Regularizing the Forward Pass

Patrick Power Shomik Ghosh

November 14, 2025

Abstract

In certain applied microeconomic settings, it's typical to view one's dataset as the realization of a stratified cluster randomized control trial: treatment is assigned at the cluster level and controls vary at both the individual and cluster level. *Locally*, this makes it more likely that observations will be from the same cluster and therefore changes how we should go about generalization. In this paper we introduce an estimation method accounts for this by partialing out non-parametric cluster effects via regularized bi-level gradient descent. We provide a python library based on JAX: https://github.com/pharringtonp19/rfp.

Keywords: Causal Inference, Deep Learning

1 Introduction

In many economic contexts, treatment varies at a level above the unit of observation. This is often desirable as it allows for estimation of partial equilibrium effects. We see this in housing, for example, when we want to understand the extent to which landlords pass the costs of eviction prevention policies onto the unhoused. This context also arises in educational settings, for instance, when we're interested in the extent to which private schools respond to the expansion of school vouchers.

This research design is also prevalent in Economics not because it is necessarily desired but rather because the only variation of the treatment is above the unit of interest. For example, when we're studying the impact of medicaid on mortality, Wyse and Meyer [2025] compare individuals in states which expanded to those in states which did not. The primary motivation wasn't to capture "spillover" effects but rather because state level variation is the most prominent source of variation in medicaid post 2010.

As we'll explain later in the paper, control variables which vary at the level of treatment are often included in regression specification in this setting to reduce selection bias. For example, in analyzing the impact of an eviction prevention policy which is initially rolled out in to zip codes with generally higher counts of eviction (as shown in Figure 1), it is reasonable control for the aggregate count of evictions at the zip code level. Doing so would improve the credibility of our assumption that the treatment can be thought of as *locally* randomly assigned.

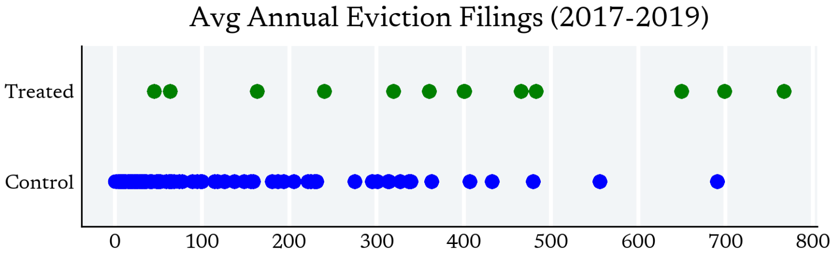


Figure 1: Assessing balance across covariate which variates at the cluster (and treatment) level

As Figure 2 illustrates however, when treatment is assigned at the cluster level and controls vary at both the individual and cluster level, it can make it more likely that *local* observations within the treatment/control group will be from the same cluster. Nonparametric methods which generate predictions via locally averaging don't account for this and as Section 5 highlights, can overfit to non-parametric cluster effects in this setting.

The central challenge that this paper explores is how to make local corrections in high dimensions to account for the cluster aspect of the data. As Bengio et al. [2000] notes, "in high dimensions, it is crucial to distribute probability mass where it matters rather than uniformly in all directions around each training point." Building from Finn et al. [2017]

and Domingos [2020], we propose training models via bi-level gradient descent (instead of gradient descent) which allows the model to partial out these non-parametric effects over the inner-gradient which adopt the model specifically to each cluster.

