# Regularizing the Forward Pass

Patrick Power Shomik Ghosh

November 16, 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 the tradeoffs associated with fitting models to finite date. In this paper we introduce an estimation method which accounts for the cluster nature of the data via regularized bi-level gradient descent. We provide a python library based on JAX: https://github.com/pharringtonp19/rfp.

**Keywords:** Clusters, 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 shows up in educational settings 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 because the only variation of the treatment is above the unit of interest. For example, when studying the impact of medicaid on mortality, Wyse and Meyer [2025] compare individuals in states which expanded Medicaid to those in states which did not. The primary motivation behind the research design isn't to capture "spillover" effects but rather to leverage the most prominent source of medicaid expansion in the years following 2010.

To reduce selection bias in this setting, it's common to include control variables which vary at the level of treatment (above the unit of observation). For example, in analyzing the impact of an eviction prevention policy that is targeted at certain zip codes with generally higher eviction counts (Figure 1) it is reasonable to include aggregate zip code evictions as a control. Doing so improves the credibility of the assumption that the treatment can be thought of as *locally* randomly assigned.

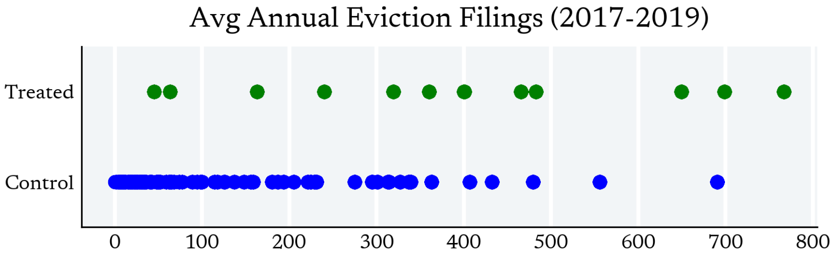


Figure 1: Assessing balance across covariate which variates at the cluster (and treatment) level

As Figure 2 illustrates however, when treatment is assigned at the cluster level and controls vary at both the individual and cluster level, it is 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, they 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 nature of the data in these settings. 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 on Finn et al. [2017] and Domingos [2020], we propose training models via regularized bi-level gradient descent (instead of gradient descent) which partials out these non-parametric cluster effects by allowing the model to adopt to each specific cluster in parallel during the inner gradient step.

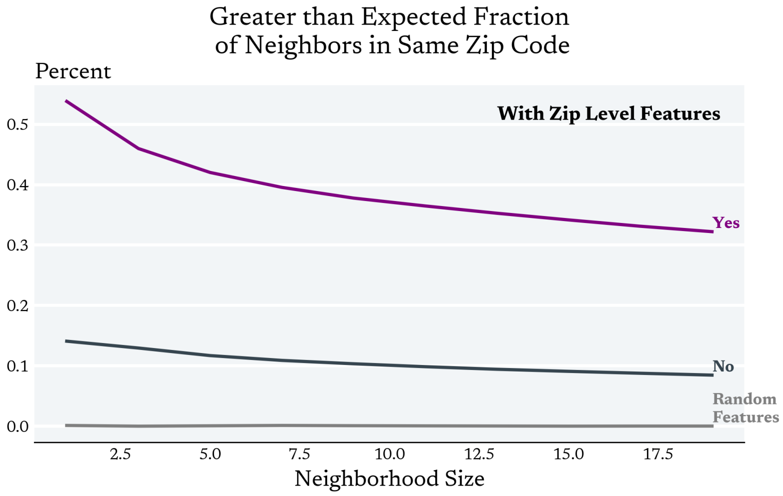


Figure 2: Within the control group, adding cluster level features increases the likelihood that the local neighborhood is overpopulated with observations from the same cluster.

The remainder of the paper is organized as follows. Section 2 illustrates how to think about Identification in this setting. Section 3 covers how cluster data changes the tradeoffs associated with fitting models to finite data. Section 4 describes the proposed estimation procedure. Section 5 compares the proposed method against two standard baselines to confirm the intuition behind our approach. Section 6 conclus.

Before moving on, it's worth pausing to consider why this "local cluster data issue" isn't discussed more frequently in empirical work. Framed another way – does this problem exist in a panel data setting where fixed effects can be substituted for the cluster level features.

