Assessing the Unintented Consequences of A Right to Counsel

Anonymous Author(s)
Affiliation
Address
email

Abstract

In line with the thinking that "nonparametric identitifcation + parameteric estimation $\implies$ causal flavor" (as mostly prominently advocated for in Mostly Harmless Econometrics), we introduce an approach to causal inference that generalizes ols, flexibly accounts for clustering effects and is inherently compositional. Based upon recent developments in deep learning (i.e Neural ODEs and MAML), our model provides an alternative to the classical approach (i.e linear model + cluster standard errors).

what is the right expression?

Contents

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

1 High Level Idea

- In low dimensions, your trying to correct of a sampling issue
- In higher dimensions it's not clear what you are doing

One the one hand we could do

$$E[Y|D,X], \quad P(D,X) = P(D|X)P(X) = U(D)P(X)$$

On the other hand, you could do a partially linear model

What's in between?

$$p(x|d) = \frac{p(d|x)p(x)}{p(d)}$$

$$\begin{aligned} p_\theta(d,x) &= p_\theta(d|x)p_\theta(x) \\ \log p_\theta(d,x) &= \log p_\theta(d|x) + \log p_\theta(x) \end{aligned}$$

- With high dimensional data, can you learn an appropriate weighting? If not, why even do a partially linear model?
- Is economic data on a low dimensional manifold?
- Is it possible to think of treatment heterogeneity as existing along a low dimensional manifold?
- The weighting depends on those around it.

The partially linear model is in some sense the best you can do if you stick to a regression only approach.

- Map neural network into kernel methods
- Suggest correction for kernel methods
- map correction into neural network training

$$K(x,x') = \left\langle \phi(x), \phi(x') \right\rangle_{\mathcal{H}}$$

2 Overview

Quarterbacks on a freshman team are often taught to look only at one side of the field when considering who to target. As they progress in their development, the begin to consider the entire field, and perhaps get to a level where the head coach is comfortable with them positioning their teamates, and making adjustments to the play based on what they observe pre-snap.

Fitting linear models with an aim towards causal inference can have a bit of a freshman quarterback feel to it. There's this narrow and often singular concern as to whether x is correlated with ε with little discussion about optimization, generative modeling, or learning topics of that nature.

Implictly what's going on is that graduate courses and seminars have said the linear model is the most appropriate way to think about causal inference and the relationship between x and ε is first order.

What we would prefer, and what we argue for in this paper is that to the extent possible, we should think about the data nonparametrically, and work with models that can be fined tuned by the applied research to address the concerns that they believe are first order.

model's power stems from the elimination of irrelevant detail, which allows the economist to focus on the essential features of the economic reality he or she is attempting to understand – Hal Varian

- Probability Space: $\left(\Omega, \mathcal{F}, \mathbb{P}\right)$
- Index Set: T
- State Space: $\left(E, \mathcal{E}\right)$

- Stochastic Process: $\{X_t : \Omega \to E\}_t$

The feature map

- A way to correct for the conditional distribution of D given X

When one is learning econometrics, it can feel like 'applied papers' exhibit inconsistencies. For instance, topics like hypothesis testing are discussed, but then there's no mention of multiple hypothesis testing. Or, an instrument will be used, but there's no mention of the L.A.T.E.. And finally, selection on observables and the overlap conditions are assumed, but then partially linear models are fit instead of fully-nonparametric.

When you have selection on observables, why do you have controls in your model? And the answer, as we are all familiar with it because they allow us to account for the conditional distribution of treatment.

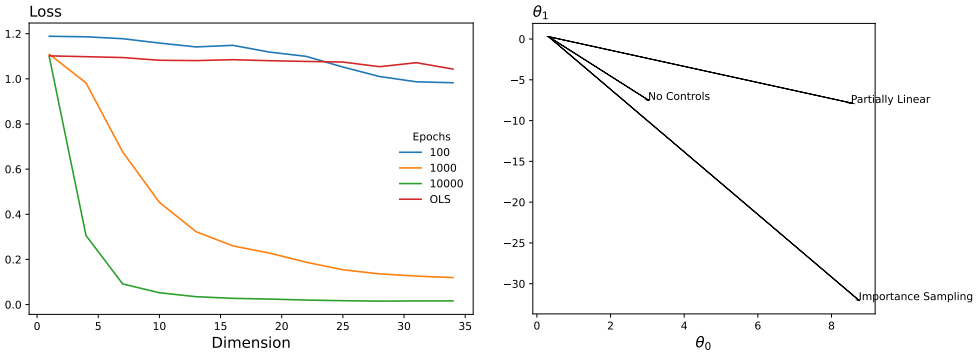


Figure 1: Reproduced Here

2.1 PML

2.1.1 Objective

Sampling From the Empirical Distribution

$$\text{key} \longmapsto f(\text{key}) = \text{jax.random.choice(key, X)}$$

Sampling from a Parameterized Distribution

$$\text{key} \longmapsto f(\text{key}) = \text{SamplingModel(key, params)}$$

2.1.2 Set-up

We have class of Gaussian random variables that are parameterized by our training data

$$\text{key} \longmapsto f(\text{key}, x_0) = \text{ForwardModel(key}, x_0)$$

- A
- Importance Sampling[1]
  - Can importance sampling backfire even with $l_\theta(t)$ as defined?

with selection on observables, a feature map's only use is to correct for the conditional distribution of treatment. It can be trivial to show that it doesn't go far enough!

[1] It can also backfire, yielding an estimate with infinite variance when simple Monte Carlo would have had a finite variance. – here

Let

$$l_\theta(t) = \big(E[Y|T=t] - f_\theta(t)\big)^2$$

Then

$$\int l_\theta(t)p(t)dt = \int \frac{l_\theta(t)p(t)}{q(t)}q(t)dt = E_q\left[\frac{l_\theta(t)p(t)}{q(t)}\right]$$

So then

$$\nabla_\theta E_q\left[\frac{l_\theta(t)p(t)}{q(t)}\right] = E_q\left[\frac{\nabla_\theta l_\theta(t)p(t)}{q(t)}\right]$$

$$\underset{\theta}{\text{minimize}}\ E_q\left[\frac{l_\theta(t)p(t)}{q(t)}\right]$$