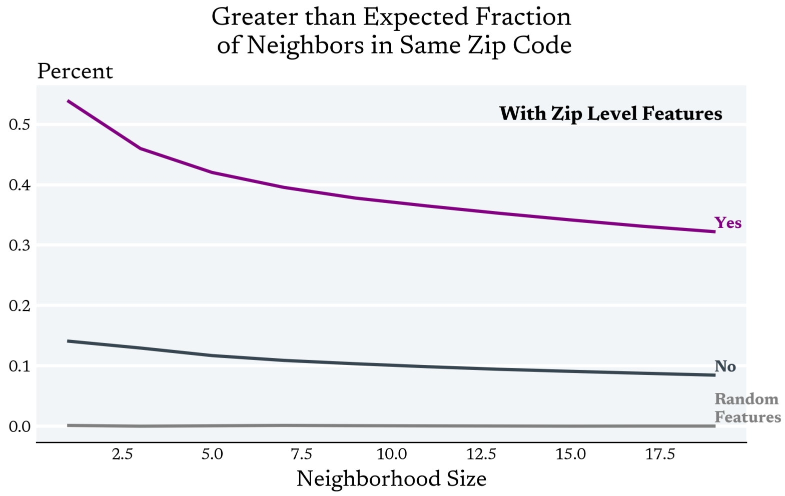


Figure 2: Within the control group, adding cluster level features increases the likelihood that the local neighborhood is overpopulated with observations from the same cluster.

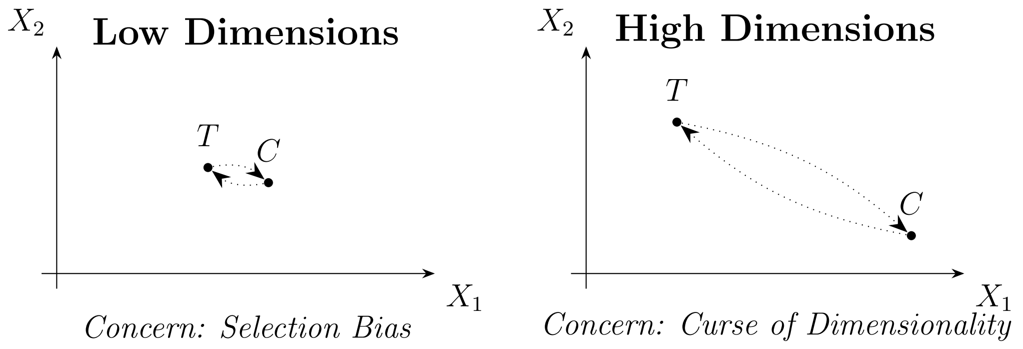
The remainder of the paper is organized as follows. Section 2, reviews the fundamental essence of causal inference, a necessary background before Section 3 illustrates how this context forces us to rethink the problem of generalization. Section 4 describes the estimation procedure and explains why this approach is reasonable approach. Section 5 compares our method to two standard baselines to confirm the intuition behind our approach. Section 6 reinforces the high level takeaways.

Throughout the paper we'll assume that the reader is familiar with the terms *potential outcomes*, *selection bias*, *probability spaces* and *kernel methods*. An accompanying set of notes which reviews these concepts can be found here.

2 The Essence of Causal Inference

Causal inference can be understood at three distinct levels. At the individual level, it's a missing data problem. We observe at most one of the potential outcomes. At the population level, the level at which most economists consider their context, it's a problem of selection bias. Can you find a set of controls such at treatment can be thought of as locally independent of (differenced) potential outcomes, $(\Delta)\tilde{Y}_i(1)$. In the finite sample, the problem is one of generalization – how do you use observations in the (control) treated group to predict the counterfactual outcomes for observations in the (treated) control group.

As figure 3 highlights, with few control variables (*Low Dimensions*), the data points between the treatment and control group are more similar with respect to observed features, yet the prediction problem is difficult because we are concerned that treatment and control observations might differ unobserved ways due to Selection Bias. Adding more control variables (*High Dimensions*) makes the observations less likely to differ in unobservable ways, but the prediction problem remains challenging because now observations differ in observed ways due to the Curse of Dimensionality.


Low Dimensions

High Dimensions

Concern: Selection Bias

Concern: Curse of Dimensionality

Figure 3: In the finite sample, the reason for why predicting counterfactual outcomes can be difficult depend on the dimensionality of the control variables.

