrace{\epsilon}_{\text{noise}}$$

Idea: Introduce confounder (the season) into regression

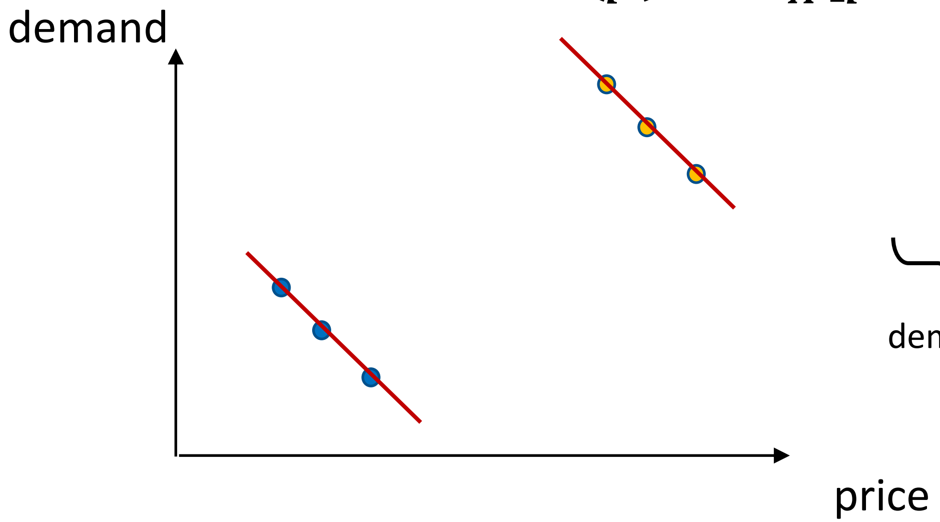
Optimal Pricing from Observational Data

We are given historical data of demand and price from a company

Goal: Find a new optimal price point based on the data

Approach: Estimate demand function $Q(p, x)$ then optimize expected revenue

$$V(p) = E_X[p \cdot Q(p, X)]$$



$$\underbrace{Y}_{\text{demand}} = \overbrace{\underbrace{\alpha_0}_{\text{elasticity}} \cdot \underbrace{P}_{\text{price}} + \underbrace{\beta_0(X)}_{\text{confounders}}}^{Q(P,X)} + \underbrace{\epsilon}_{\text{noise}}$$

Problem: What if there are 100s or 1000s of potential confounders?

Can we get $\sqrt{n}$-rates for the optimal price, even if β_0 is not $\sqrt{n}$ consistent?

Personalized/Contextual

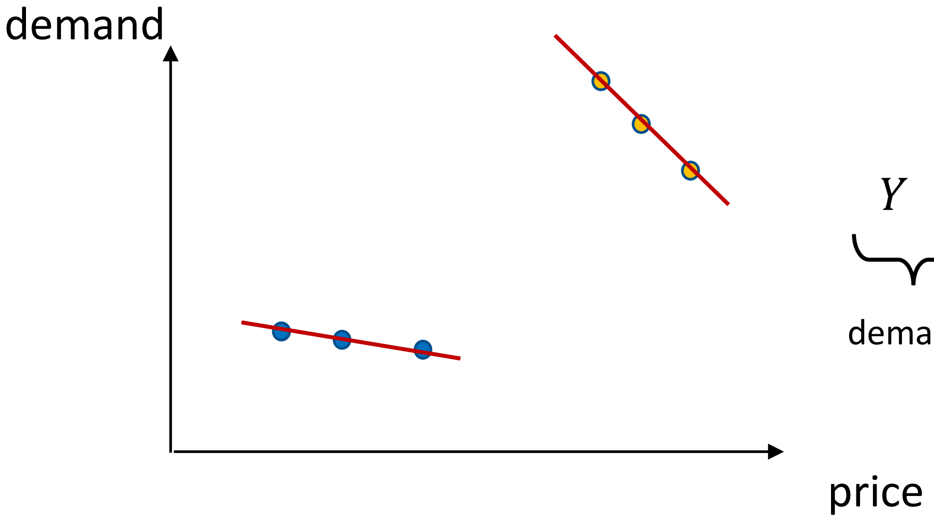
Optimal Pricing from Observational Data

We are given historical data of demand and price from a company

Goal: Find a new optimal price **policy** based on the data

Approach: Estimate demand function $Q(p, x)$ then optimize policy revenue

$$V(\pi) = E[\pi(X) \cdot Q(\pi(X), X)]$$



$$\underbrace{Y}_{\text{demand}} = \overbrace{\underbrace{\alpha_0(X)}_{\substack{\text{contextual}\\ \text{elasticity}}} \cdot \underbrace{T}_{\text{price}} + \underbrace{\beta_0(X)}_{\substack{\text{baseline}\\ \text{demand}}}}^{Q(P,X)} + \underbrace{\epsilon}_{\text{noise}}$$

Can we get $\sqrt{n}$-rates for the optimal policy if policy space is simple, even if β_0 and α_0 is not $\sqrt{n}$ consistent?