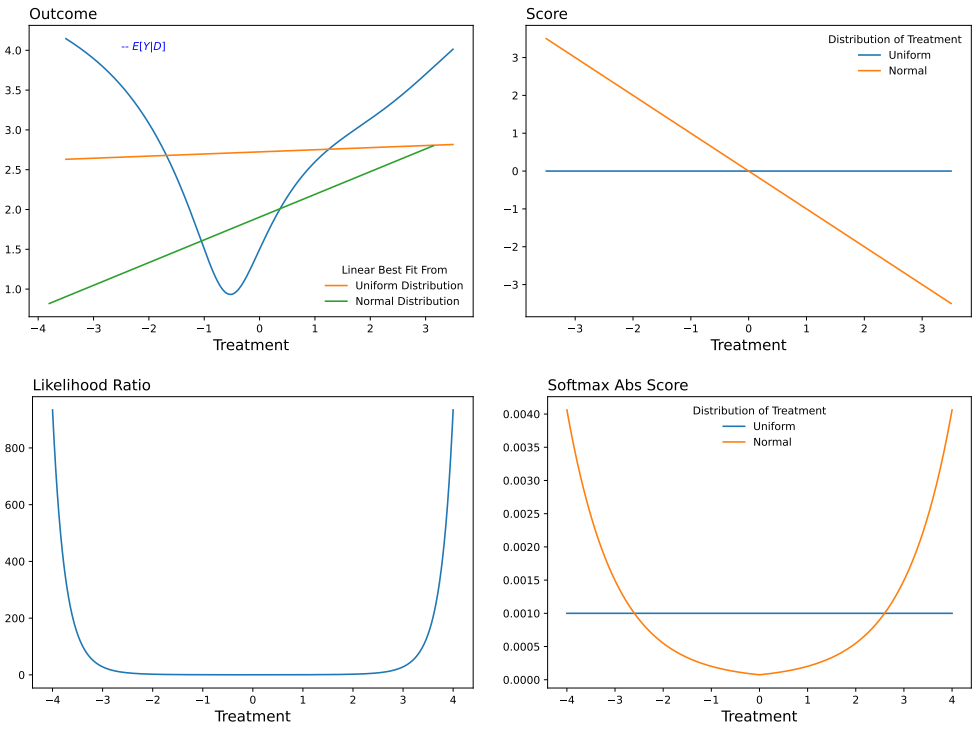


Figure 2: Reproduced Here

2.2 Problem

Which is to say

And so to put it a bit too simplicitly, the concerns of the Econometrician (nonparametric identification, consistency, asymptotic normality) are often not first order issues.

The selection process of treatment is not fully revealed. Nonparametric identification is therefore usually not attainable, and perhaps even more often, not applicable.

The starting point of applied microeconomic seminars, and the way we'll start this paper is how can I enhance the causal flavor of my estimator.

Which is perhaps another way of saying that the concerns of the applied microeconomist The sobering reality of applied causal inference is that you often find yourself betting on something you don't

believe is true, don't care about, or some combination of the two. No approach to causal inference can avoid this reality.

The hope, and to be sure, it is a hope, is that we might be able to learn something from the data. Not to convince someone else but rather to pick of. From the applied microeconomist's perspective, the question is how do I enhance the causal flavor of my estimator.

- A sinlge method that can apply to nonparameteric and parametric methods as well as cases where there is no variation of treatment within cluster
- The interpretability of a partially linear model! #causalflavor
- Nests OLS/feature models

Binary Treatment: For each level of treatment, we partial out the cluster effect

$$\begin{aligned} E(Y|D,X) &= (1-D)E(Y|D=0,X) + DE(Y|D=1,X) \\ &= (1-D)f_{\theta_1}(X) + Df_{\theta_2}(X) \end{aligned}$$

Parametric (Continuous Treatment) :

$$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2^T \phi_\theta(X_i)$$

$$E(D|X) = \beta^T \phi_\theta(x)$$

$$E(Y|D,X) \approx \alpha_0 + \alpha_1 D + \phi_\theta(x)$$

Model 1: This model breaks down if there is no within-cluster treatment variation

$$Y_i = \alpha_0 + \alpha_1 D_i + \phi_{\theta_c}(x) + \varepsilon_i$$

Interpretation 1:

$$Y_i = \alpha_1 (D_i - \underbrace{\beta^T \phi_{\theta_c}(x)}_{\approx E(D|X,C)}) + \varepsilon_i$$

Interpretation 2:

$$\begin{aligned} v_i &= Y_i - \beta_1^T \phi_{\theta_c}(x) \\ \mu_i &= D_i - \beta_2^T \phi_{\theta_c}(x) \\ v_i &= \theta \mu_i + \varepsilon_i \end{aligned}$$

No variation of treatment within cluster, but within same level of treatment, multiple clusters:

- How sensitive is the partially linear model to $\mathbb{P}(D|X)$ (good enough for us!)
- In the finite sample, we might prefer to partial out versus average over (aren't these the same?). That is I think what we want to run is the following

$$\begin{aligned} v_i &= Y_i - E(Y|X,C) \\ \mu_i &= D_i - E(D|X) \\ v_i &= \theta \mu_i + \varepsilon_i \end{aligned}$$

- $E(D|X)$ should be known

2.3 Overview of paper

We introduce a relatively simple, yet general way to adjust one's estimator for the clustered nature of microeconomic data. As applied microeconomists, we use our approach to examine the intended and unintended consequences of Connecticut's adoption of a Right to Counsel. The generality of

our approach allows us to estimate both parametric and nonparametric estimands within the same conceptual framework.

$$E(Y|X) = f_\theta(X)$$
$$E(Y|X, C) = f_{\theta_c(\theta)}(X)$$

Parametric estimation actually nests nonparametric estimation Like the majority of Economic papers, we are interested in reporting the best linear predictor when the To be precise, we're interested in the OLS estimates We are interested in estimating linear models because the coefficients associated with these models are easy to communicate to other interested parties. This is distinct from the idea that linear models are easy or straightforward to interpret. We are not comfortable, though, with the idea of using a linear model to justify estimating a linear model. Such conversations dissolve into debates over whether the causal variable of interest D_i is correlated with ε_i. Because the coefficient on D_i is unknown, then ε_i is unknown as well, and so unless one imposes a constant treatment effect assumption on top of the potential outcome framework it's not clear what people are in fact arguing about. Just with repect to a probability measure from To make this clear, we define it as the solution to the problem that we would like to solve if we has access to the probability measure of our choice.[2]

$$\theta^*(\mathbb{P}) = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{\mathbb{P}}\left[(Y_i - \theta_0 - \theta_1 D)^2\right]$$

Following Mostly Harmless Econometrics the pillars of our economic analysis are nonparametric identification and parametric estimation.

Nonparametric identification + parametric estimation $\implies$ causal flavor.

Within this general approach to causal inference, we introduce a method that enhances the causal flavor as it generalizes OLS, flexibly accounts for the cluster nature of data, and is inherently compositional.