As we highlighted at the beginning of the paper either because of an interests in partial equilibrium effects (or out of necessity), it's typical to have variation across in the density of the treatment. Certain zip code or states get treated. Having variation in the density of treatment across space makes it more difficult to local variation of treatment, and to have a low dimensional set of controls. A more complete understanding of the essence of causal inference would capture the fundamental tradeoff between each of these components.

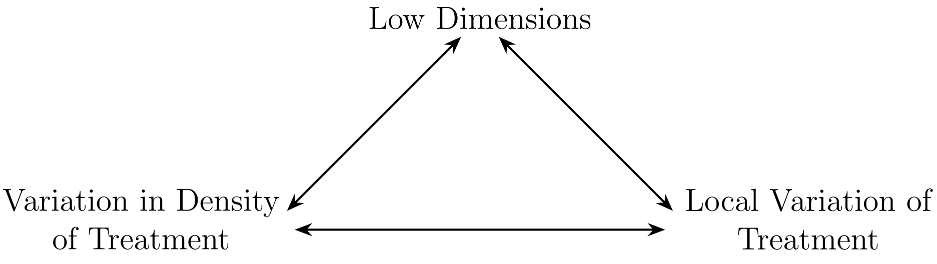

Low Dimensions
Variation in Density of Treatment
Local Variation of Treatment

Figure 4: Illustration of the relationship between dimensionality, density of treatment, and local treatment variation.

2.1 Formal Identification

It's important to note that the selection on observable identification assumptions remains unchanged when treatment is assigned at the cluster level and control variable vary at both the cluster an individual levels. To see this, we'll walk through the ideas of the conditional expectation function, conditional independence, and how they relate to the selection on observable assumption.

2.1.1 Conditional Expectation

To do so, let's define (using curried function notation) a conditioning function as follows which takes a probability measure, an event with positive probability and returns as new probability measure.

$$\mathcal{C} : \mathcal{M} \to \mathcal{F}_* \to \mathcal{M} \tag{1}$$

And let's define the expectation function as taking a probability measure, an integrable function and returns a real number.

$$E : \mathcal{M} \to \mathcal{H}_* \to \mathcal{R} \tag{2}$$

We can compose these functions:

$$E \circ \mathcal{C} : \mathcal{M} \to \mathcal{F}_* \to \mathcal{H}_* \to \mathcal{R}, \quad (E \circ \mathcal{C})_{\mathbb{P},B}(f) = \int f d\mathbb{P}_B \tag{3}$$

Now, we can extend this conditioning function, by conditioning on elements of the sample space.

$$\tilde{\mathcal{C}}_{\sigma(X)} : \mathcal{M} \to \Omega \to \mathcal{M}, \quad \forall B \in \sigma(X), \int_B \tilde{\mathcal{C}}_{\sigma(X),\mathbb{P},\cdot}(A) d\mathbb{P} = \mathbb{P}(A \cap B) \tag{4}$$

And now composing this augmented conditional expectation function with the expectation function produces the standard conditional expectation function.

$$E \circ \tilde{\mathcal{C}}_{\sigma(X)} : \mathcal{M} \to \Omega \to \mathcal{H}_* \to \mathcal{R} \tag{5}$$

2.1.2 Independence

In this subsection, we'll build up to what it means for two random variables to be conditional independent of a third random variable. We can start though by defining independence as follows.

$$Y \perp X \equiv \forall A, \forall B \in \sigma(X), \sigma(Y), \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \tag{6}$$

For conditional independence with respect to an event, W, we can replace the probability measure $\mathbb{P}$ with the conditional probability measure defined above: $\mathcal{C}_{\mathbb{P},W}(\cdot)$.

$$Y \perp X | W \equiv \forall A, \forall B \in \sigma(X), \sigma(Y), \quad \mathcal{C}_{\mathbb{P},W}(A \cap B) = \mathcal{C}_{\mathbb{P},W}(A)\mathcal{C}_{\mathbb{P},W}(B) \tag{7}$$