Policy Optimization from Observational Data with Continuous Actions

- Given observational data with n samples of triplets (Y, T, X) of outcomes Y, actions T and features X
- Assume conditional exogeneity: making an intervention and setting $T = \mathrm{t}$, at $X = x$ yields value
$$V(t, x) = E[Y | T = t, X = x]$$
- Given a treatment policy space Π: find a policy $\hat{\pi}$ with good regret
$$R(\Pi, n) = \underbrace{\sup_{\pi \in \Pi} E_X[V(\pi(X), X)]}_{\text{Expected Value of Optimal Policy}} - \underbrace{E_X[V(\hat{\pi}(X), X)]}_{\text{Value of Chosen Policy}}$$

Results extend if we optimize $\rho(t, x) \cdot V(t, x) + c(t, x)$ for any known functions ρ, c or when Y is vector and optimize $\rho(t, x)' V(t, x) + c(t, x)$

Main Assumption

Linearity in known feature space:

$$V(t,x) = \langle \theta_0(x), \phi(t,x) \rangle$$

for some known feature vector function ϕ but unknown functions θ_0

Example. pricing; Y=demand, T=price

$$V(t,x) = \theta_0(x) \cdot t + g_0(x)$$

and we optimize $t \cdot V(t,x)$

Example. investment allocation; Y=ROI, T=vector of investments

$$V(t,x) = \langle \theta_0(x), t \rangle$$

and we optimize $V(t,x) - c(t)$ for some known investment cost c

Mis-specification

Even if assumption is violated, we achieve regret wrt to best linear projection $V_p(t,x) = \left\langle \theta_p(x), \phi(t,x) \right\rangle$, where

$$\theta_p(x) = \operatorname{argmin}_\theta E[\,(V(t,x) - \langle \theta, \phi(t,x) \rangle)^2 \mid x\,]$$

Main Question

Can we get $\sqrt{n}$-rates for the optimal policy if policy space is simple, even if θ_0 is not $\sqrt{n}$ consistent?

Attempt 1: Direct Approach

- Estimate $\hat{\theta}$ by regressing $Y \sim T, X$ on one half of the data

- Optimize on the half part:

$$\hat{\pi} = \sup_{\pi \in \Pi} E_n[\langle \hat{\theta}(X), \phi(\pi(X), X)]$$

- Problem: estimate of policy value heavily depends on estimate of $\hat{\theta}$
- If estimate of $\hat{\theta}$ has RMSE of ϵ_n, then regret incurs an error of ϵ_n

Contribution 1: A Doubly Robust Method for Continuous Actions

- Estimate $\hat{\theta}$ by regressing $Y \sim T, X$ on one half of the data
- Estimate conditional covariance matrix on one half of the data

$$\hat{\Sigma}(x) = E[\,\phi(T,X)\phi(T,X)' \mid X = x\,]$$

- One the other half: construct a doubly robust estimate of the value coefficients:

$$\theta_{DR}(X) = \underbrace{\hat{\theta}(X)}_{\text{Direct Regression Estimate}} + \underbrace{\hat{\Sigma}^{-1}(X)\,\phi(T,X)\left(Y - \langle\hat{\theta}(X), \phi(T,X)\rangle\right)}_{\text{Doubly Robust Correction}}$$

- Optimize:

$$\hat{\pi} = \sup_{\pi\in\Pi} E_n[\langle\theta_{DR}(X), \phi(\pi(X), X)\rangle]$$

Double Robustness

$$\theta_{DR}(X) = \hat{\theta}(X) + \hat{\Sigma}^{-1}(X)\, \phi(T,X)\, \left(Y - \langle \hat{\theta}(X), \phi(T,X) \rangle\right)$$

If $\hat{\theta}$ is correct, then

$$E\left[\, \phi(T,X)\, \left(Y - \langle \hat{\theta}(X), \phi(T,X) \rangle\right) \mid X \,\right] = 0$$