To address this point, and to keep things concrete, let's extend the example regarding an eviction prevention policy which is initially targeted at certain zip codes which generally had higher levels of eviction counts. Let's consider a panel dataset which includes individual level observations both before and after the policy rollout. In this context, individual level outcomes ($Y$) are typically regressed on zip code ($z$) and time ($t$) fixed effects and individual level controls ($X$).

$$Y_{izt} = \tau I\{t \geq t_*\} + \gamma_z + \gamma_t + X_{ist} + \varepsilon_{izt} \tag{1}$$

Via the Frish-Waugh-Lovel, theorem, the coefficient of interest, $\tau$, in Equation (1) is the same as the coefficient in the single variable regression in Equation (2) where the right hand side variable of interest is the difference between the indicator for treatment and the predicted indicator of treatment.

$$Y_{izt} = \tau(I\{t \geq t_*\} - \hat{I}) + v_{izt}, \quad \hat{I} = \phi_z + \phi_t + X_{izt} \tag{2}$$

This residual variation goes to zero, though, if we were to substitute a non-parametric conditional expectation function for the linear model in the prediction step because treatment is a function of the zip code ($z$) and time ($t$).

$$Y_{izt} = \tau \underbrace{(I\{t \geq t^*\} - \mathbb{E}[I|z,t,X_{izt}])}_{=0} + \eta_{izt} \tag{3}$$

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

In some regards, central aim of this paper isn't to present a novel estimation technique. As we'll highlight later in the paper, our method has its own drawbacks. It introduces additional hyperparameters which results can be sensitive to. Furthermore, the aim of this paper isn't to label the current approach as wrong. Exploiting functional form induced variation seems problematic, but maybe in a specific context, it's the best thing to do. Rather, the aim of this paper is to present a conceptual framework so that the applied researcher working on causal inference problems with clustered data can better evaluate what the right tradeoffs to make are in their context.

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

## 2 Identification

Causal inference can be understood at three distinct levels. In this section, we address the first two. At the individual level, causal inference is 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. Is there a set of controls such that treatment can be thought of as locally independent of the potential outcomes.

It's worthwhile to clarify the necessary assumptions for identification in this setting where treatment is assigned at the cluster level and control variable vary at both the cluster an individual levels. To do so, we'll formally define the conditional expectation function, conditional independence, and thereby have the necessary components to discuss selection on observables with cluster data.

Let's define (using curried function notation) a conditioning function as follows which takes a probability measure, an event with positive probability as inputs and returns as new probability measure. The subscript $*$ indicates that the feasible event space, $\mathcal{F}$, depends on the probability measure selected from $\mathcal{M}$.

$$\mathcal{C} : \mathcal{M} \to \mathcal{F}_* \to \mathcal{M} \tag{4}$$

Let's further define the expectation function as taking a probability measure and an integrable function as inputs and returning a real number. Again, the subscript $*$ denotes the fact the set of measurable functions depends on the probability measure selected from $\mathcal{M}$.

$$E : \mathcal{M} \to \mathcal{H}_* \to \mathcal{R} \tag{5}$$

We can compose these functions, where now the subscripts, $\mathbb{P}$ and $B$, denote the partial application of a function.

$$E \circ \mathcal{C} : \mathcal{M} \to \mathcal{F}_* \to \mathcal{H}_* \to \mathcal{R}, \quad (E \circ \mathcal{C})_{\mathbb{P},B}(f) = \int f d\mathbb{P}_B \tag{6}$$

We can augment this conditioning function, so that we can "conditioning" on elements of the sample space instead of just events by adopting the signature of the function and imposing a constraint. Here we subscript the augmented conditioning function with the $\sigma$-algebra generated by the random variable $X$.

$$\tilde{\mathcal{C}}_{\sigma(X)} : \mathcal{M} \to \Omega \to \mathcal{M} \tag{7}$$

$$\forall A \in \mathcal{F}, \forall B \in \sigma(X), \int_B \tilde{\mathcal{C}}_{\sigma(X),\mathbb{P},\cdot}(A) d\mathbb{P} = \mathbb{P}(A \cap B) \tag{8}$$

We can then compose this augmented conditional function with the expectation function to produce the standard conditional expectation function.

$$E \circ \tilde{\mathcal{C}}_{\sigma(X)} : \mathcal{M} \to \Omega \to \mathcal{H}_* \to \mathcal{R} \tag{9}$$

$$\forall B \in \sigma(X), \int_B E \circ \tilde{\mathcal{C}}_{\sigma(X),\mathbb{P},\cdot}(Y)\, d\mathbb{P} = \int_B Y\, d\mathbb{P} \tag{10}$$

A preliminary challenge in discussing identification, though, is to clarify 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) and the random variable of interest is the individual level treatment effects.

$$(\Omega, \mathcal{F}, \mathbb{P}) \xmapsto{\tau} (\mathcal{R}, \mathcal{B}(\mathcal{R}), \mathbb{P} \circ \tau^{-1}) \tag{11}$$

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

The second probability space has a sample space of all possible datasets, and 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 probability measure instead of as an exponent which would suggest a product measure.

$$\left(\Omega_n, \mathcal{F}_n, \mathbb{P}_n\right) \xmapsto{\mathcal{A}} \left(\mathcal{R}, \mathcal{B}(\mathcal{R}), \mathbb{P} \circ \mathcal{A}^{-1}\right) \tag{12}$$

Every identification strategy in Economics is based on the idea of treatment being conditional independent of a potential outcome. We can express this in one line with respect to the sample probability measure, $\mathbb{P}_n$, as follows.