And finally, as the pattern suggests, we express conditional independence with respect to a random variable as follows using the notion of the augmented conditional probability measure:

$$\forall A, \forall B, \forall U \in \sigma(X), \sigma(Y), \sigma(W), \quad \int_U \tilde{\mathcal{C}}_{\sigma(W),\mathbb{P},\cdot}(A \cap B)\, d\mathbb{P} \tag{8}$$

$$= \int_U \tilde{\mathcal{C}}_{\sigma(W),\mathbb{P},\cdot}(A)\, d\mathbb{P} \int_U \tilde{\mathcal{C}}_{\sigma(W),\mathbb{P},\cdot}(B)\, d\mathbb{P} \tag{9}$$

2.1.3 Selection On Observables

A preliminary challenge in discussing identification is clarying what we mean by the term. A recent paper on the topic (Lewbel [2019]) writes, “Econometric identification really means just one thing: model parameters or features being uniquely determined from the **observable population** that generates the data.” . With clustered data, though, the **observable population** is not well defined.[1]

The way to think about identification is within a probability space framework. In every empirical paper there are two probability spaces of interest. The first is at the population level where the sample space is the collection of all people (or firms) of interest and the random variable in this context is the individual level treatment effects.

$$\left(\Omega, \mathcal{F}, \mathbb{P}\right) \xmapsto{\tau} \left(\mathcal{R}, \mathcal{B}(\mathcal{R}), \mathbb{P} \circ \tau^{-1}\right) \tag{10}$$

The second probability space has a sample space of all possible datasets, and where the random variable of interest is the estimator. Note, we don't assume the data is $i.i.d$ which is why we denote the n in the subscript on the sample probability measure instead of as an exponent which would suggest a product measure.

$$\left(\Omega_n, \mathcal{F}_n, \mathbb{P}_n\right) \xmapsto{\mathcal{A}} \left(\mathcal{R}, \mathcal{B}(\mathcal{R}), \mathbb{P} \circ \mathcal{A}^{-1}\right) \tag{11}$$

With this setup, we can know think of identification in terms of mathematical objects rather than in undefined notions such as the **observable population**. In can be *roughly* argued that every identification strategy is based on the idea of treatment being conditional independent of a potential outcome. We can now express this conditional independence with respect to the sample probability measure: $\mathbb{P}_n$

$$\tilde{Y}_i(1) \perp_{\mathbb{P}_n} D_i | X_i, \quad \Delta\tilde{Y}_i(1) \perp_{\mathbb{P}_n} D_i | X_i \tag{12}$$

Given our previous work, we understand this this assumption formally means the following.

$$\forall A, \forall B, \forall U \in \sigma(\tilde{Y}_i(1)), \sigma(D_i), \sigma(X_i), \quad \int_U \tilde{\mathcal{C}}_{\sigma(X_i),\mathbb{P}_n,\cdot}(A \cap B)\, d\mathbb{P}_n \tag{13}$$

$$= \int_U \tilde{\mathcal{C}}_{\sigma(X_i),\mathbb{P}_n,\cdot}(A)\, d\mathbb{P}_n \int_U \tilde{\mathcal{C}}_{\sigma(X_i),\mathbb{P}_n,\cdot}(B)\, d\mathbb{P}_n \tag{14}$$

[1] We bold the text to make the expression jump out to the reader

Under this assumption the following conditional expectation function

$$E \circ \tilde{\mathcal{C}}_{\sigma(X),C_{\mathbb{P}_{n,\{D_i=1\}}},\cdot}(Y_i) \overset{a.s}{=}_{\sigma(X_i)} E \circ \tilde{\mathcal{C}}_{\sigma(X),C_{\mathbb{P}_n},\cdot}(\tilde{Y}_i(1)) \tag{15}$$

3 Rethinking Generalization

It's worth considering why this *issue* isn't discussed more frequently in empirical work. Framed another way, why wouldn't a panel data setting remove this issue when one can substitute a cluster level feature with a cluster fixed effect.