- **Generalizing OLS**: OLS can be refactored into the composition of identity functions. Our method simply replaces these identity functions with suitable alternatives, thereby producing reasonable estimates on benchmark examples where the OLS model breaks down
- **Flexible Clustering Effects**: This highlights that the cluster map, this attempt to correct for cluster sampling, can be thought of as a gradient correction
- **Inherently Compositional**: Even with the data-dependent form of regularization that we make use of our method remains compositional, which means that conceptually, the model remains simple.[a][b]

[a] The composition of partially evaluated functions

[b] Perhaps what we mean by this is that it can be shown to be compositional which provides us with a simple way to understand the model. The fact that something can be expressed in a compositional manner is a relatively weak statement on its own. Consider how one might define 3 in agda[c]

```
three : ℕ
three := suc (suc (suc zero))
```

We apply our method to ...

[2] We are greatly inspired in this introduction by the presentation of the statistical learning problem as presented by Professor Lorenzo Rosasco: see here

2.4 Implementation

2.5 Parallel Execution

- tree_map: However, JIT compiler may execute code without data dependency in parallel.[3]

Our belief is that if we can show you how each submodel of our model can be thought of as a morphism in a category where the objects are types then either our model is simple, or we have provided a simple way to conceptualize our model.[4]

I think we have more structure than just a monad, because the function retains the same output when its a morphsim in the category SET

Of these transformations (is that the right word), perhaps the most interesting is regularzation which can be thought of as a notion of computation and defines a monad.[5]

- Our model is a function from A to $T(B)$.

$$\text{A} \rightarrow \text{T(B)}$$

- We can think of this though as a morphism from A to B

$$\text{A} \rightsquigarrow \text{B}$$

- We can define composition of arrows as $>=>$
- Then $(\mathcal{A}, \rightsquigarrow, >=>)$[6]
- Arrows represent the partial evaluation of functions

Below we capture the above points by writing the model in pseudo Haskell[7] code.[8]

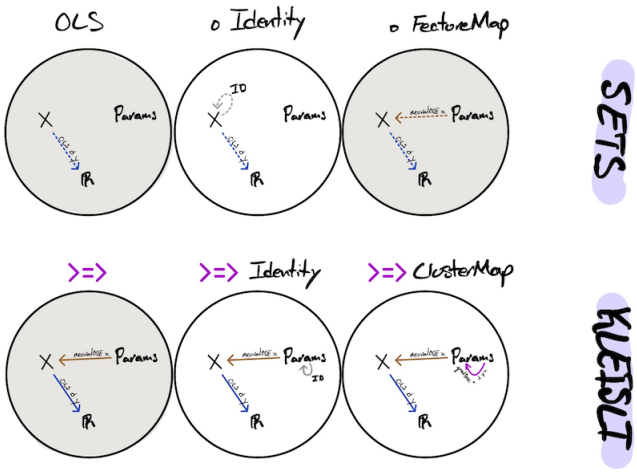


Figure 3: This can be seen as a complete refactoring of OLS by introducing the identity functor (this part is not reflected in the refactor arrow). If we look at this diagram vertically, we see that we are workign with two categories, and our "model" is attained by selecting the appropriate morphism between these two categories (where the morphisms are functors)

[3]

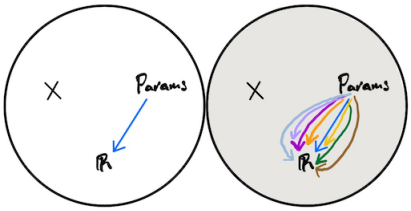
[4] You are of course free to choose either of these two definitions or to disagree, and we would be happy to hear which you choose

[5] A type constructor and an extendor

[6] If the category of T-programs is constructed using a Kleisli triple, as depicted on the next slide, then T defines a monad. – She says this should be attributed to Main?

[7] the feature and cluster maps both take as inputs data and parameters. The dollar sign denotes partial application (i.e. currying) of a function. ∘ denotes composition

[8] A reader familiar with Haskell/Category theory will recognize that the act of adding regularization to our model conceptually moves us from the category of Sets to the Kleisi Category. A full description of this change following the layout of Category Theory for Programmers is described in the appendix



$$\begin{aligned}
\text{ols data} &= \text{linearModel data} \\
\text{ols data _} &= \Big(\text{linearModel} \circ \$\ \text{identityMap data} \circ \$\ \text{identityMap data}\Big) \\
\text{fwd_pass data params} &= \Big(\text{linearModel} \circ \$\ \text{featureMap data} \circ \$\ \text{clusterMap data}\Big)\ \text{params} \\
\text{fwd_pass data params} &= \Big(\text{linearModel} \circ \$\ \text{neuralODE data} \circ \$\ \text{MAML data}\Big)\ \text{params} \\
\text{reg_fwd_pass data params} &= \Big(\text{linearModel} >\!=\!> \$\ \text{regNeuralODE data} >\!=\!> \$\ \text{regMAML data}\Big)\ \text{params}
\end{aligned}$$

$$\begin{aligned}
&\text{regNeuralODE} :: \text{Data} \to \text{Params} \to \big(\text{Data}, \text{Float}\big) \\
&\text{regNeuralODE _ } x \text{ _ } \theta := x + \int f(t, x(t), \theta) dt, \quad \int \left\| \frac{\partial^k}{dt^k} f(t, x(t), \theta) \right\| dt \\
&\qquad\qquad \text{where} \quad x(0) = x
\end{aligned}$$

$$\begin{aligned}
&\text{regMAML} :: \text{Data} \to \text{Params} \to \big(\text{Params}, \text{Float}\big) \\
&\text{regMAML data } \theta := \Big(\text{Update}_m \circ \text{Update}_{m-1} \cdots \circ \text{Update}_1\Big)\ \theta, \quad \mathcal{L}_c(\text{data}, \theta) \\
&\qquad\qquad \text{where} \quad \text{Update}_t\ \theta = \theta - \alpha_t \nabla \mathcal{L}_c(\text{data}, \theta)
\end{aligned}$$

3 The Why

- Feature Map
  - We would like the function space perspective, but we'll settle for an input space perspective in contrast to regularizing the parameter space
- Cluster Effects
  - Gradient Correction that favors early stopping at the cluster level
  - Parametrics: We want to be able to control the influence of certain variables (like cluster indicators)
  - Nonparametrics: We want to partial out the cluster effects! (We want to use the cluster level information in the training set to produce a cluster-free model) AKA generalizing across clusters
- Defining the problem (need to add the clustering part!)
- Defining conditional independence
- Defining the linear approximation