And

$$\mathrm{E}[\, \theta_{DR}(X) \mid X \,] = \hat{\theta}(X) = \theta_0(X)$$

If $\hat{\Sigma}$ is correct, then

$$\hat{\Sigma}^{-1}(X) \cdot E\left[\phi(T,X)\, \langle \hat{\theta}(X), \phi(T,X) \rangle \mid X\right] = \hat{\theta}(X)$$

and

$$\mathrm{E}[\, \theta_{DR}(X) \mid X \,] = \hat{\Sigma}^{-1}(X) E[\phi(T,X)\, \mathrm{E}[\, Y \mid T, X \,] \mid X] = \Sigma_0^{-1}(X) E[\, \phi(T,X) \phi(T,X)' \mid X \,] \theta_0(X) = \theta_0(X)$$

Contribution 2: Semi-Parametric Efficiency

Theorem. If we let

$$\theta_{DR}^{0}(X) = \theta_0(X) + \Sigma_0^{-1}(X)\, \phi(T, X)\, (Y - \langle \theta_0(X), \phi(T, X) \rangle)$$

For any policy π the variance of the quantity off policy estimate

$$E_n[\langle \theta_{DR}^{0}(X), \phi(\pi(X), X)]$$

is the best variance statistically achievable, without making further assumptions on the functions θ_0; aka semi-parametric efficiency bound

* This holds either when the errors in the Y regression are homoscedastic, or when the model is mis-specified and θ_0 is the best linear projection

Contribution 3a: Robust Regret

Theorem. If the RMSE of $\hat{\theta}$ and $\hat{\Sigma}^{-1}$ are ϵ_n, then policy optimization based on the doubly robust estimate, achieves regret:

$$R(\Pi, n) = O(Rademacher(\Pi) + \epsilon_n^2)$$

If $Rademacher(\Pi) = \mathrm{O}\left(\frac{1}{\mathrm{sqrt(n)}}\right)$ then as long as $\epsilon_n = o\left(n^{-1/4}\right)$ the impact from the estimates of θ and Σ does not affect the leading regret term.

Contribution 3: Variance Based Robust Regret

Out-of-sample Regularized ERM:

Split your final sample in two.

Estimate optimal policy on first sample using the DR estimate

Consider the subset of the policy space that on the first sample has DR value at most the first solution plus some error μ_n

Find the best policy within the subset based on the DR estimate on the second sample

Final Contribution: Variance Based Regret

Theorem. Consider the entropy integral

$$\kappa(\mathrm{r}, \Pi) \approx \int_0^r \sqrt{\frac{H_{2(\epsilon,\Pi,n)}}{n}}\, d\epsilon$$

Let V_2^0 denote the worst-case semi-parametric optimal variance of the difference between any two policies that are within μ_n of the true optimal.

Then the regret of out-of-sample regularized ERM is:

$$R(\Pi, n) = O\left(\kappa\left(\sqrt{V_2^0}, \Pi\right) + \sqrt{\frac{V_2^0}{n} + \epsilon_n^2}\right)$$

Example: for policies with constant VC dimension d: $R(\Pi, n) = O\left(\sqrt{V_2^0 \frac{d}{n} + \epsilon_n^2}\right)$