To address this point, and to keep things concrete, let's extend the example mentioned above regarding evictions. Let's continue with the idea that a subset of zip codes in a state adopt an eviction prevention policy and but now extend our dataset so that observe individual level both before and after the policy rollout. The typical regression model for this context, then has the following form, where we regress individual level outcomes on zip code and time fixed effects and individual level controls.

$$Y_{izt} = \tau I\{t \geq t_*\} + \gamma_z + \gamma_t + X_{ist} + \varepsilon_{izt}$$

Via the Frish-Waugh-Lovel, theorem, the coefficient of interest, τ, is the same as in the following single variable regression where the right hand side variable of interest is the difference between the indicator for treatment and the predicted indicator of treatment where the prediction is based on a linear model.

$$Y_{izt} = \tau(I\{t \geq t_*\} - \hat{I}) + v_{izt}, \quad \hat{I} = \phi_z + \phi_t + X_{izt}$$

This residual variation goes to zero though if we were to substitute a non-parametric conditional expectation function for the linear model.

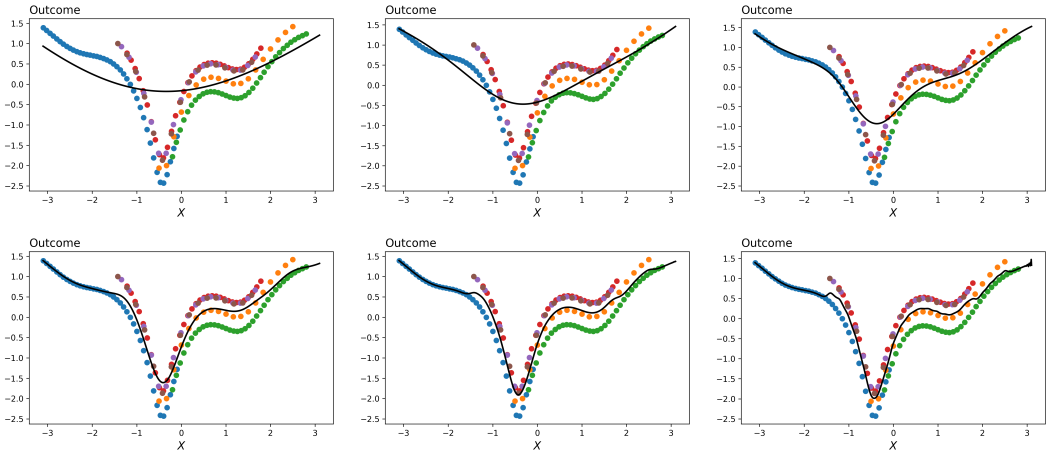
$$Y_{izt} = \tau \underbrace{(I\{t \geq t_z^*\} - \mathbb{E}[I|s, t, X_{izt}])}_{=0} + \eta_{izt}$$

To address our original question this problem is not typically highlighted in empirical work because papers are exploiting function form induced variation. That is, they're exploiting variation in the treatment which only exists because they're fitting a linear model. If they had relaxed the function space of their models, the variation would go to zero.

Exploiting functional form induced variation seems problematic. To avoid this problem, one can control not for a cluster fixed effect, but rather a cluster level feature. Hence the focus of our paper.