3.1 The Problem We Would Like to Solve

- "At the level of this book, the theory would be more 'elegant' if we regarded a random variable as an equivalence class of measurable functions on the sample space, two functions belonging to the same equivalence class if and only if they are equal almost everywhere"[9]
- We begin by defining our probability space of interest as follows

$$
\begin{aligned}
&\big(\Omega_1, \mathcal{F}_1, \mu_1\big), \quad \omega \in \Omega_1 \longmapsto Y_{po}(\omega) \in C(\mathbb{R}) \\
&\big(\Omega_2, \mathcal{F}_2, \mu_2\big), \quad \omega \in \Omega_2 \longmapsto D(\omega) \in \mathbb{R} \\
&\big(\Omega, \mathcal{F}, \mu\big) = \big(\Omega_1 \times \Omega_2, (\mathcal{F}_1 \times \mathcal{F}_2), \mu_1 \times \mu_2\big) \\
&\omega = (\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 \longmapsto Y_{po}(\omega_1)(D(\omega_2)) = Y(\omega) \in \mathbb{R}
\end{aligned}
$$

- And our parameter of interest (where H is the set of of functions. In the appendix we walk through this derivation in more detail.

Is this the set of affine, $\sigma(D)$ measurable, square integrable functions?

$$
\begin{aligned}
\underset{g \in \mathcal{H}}{\operatorname{argmin}} \, \mathbb{E}_{\mu}\big[(Y - g_\theta(D))^2\big] &= \underset{g \in \mathcal{H}}{\operatorname{argmin}} \, \mathbb{E}_{\mu}\big[(E(Y|D) - g_\theta(D))^2\big] \\
&= \underset{g \in \mathcal{H}}{\operatorname{argmin}} \, \mathbb{E}_{\mu_2}\big[(E(Y|D) - g_\theta(D))^2\big] \\
&= \underset{g \in \mathcal{H}}{\operatorname{argmin}} \, \mathbb{E}_{\mu_2}\big[(E(Y_D) - g_\theta(D))^2\big]
\end{aligned}
$$

3.2 The Problem We Are Dealt (1)

3.2.1 Treatment Assignment Mechanism

As we'll now explain, we don't have access to realizations from the product measure[10]. Instead our measure v can be constructed as follows.[11]

$$
\mu \ll v, \quad \mu \overset{T}{\longmapsto} v; \quad v(A) = \int_A f d\mu \quad \forall A \in \mathcal{F}
$$

With the additional condition that for all

$$
\omega_1 \longmapsto \mathbb{E}[f(\omega_1, \omega_2)] \overset{a.s.}{=} \omega_1 \longmapsto 1
$$

[9] Williams

[10] We need these measures to by σ -finite

[11] We model the treatment assignment mechanism as a higher order function on the set of measures

As highlighted below

$$\begin{aligned} v(A_1 \times \Omega_2) &= \int_{A_1 \times \Omega_2} f d\mu \quad \text{by Radon Nykodym} \\ &= \int_{A_1} \int_{\Omega_2} f(\omega_1, \omega_2) d\mu_2 d\mu_1 \quad \text{by Tonelli's theorem} \\ &= \int_{A_1} \mathbb{E}[f(\omega_1, \omega_2)] d\mu_1 \quad \text{by definition} \\ &= \mu(A_1) \end{aligned}$$

This is an important condition because it tells us that we just need to work to get the following:

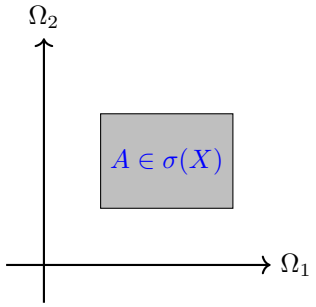
We know that

$$E_v(Y_D|X) = E_\mu(Y_D|X)$$

What we have now is

$$E_v\big[E_v(Y_D|X)\big] = E_\mu\big[E_\mu(Y_D|X)\big]$$

- **Clusters**:
  - In what follows, we assume that the number of clusters is known ex ante



- Consider
  1. Conditional on X treatment is randomly assigned, but the conditional distribution of treatment might not match the desired one!
  2. Are we defining equality of random variables the right way? See the appendix

  By SUTVA

$$\int_A E(Y|D,X)dv = \int_A E(Y_D|D,X)dv \quad \forall A \in \sigma(D,X)$$

  By Conditional Independence Assumption

$$= \int_A E(Y_D|X)dv \quad \forall A \in \sigma(D,X)$$

  By the restriction on f

$$D \longmapsto \int_{\Omega_1} E(Y|D,X)dv \iff D \longmapsto \underbrace{\int_{\Omega_1} E(Y_D|X)d\mu_1}_{=\mathbb{E}_{\mu_1}[Y_D]}$$

  As

$$\int_{\Omega_1} E(Y|D,X)dv = \int_{\Omega_1} E(Y|D,X) f d\mu_2 \mu_1$$

Therefore, we can define an objective function which we observe who shares the same minimum as our desired objective function:

$$= \underset{g \in \mathcal{H}}{\text{argmin}} \int_{\Omega_2} \Big(\int_{\Omega_1} E[Y|D,X] d\mu_2 - g_\theta(D) \Big)^2 d\mu_1$$
$$= \underset{g \in \mathcal{H}}{\text{argmin}} \int_{\Omega_2} \Big(E[Y_D] - g_\theta(D) \Big)^2 d\mu_1$$

and defining my objective function in terms of this function.

$$\theta^* = \underset{\theta}{\text{argmin}} \int_{\Omega_2} \big(h_X \circ D - g(D) \big)^2 d\mu_2$$

$$E[1_U 1_V | X] = E[1_U | X] E[1_V | X] \quad \forall U \in \sigma(Y_D), V \in \sigma(D)$$

- We start by introducing our probability space of interest $\big(\Omega, \mathcal{F}, \mathbb{P}\big)$ and our working probability space: $\big(\Omega, \mathcal{F}, v\big)$ with the following random variables, Y_i, X_i, D_i, C_i, defined on the measurable space $\big(\Omega, \mathcal{F}\big)$.
- Parameter of Interest
  - Let $W_i = (Y_i, D_i)$ be a continuous random variable on $\big(\Omega, \mathcal{F}, \mathbb{P}\big)$ with pdf f_W.
  - Let $g_\theta(W_i) = (Y_i - \theta_0 - \theta_1 D_i)^2$, be an $\mathcal{F}$-measurable function.[12]

$$\theta^*(\mathbb{P}) = \underset{\theta}{\text{argmin}}\, \mathbb{E}[g_\theta \circ W] = \underset{\theta}{\text{argmin}} \int g_\theta d\mathbb{P}_W = \underset{\theta}{\text{argmin}} \int g_\theta f_W d\lambda$$