Optimal Pricing from Observational Data

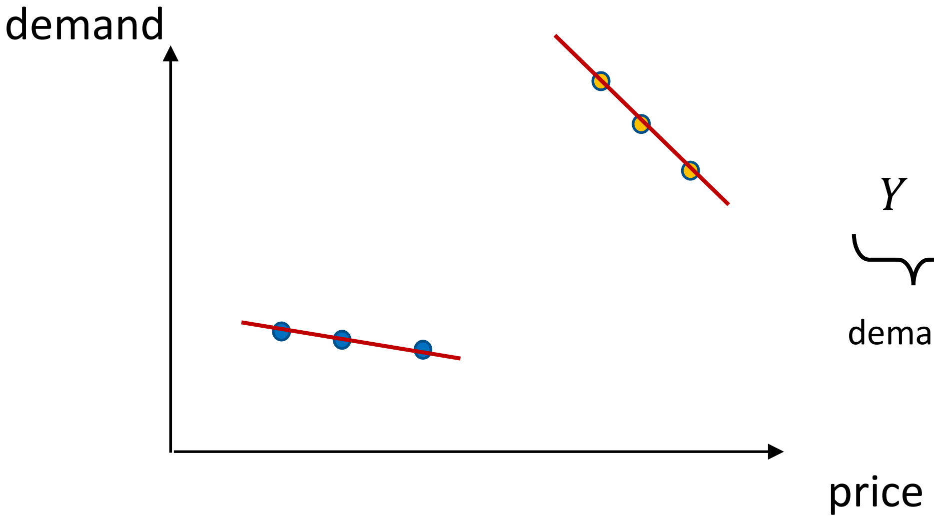
Personalized/Contextual

We are given historical data of demand and price from a company

Goal: Find a new optimal price **policy** based on the data

Approach: Estimate demand function $Q(p, x)$ then optimize policy revenue

$$V(\pi) = E[\pi(X) \cdot Q(\pi(X), X)]$$



$$\underbrace{Y}_{\text{demand}} = \overbrace{\underbrace{\alpha_0(X)}_{\text{contextual elasticity}} \cdot \underbrace{T}_{\text{price}} + \underbrace{\beta_0(X)}_{\text{baseline demand}}}^{Q(P,X)} + \underbrace{\epsilon}_{\text{noise}}$$

Can we get $\sqrt{n}$-rates for the optimal policy if policy space is simple, even if β_0 and α_0 is not $\sqrt{n}$ consistent?

Back to Pricing

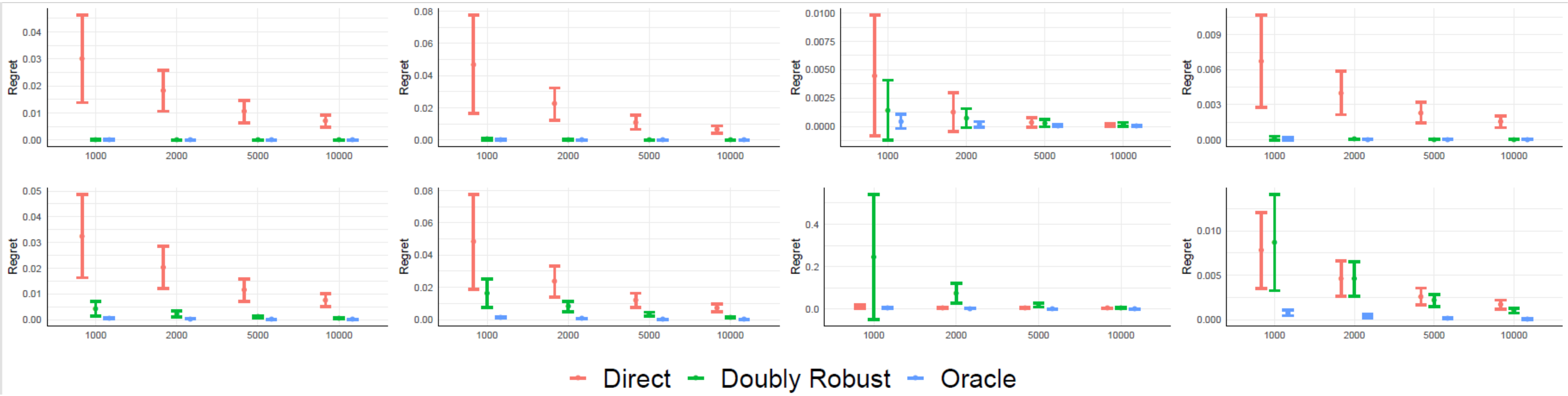
Under homoscedastic observational policy

$$\alpha_{DR}(X) = \hat{\alpha}(X) + \frac{T - E[T|X]}{Var(T - E[T|X])}\left(T - \hat{\alpha}(X)T - \hat{\beta}(X)\right)$$

$$\beta_{DR}(X) = \hat{\beta}(X) + \left(1 + \frac{(T - E[T|X])E[T|X]}{Var(T - E[T|X])}\right)\left(T - \hat{\alpha}(X)T - \hat{\beta}(X)\right)$$

So only need to regress $T \sim X$ and estimate the variance of the residuals of this regression.

Sneak Peak of Experimental Results



Conclusions

- Addressed off policy optimization from observational data with continuous actions
- Under a linear of value assumption provided novel Doubly Robust off-policy estimate
- Showed semi-parametric efficiency of the variance of estimate
- Provided novel out-of-sample regularized ERM algorithm
- Showed variance-based regret with second order dependence from first stage regression and policy estimates