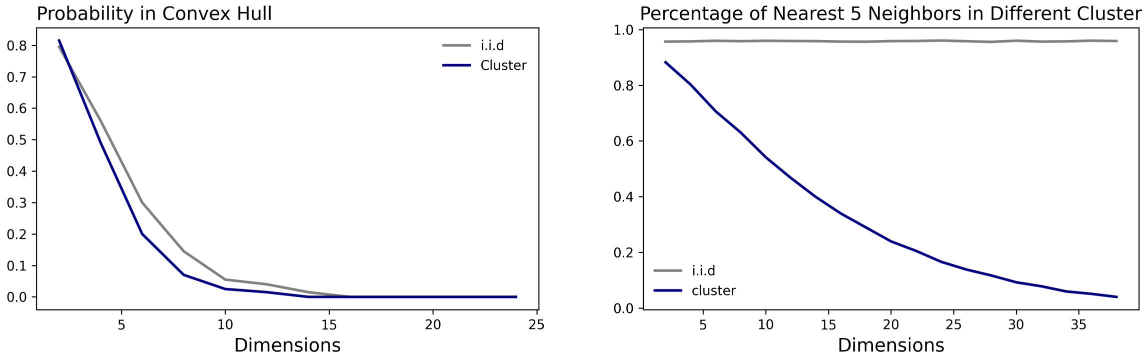
The typical way to balance these tradeoff is, again as referenced bfore, to fit models wich control for cluster level features. Figure 5 captures the behavior of a typical smoothing estimator in this context. A standard nonparametric model (black line) with a "smoothing" hyperparameter struggles: in order to fit the 'v'-shaped component of the data where there is general consensus across the clusters, the bandwidth of the estimator must be small. In doing so, though, it over fits the tails. Intuitively what's needed in this context is an estimator that is "locally" aware of the cluster structure of the data.

Figure 5: The Tragic Triad of Clustered Data



Reproduced Here: In this figure, we assume away within-cluster variation. Each dot corresponds to an observation. The different colors highlight the various clusters.

It's interesting to note that these issues are perhaps only magnified as we increase the dimensionality of the data. Extending the work of Balestriero et al. [2021], we illustrate in figure 7a that clustered sampling doesn't change the fundamental issue of learning in high dimensions (extrapolation) so much as it motivates us to reconsider how we go about it, as highlighted in figure 7b.



(a)

(b)

Figure 7: (a) Probability that a observation in the test set is in the convex hull formed by the training set (b) Fraction of the five nearest neighbors in a different cluster. Data consists of 25 cluster and 25 observations per cluster. Clusters differ only in the mean which is drawn from an isotropic gaussian distribution: Reproduced Here and Here

We highlight in figure 8 that subject to the typical caveats of hyperparameter tuning, our model fits the 'v'-shaped nature of the data where there is consensus across the clusters without overfitting in the tails." That is, our model, which we formally define in the next section, is implicitly locally aware of the cluster structure of the data.

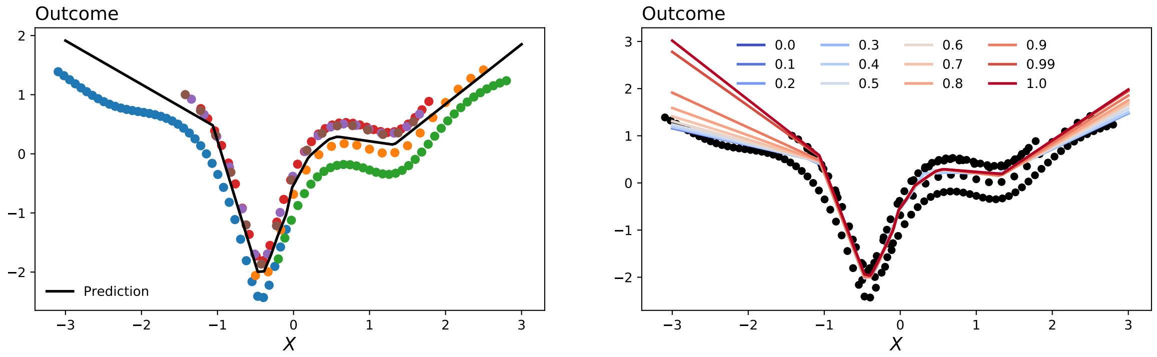


Figure 8

4 Method

4.1 Parameter Gradients

As we alluded to previously, intuitively we would like the model to locally partial out the cluster effects. The challenge is that local methods don't work in high dimensions. The "success" of neural networks in high dimensions is there ability to learn "feature representations". So in some sense, our aim is to partial out the cluster effects from this learnt feature representation.