- **Clustered Data**:

$$C : \mathcal{F} \to \{\mathbb{P}\} \to \{\mathbb{P}\}$$
$$\mathbb{P}, A \longmapsto \mathbb{P}_A$$
$$\text{where} \quad \mathbb{P}_A(B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$

We define the event of interest, A_m, as follows (where m denotes the number of clusters)[13]

$$H : \Omega \to \mathbb{N}$$
$$\omega \longmapsto H = \lim_{n \to \infty} \text{numUnique}(C_{i \leq n})$$
$$A_m \in \mathcal{F} := \{\omega \in \Omega | H(\omega) = m\}$$

- Treatment Assignment Mechanism:

$$T : \{\mathbb{P}\} \to \{\mathbb{P}\}$$

$$v = (C\ A_m) \circ T\ \mathbb{P}$$

- Identification

$$Y_i(D_i) \perp\!\!\!\perp D_i \mid X_i$$

- **Conditional Expectation**

[12] We don't need to assume that it is absolutely integrable because it is a nonnegative function – ref

[13] In haskell, we might define this function as done in LearnYouaGoodHaskell:

```
numUniques :: (Eq a) => [a] -> Int
numUniques = length . nub
```

- TL;DR[14] (a version of)[15]the conditional expectation of a random variable X with respect to another random variable, Y, is (a/the) random variable, $E_Y(X)$, whose integral over any element in $\sigma(Y)$ is equal to that of X. That is

$$\int_A E_Y(X) d\mathbb{P} = \int_A X d\mathbb{P} \quad \forall A \in \sigma(Y)$$

Moreover, we can see that any $\sigma(Y)$-measurable and integrable function $g \circ Y$

$$\begin{aligned} & \int_A (g \circ Y) E_Y(X) d\mathbb{P} = \int_A (g \circ Y) X d\mathbb{P} \quad \forall A \in \sigma(Y) \\ \iff \quad & \int_A (g \circ Y) \big(E_Y(X) - X\big) d\mathbb{P} = 0 \end{aligned}$$

- The **Law of Iterated Expectation** is just a special case of the above statement

$$\begin{aligned} \mathbb{E}\big[E_Y(X)\big] &:= \int_\Omega E_Y(X) d\mathbb{P} \\ &= \int_\Omega X d\mathbb{P} \\ &= \mathbb{E}\big[X\big] \end{aligned}$$

- What I want I think integrating over X-measurable D-constant subsets seems what we want to do (ish)

$$E_\mu[Y|D, X] = E_{\mathbb{P}}[Y_D|X]$$

- Conditional Probability
  - Partially evaluated, this is a continuous linear function[16]
  - The atoms of a countably generated sigma algebra is measurable

$$E :: S(\mathcal{F}) \to L^1(\Omega, \mathcal{F}, \mathbb{P}) \to L^1(\Omega, \mathcal{F}, \mathbb{P})$$

such that

$$\int E_{\mathcal{A}}(f) d\mathbb{P} = \int_A f d\mathbb{P} \quad \forall A \in \mathcal{A}$$

[14]The conditional expectation of X with respect to either Y or Z will be the same random variable if $\sigma(Y) = \sigma(Z)$

[15]https://web.ma.utexas.edu/users/gordanz/notes/conditional_expectation.pdf

[16]pf. 177

4 The What

Hi

$$E(D|X)$$

5 Understanding the Effect(s) of Right to Cousel

5.1 Overview

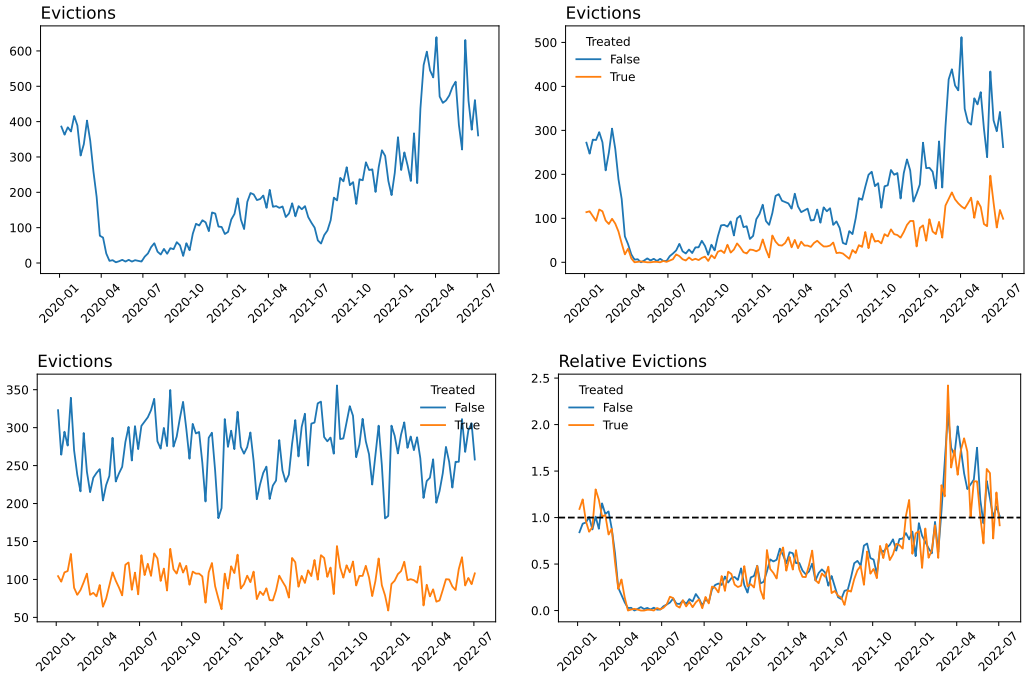


Figure 4: Reproduced Here

5.2 Effects on low Income locations

5.3 Unintended Consequences

- The resemblance between equations (3.1) and (3.4) is clear. The essential difference is that equation (3.4) is driven by the data process x, whilst equation (3.1) is driven only by the identity function $R \to R$.[17]
- In Connecticut, 80% of landlords have legal representation in eviction cases, while only 7% of tenants have the same benefit.[18]
- Before the pandemic, an average of 20,000 residents faced eviction in Connecticut each year. And four cities in Connecticut were in the top 100 in the nation for most evictions in 2016.[19]
- "If the system works from start to finish in 45 days, landlords would take a chance on anybody for the most part," Souza said. "At the end of the next month, if the whole thing is resolved, I have a smaller chance of losing money. If it takes me 6 months to resolve someone not paying, I'm losing six months and can't recover the money."[20]
- This program is offered without pre-conditions like employment, income, criminal record, or sobriety.[21]