Our approach is motivated in part by Domingos [2020] which illustrates how neural networks can be understood within the framework of kernel machines. Specifically, Domingos [2020] shows that in the gradient flow regime, neural networks can be understood as kernel machines where the kernel is formed by integrating the inner product of gradient of the neural network evaluated at the corresponding points of the domain. That is, the learnt similarity is a function of the similarity of the parameter gradients.

$$K_\theta(x, x') = \int_0^1 \Big\langle \nabla_\theta f_x\big(\theta(t)\big), \nabla_\theta f_{x'}\big(\theta(t)\big) \Big\rangle dt$$

Motivated by this observation, we propose a method which partialls out the non-parametric cluster effects in a way by augmenting the parameter gradients.

4.2 Proposed Method

Building off the popular meta-learning algorithm MAML (Finn et al. [2017]), we train a neural network via bi-level gradient descent. In the inner level, we allow the parameters of the neural network to adopt to each cluster in parallel. By adding a regularization penalty which differentiates our approach from MAML, the model intuitively learns a representation which has a low prediction loss, but where given a few steps of gradient descent can significantly improve its loss on any cluster. In this way, we partial out the potentially nonparametric effects.

We can express our objective function as the weighted average of a loss function evaluated at two different parameter values.

$$R_k(\theta) := \frac{1}{N}\sum_i \alpha L_i(\theta) + (1-\alpha)L_i(\theta_c^k(\theta)) \tag{16}$$

The first term captures the typical squared prediction loss. Here we assume that the outcome is a real-valued, but the same approach works for discrete outcomes as well.

$$L_i(\theta) := (y_i - f(\theta, x_i))^2 \tag{17}$$

The second term contains the same loss function, but the parameter corresponds to the learnt parameter based on k steps of gradient descent on observations from the same cluster.

$$\theta_c^k(\theta) := \theta^{k-1} - \alpha\nabla_\theta \frac{1}{n_c}\sum_{i\in c} L_i(\theta^{k-1}), \quad \theta^0 = \theta \tag{18}$$

Together, this can be understood as a regularized version of the popular metal-learning algorithm MAML. We highlight the importance of the regularization parameter in the results section.

5 Results

We illustrate the relative importance of our design choices in figure **??**. First, standard methods ignore the clustered nature of the data and therefore tend to overfit in the tails where the local observations are from the same cluster. Second, MAML - which is bi-level gradient descent with out the regularization terms - learns parameters values from which it can "quickly" minimize the cluster specific loss values. This learnt initialization though has no guarantee of in sample performance. As figure **??** illustrates, MAML essentially ignores one of the clusters. Importantly though, this issue isn't apparent when there are multiple clusters with significant overlap. Therefore, the regularization parameter appears important when the clusters don't overlap significantly, which we showed previously can occur with clustered data.

6 Conclusion

Why do we have estimators? It's because we cannot draw in high dimensional spaces.

In this paper we show that clustered treatment assignment together with cluster level controls can increase the variance of standard smoothing estimators. We illustrate that a regularized version of the popular meta-learning algorithm (MAML) is a potentially attractive estimation method in this context in that it partials out the potentially nonparametric cluster effects via bi-level gradient descent. Intuitively, this is akin to early stopping at the cluster level.

References

Randall Balestriero, Jerome Pesenti, and Yann LeCun. Learning in high dimension always amounts to extrapolation. *arXiv preprint arXiv:2110.09485*, 2021.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

Pedro Domingos. Every model learned by gradient descent is approximately a kernel machine. *arXiv preprint arXiv:2012.00152*, 2020.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Arthur Lewbel. The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4):835–903, 2019.

Angela Wyse and Bruce D. Meyer. Saved by medicaid: New evidence on health insurance and mortality from the universe of low-income adults. NBER Working Paper 33719, National Bureau of Economic Research, 5 2025. URL https://www.nber.org/papers/w33719.