[17] https://arxiv.org/pdf/2202.02435.pdf

[18] https://www.ctpublic.org/news/2022-01-30/some-residents-facing-eviction-could-now-be-eligible-for-free-legal-aid

[19] https://www.ctpublic.org/news/2022-01-30/some-residents-facing-eviction-could-now-be-eligible-for-free-legal-aid

[20] https://ctnewsjunkie.com/2022/01/31/right-to-counsel-launches-to-help-fight-evictions/

[21] https://www.newreach.org/programs/housing

Program Type	Count
Coordinated Assessment	9408
Emergency Shelter (ES)	1803
Services Only	963
Homelessness Prevention	899
Not a HUD HMIS Project	752
Street Outreach	254
Transitional Housing (TH)	242
Safe Haven (SH)	19
Other	1
Day Shelter	1

Table 1: Programs listed by Usage

6 Conclusion

The sobering reality of applied causal inference is that you often find yourself betting on something you don't believe is true, don't care about, or some combination of the two. No approach to causal inference can avoid this reality. The best we can do is to evaluate the tradeoffs of our approach and be cautious in our interpretations.

7 Deep Controls

- **Starting Point**: Out of distribution, we are interested in fitting a linear model w
- Underlying probability space: $\big(\Omega, \mathcal{F}, \mathbb{P}\big)$[22]
- When we "sample" ω, it's as if we are sampling a single individual from the population of interest

$$\theta_0 = \underset{\theta}{\text{argmin}} \int \int \big(Y(d,\omega) - \theta_0 - \theta_1 d\big)^2 d\lambda d\mathbb{P}$$

- The story I have told so far, is that the ideal would be to have a fully flexible model of $g(x)$
- But this isn't exactly true – see here

7.1 Limitations of OLS

Partial Non-parametric Identification

$$Y_i(D_i) \perp\!\!\!\perp D_i \mid X_i$$

Linear Approximation

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i + \hat{\beta}_2 X_i$$

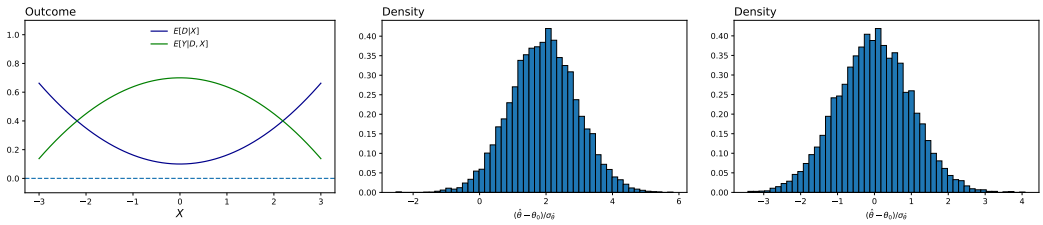


Figure 5: Reproduced Here

$$Y_i = \beta_0 + \beta_1 D_i + g(X_i) + \varepsilon_i$$

If you have non-parametric identification, and if you are motivated to use this partially linear model, then the question becomes what's the "right" functional prior on g?

7.2 Sensitivity of Estimate to the Conditional Treatment Distribution: Notebook

- $D \in [0,1]$
- $\bar{Y}(D) \in [-c, c]$
- $\tilde{D} : \{\mathbb{P}\} \times [0,1]$

$$\alpha^*(\bar{Y}, \mathbb{P}), \_ = \underset{\alpha,\gamma}{\text{argmin}} \int \big(\bar{Y}(D) - \gamma - \alpha D\big)^2 d\mathbb{P}$$

$$\begin{aligned}
\alpha^*(\bar{Y}, \mathbb{P}) &= \underset{\alpha}{\text{argmin}} \int \big(\bar{Y}(D) - \alpha \tilde{D}(\mathbb{P}, D)\big)^2 d\mathbb{P} \\
&= \underset{\alpha}{\text{argmin}} \int \big(\bar{Y}(D) - \alpha\big(D - \int D d\mathbb{P}\big)\big)^2 d\mathbb{P} \\
&= \underset{\alpha}{\text{argmin}} \int \big(\bar{Y}(D) - \alpha D + \alpha \mathbb{E}[D]\big)^2 d\mathbb{P}
\end{aligned}$$

- Can I think about changing the distribution of treatment as integrating over the outcome with respect to a different measure?

[22] Where for recall, $\mathbb{P} : \mathcal{F} \to \bar{\mathbb{R}}^{+0}$ – ref

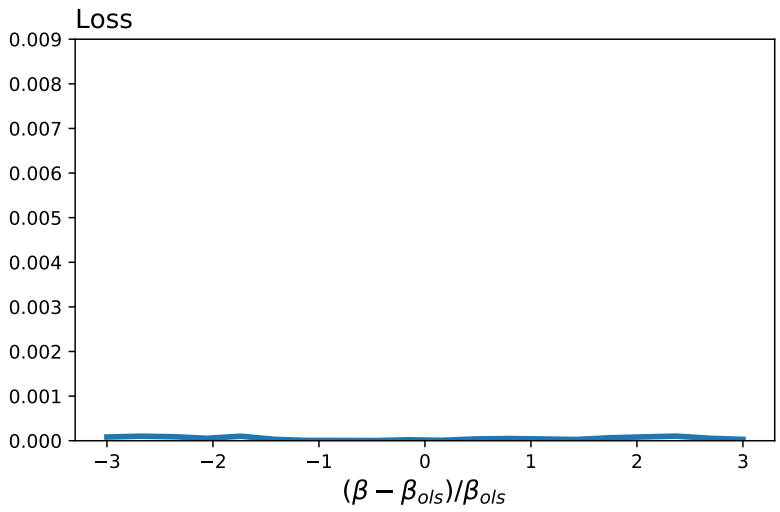


Figure 6: Reproduced Here

7.3 Weakness of Instruments

- Let $Y_i(Z)$ be person i's potential outcome function with respect to the instrument Z
- Let $\pi_i(Z)$be person i's potential treatment function with respect to the instrument Z

$$Y_i(1) = \tilde{Y}_i\big(\pi_i(1)\big)$$

- Doing so, we see that we can decompose the the ITT as follows, where we assume one-sided non compliance.

$$\begin{aligned}
\text{ITT} &= \int Y_i(1) - Y_i(0) d\mu \\
&= \int \tilde{Y}_i\big(\pi_i(1)\big) - \tilde{Y}_i\big(\pi_i(0)\big) d\mu \\
&= \int \tilde{Y}_i\big(\pi_i(1)\big) - \tilde{Y}_i\big(0\big) d\mu \quad \text{by one-sided noncompliance}
\end{aligned}$$

Which is just the average policy effect

7.4 Partially Linear Models

$$\phi_\theta(x) = x + \int f(t, x(t), \theta) dt, \quad x(0) = x$$

8 Full Model

8.1 Nesting OLS

TLDR; Our method nests OLS by showing that (1) OLS can be refactored into the composition of identity functions, and our method simply replaces these identity functions with suitable alternatives.

- We want to show how our estimation approach nests OLS
- We begin with a table which highlights how learning frameworks differ across key characteristics. Importantly, these are composable characteristics.

Table 2: Learning Frameworks

Framework	Top Model	Feature Map	Cluster Map	Trainable Params
OLS	Linear	Identity	Identity	False
Kernel Methods	Linear	Non-Identity	Identity	False
Supervised Learning	Linear	Non-Identity	Identity	True
Cluster Supervised Learning	Linear	Non-Identity	Non-Identity	True

- Which means that to go from OLS to Cluster Supervised Learning we need to simply refactor OLS and replace the newly introduce identity function.

8.2 Does Clustering Still Matter?

$$\theta_c^*(\theta) := \underset{\theta_c}{\operatorname{argmin}} f_1(\theta, \theta_c)$$

$$\Phi_{\theta_c^*(\theta)}(x) = \int f_2(t, \hat{x(t)}, \phi_c^*(\theta))dt \quad x(0) = x$$

$$\underset{\beta}{\operatorname{argmin}} \tfrac{1}{N_c} \sum \left(y_i - \beta_0 - \hat{\beta_1 d_i} - \beta_2 \Phi_{\theta_c^*(\theta)}(x)\right)^2$$

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points. Times New Roman is the preferred typeface throughout, and will be selected for you by default. Paragraphs are separated by ½ line space (5.5 points), with no indentation.

The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow ¼ inch space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

For the final version, authors' names are set in boldface, and each name is centered above the corresponding address. The lead author's name is to be listed first (left-most), and the co-authors' names (if different address) are set to follow. If there is only one co-author, list both author and co-author side by side.

Please pay special attention to the instructions in Section 9 regarding figures, tables, acknowledgments, and references.

9 Citations, figures, tables, references

These instructions apply to everyone.

9.1 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

> Hasselmo, et al. (1995) investigated...

If you wish to load the `natbib` package with options, you may add the following before loading the `neurips_2022` package:

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{neurips_2022}
```

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]," not "In our previous work [4]." If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form "A. Anonymous."

9.2 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number[23] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

Note that footnotes are properly typeset *after* punctuation marks.[24]

9.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

[23] Sample of the first footnote.

[24] As in this example.

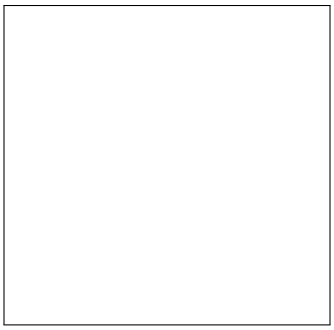


Figure 7: Sample figure caption.

Table 3: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

9.4 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 3.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

https://www.ctan.org/pkg/booktabs

This package was used to typeset Table 3.

10 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

11 Preparing PDF files

Please prepare submission files with paper size "US Letter," and not, for example, "A4."

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You should directly generate PDF files using `pdflatex`.
- You can check which fonts a PDF files uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdffonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- The IEEE has recommendations for generating PDF files whose fonts are also acceptable for NeurIPS. Please see http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf

- `xfig` "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

  ```
  \usepackage{amsfonts}
  ```

  followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for $\mathbb{R}$, $\mathbb{N}$ or $\mathbb{C}$. You can also use the following workaround for reals, natural and complex:

  ```
  \newcommand{\RR}{I\!\!R} %real numbers
  \newcommand{\Nat}{I\!\!N} %natural numbers
  \newcommand{\CC}{I\!\!\!\!C} %complex numbers
  ```

  Note that `amsfonts` is automatically loaded by the `amssymb` package.

If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

11.1 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the `graphicx` package. Always specify the figure width as a multiple of the line width as in the example below:

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

See Section 4.4 in the graphics bundle documentation (http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command when necessary.

References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the Reference section does not count towards the page limit.

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **[TODO]**
   (b) Did you describe the limitations of your work? **[TODO]**
   (c) Did you discuss any potential negative societal impacts of your work? **[TODO]**
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **[TODO]**
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? **[TODO]**
   (b) Did you include complete proofs of all theoretical results? **[TODO]**
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[TODO]**
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **[TODO]**
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[TODO]**
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[TODO]**
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? **[TODO]**
   (b) Did you mention the license of the assets? **[TODO]**
   (c) Did you include any new assets either in the supplemental material or as a URL? **[TODO]**
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[TODO]**
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[TODO]**
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[TODO]**
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[TODO]**
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[TODO]**

A Appendix

A.1 Controlling for Time

$$y(0) = \zeta_\theta(x(0)), \quad y(t) = y(0) + \int_0^t f_\theta(y(s))dx(s)$$

$$\theta, x \longmapsto y$$

$$y_t = x + \int f(s, x(s), \theta)ds, \quad x(0) = x$$

$$\theta, x, t \longmapsto y_t$$
$$\begin{aligned} y &= x + \int h(t, s, x(s), \theta) ds, \quad x(0) = x \\ &= x + \int f(s, x(s), \theta_1) g(t, \theta_2) ds, \quad x(0) = x \end{aligned}$$

A.2 Questions

- This allows us to control the influence of time
- Does this map into a Controlled Neural ODE? (this would be efficient, right?

A.3 Measure Theory Notes

- Alternative Notation for of Parameter of Interest

$$\begin{aligned} \theta &= \underset{\theta}{\operatorname{argmin}} \int_{\Omega_2} \int_{\Omega_1} f_\theta(\omega_1, \omega_2)^2 d\mu_1 d\mu_2 \\ \text{where} \quad f_\theta(\omega_1, \omega_2) &:= Y(\omega) - \theta_0 - \theta_1 D(\omega_2) \end{aligned}$$

- 1st Equality Sign

$$\begin{aligned} \mathbb{E}\big[(Y - g(D))^2\big] &= \mathbb{E}\big[(Y - E[Y|D] + E[Y|D] - g(D))^2\big] \\ &= \mathbb{E}\big[(Y - E[Y|D])^2\big] + \mathbb{E}\big[(E[Y|D] - g(D))^2\big] \\ &\quad + \overbrace{\mathbb{E}\big[(Y - E[Y|D]) \underbrace{(E[Y|D] - g(D))}_{\phi(D)}\big]}^{=0} \end{aligned}$$

- The second equality sign is related to the following:[25]

$$\int_A E(Y|D) d\mu = \int_A Y d\mu \quad \forall A \in \sigma(D)$$

$$\begin{aligned} A_{\omega_2} &= \{\omega_1 \in \Omega_1 : (\omega_1, \omega_2) \in A \subset \Omega\} \subset \Omega_1 \in \mathcal{F}_1 = \Omega_1 \\ \mu(A) &= \int_{\Omega_2} \mu_1(A_{\omega_2}) d\mu_2 \\ &= \int_{\Omega_2} \mu_1(\Omega_1) d\mu_2 \\ &= \int_{\Omega_2} d\mu_2 \end{aligned}$$

- The fourth equality sign is related to the following

$$\int_A E(Y|D) d\mu_2 = \int_A E[Y_D] d\mu_2 \quad \forall A \in \sigma(D)$$

- A measureable function is integrable if

$$\int |f| d\mu < \infty$$

- Space of Integrable Functions

$$L^1(\Omega, \mathcal{F}, \mathbb{P}) := \{f : \Omega \to \mathbb{R} | f \text{ is measurable}, \int |f| d\mathbb{P} < \infty\}$$

[25] helpful lecture

- **Question**: If we have two random variable f and g whose integration matches on $\mathcal{F}$ then we can replace one for the other in our objective function?
- Proof by contradiction: Assume the following holds[26]

$$\int_A f d\mu = \int_A g d\mu \quad \forall A \in \mathcal{F}$$
$$\int F(f) d\mu \neq \int F(g) d\mu$$

By definition:

$$\int F(f) d\mu = \underset{h \in S(F(f))}{\text{argmax}} \sum_{i=1}^{n} a_i \mu(A_i)$$

A.4 Kleisil Category

- From a programmer's perspective, computing the empirical loss of our model stays within the category of types and functions[27]
- To add regularzation to our model, we have to
  1. Introduce a new type constructor[28]
  2. Augment the output of our functions [29]
  3. Redefine Composition[30]
- Doing so generates a new category (the Kleisli Category)where , but where
  1. Objects are types (just as they were before)
  2. Morphisms between A, B are now functions from A to Kleisi B
  3. Composition is defined using the $>=>$ operator

B Full Model

The choice of regularization can be thought of as a choice of composition!

B.1 Empirical Loss

High Level Summary: At the "batch" level, it's just composition

$$\mathcal{L}_c(\theta, \mathcal{D}) := \mathcal{E}(\theta_c^*(\theta, \mathcal{D}), \mathcal{D})$$

[26] Helpful proof description

[27] In Haskell with a bit of handwaving, we can consider this category to be SET

[28]
type Kleisi a = (a, Float)

[29]
f :: a $\rightarrow$ Kleisi b

[30]
$$f >=> g = x \rightarrow$$
$$\text{let } (a, a_1) = f\ x$$
$$(b, b_1) = g\ a$$
$$\text{in } (b, a_1 + b_1)$$

We start with a Model, which we denote by M, which outputs predictions for a given set of controls x and and variable(s) of interest, d.

$$M(\theta, x_i, d_i)$$

Given a model, we can measure the fit as follows:

$$\mathcal{E}(\theta, \mathcal{D}) = \sum_{i}^{n} \big(y_i - M(\theta, x_i, d_i)\big)^2$$

As we are working with clustered data, we can assess the fit of our model on each cluster as follows:

$$\mathcal{E}_c(\theta, \mathcal{D}_c) = \sum_{i \in c} \big(y_i - M(\theta_c^*(\theta), x_i, d_i)\big)^2$$

$$\theta_c^*(\theta) = \theta + \int -\nabla \mathcal{E}(\theta(t), \mathcal{D}_c) dt, \quad \theta(0) = \theta$$

Our Empirical Loss function is then

$$\mathcal{L}(\theta, \mathcal{D}) = \sum_{c} \mathcal{E}_c(\theta, \mathcal{D}_c)$$

B.2 Regularization

High Level Summary: Modifying the functions

$$\theta, \mathcal{D}_c \longmapsto (\theta_c, \int -\nabla \mathcal{E}(\theta(t), \mathcal{D}_c) d\mu$$

$$\theta, \mathcal{D}_c \longmapsto (M(\theta, x_i, d_i), \int \Big\| \frac{d^k x(t)}{d^k} \Big\|_2^2 dt)$$

$$\hat{Y}, \mathcal{D}_c \longmapsto (\text{SqrLoss}(\hat{Y}, \mathcal{D}_c), 0)$$

Admittedly, displaying the type signature of these functions would be clearer, as the top function takes a set as an input, while the lower function works on individual observations.

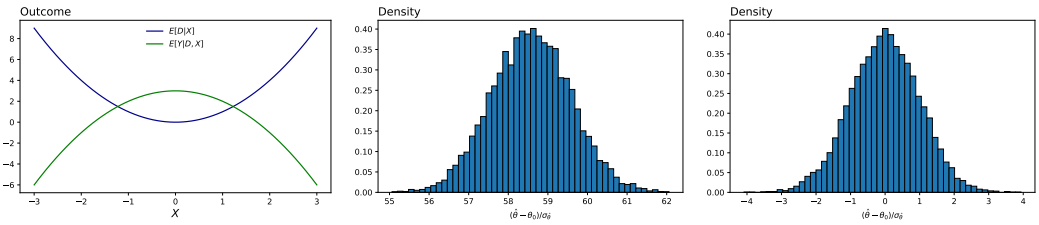


Figure 8: Reproduced Here

$$\mathcal{L}(\text{Model}, \text{Data}) = \underbrace{\mathcal{E}(\text{Model}, \text{Data})}_{\checkmark} + \mathcal{R}(\text{Model}, \text{Data})$$

```
def compose(f: Callable, g: Callable): Callable

    def g_after_f(x):
        a = f(x)
        b = g(a)
        return b

    return g_after_f
```

```
def compose_with_regularization(f: Callable, g: Callable): Callable

    def g_after_f(x):
        a, a_penalty = f(x)
        b, b_penalty = g(a)
        return b, (a_penalty + b_penalty)

    return g_after_f
```