$$\tilde{Y}_i(1) \perp_{\mathbb{P}_n} D_i | X_i, \quad \Delta\tilde{Y}_i(1) \perp_{\mathbb{P}_n} D_i | X_i \tag{13}$$

Formally, this can be understood in terms of the augmented conditioning function defined previously. When treatment varies at the cluster level, and covariates vary at both the cluster and individual level, the validity of selection on observables assumption depends on whether the augmented conditioning function "products out".

$$\forall A, \forall B, \forall U \in \sigma(\tilde{Y}_i(1)), \sigma(D_i), \sigma(X_i), \quad \int_U \tilde{\mathcal{C}}_{\sigma(X_i),\mathbb{P}_n,\cdot}(A \cap B)\, d\mathbb{P}_n \tag{14}$$

$$= \int_U \tilde{\mathcal{C}}_{\sigma(X_i),\mathbb{P}_n,\cdot}(A)\, d\mathbb{P}_n \int_U \tilde{\mathcal{C}}_{\sigma(X_i),\mathbb{P}_n,\cdot}(B)\, d\mathbb{P}_n \tag{15}$$

Under this assumption, the conditional expectation function defined with respect to the sampling probability measure is equal to the conditional expectation function defined with respect to the population probability measure.

$$E \circ \tilde{\mathcal{C}}_{\sigma(X),C_{\mathbb{P}_{n,\{D_i=1\}}},\cdot}(Y_i) \overset{a.s}{=}_{\sigma(X_i)} E \circ \tilde{\mathcal{C}}_{\sigma(X),C_{\mathbb{P}},\cdot}(\tilde{Y}_i(1)) \tag{16}$$

# 3 Generalization

The third level of causal inference is the finite sample. Here the problem is one of generalization – how best leverage the observations in the (control) treated group to predict the counterfactual outcomes for observations in the (treated) control group. It is this third level – the finite – sample, which deserves further consideration.

As figure 3 highlights, with few control variables (*Low Dimensions*), the data points in the treatment and control group are typically similar with respect to observed features, yet the prediction problem is often difficult because the treatment and control observations might differ in 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, again, can remain challenging because observations differ in observed ways due to the Curse of Dimensionality.

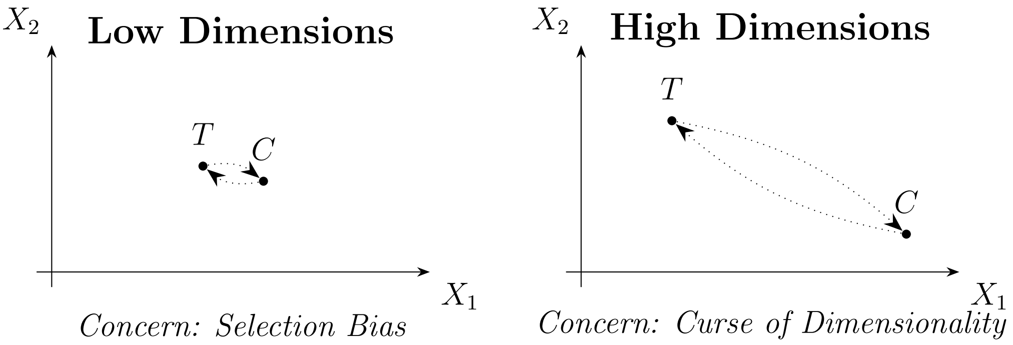


Figure 3: In the finite sample, the challenge of predicting counterfactual outcomes depends on the dimensionality of the control variables.

Working with clustered data adds a third "dimension" to this tradeoff. As highlighted at the beginning of the paper, either because of an interests in partial equilibrium effects or out of necessity, it's common to have variation in the density of the treatment: only certain zip code or states get treated for example. This makes it more difficult to exploit local variation of treatment and to condition on a low dimensional set of controls because there is no within cluster variation of the treatment and conditioning on a cluster level feature expands the control set. Figure 4 illustrates highlights the tradeoffs associated with these competing aims.

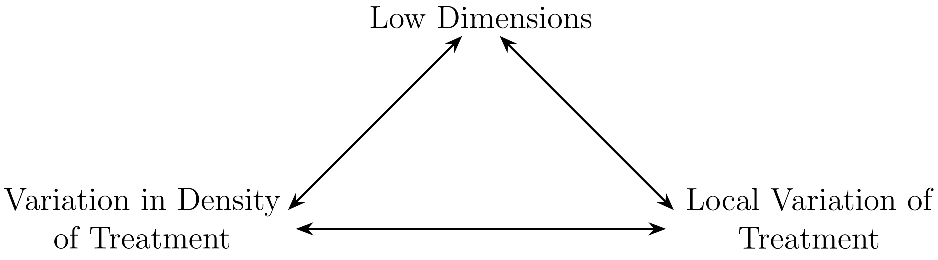
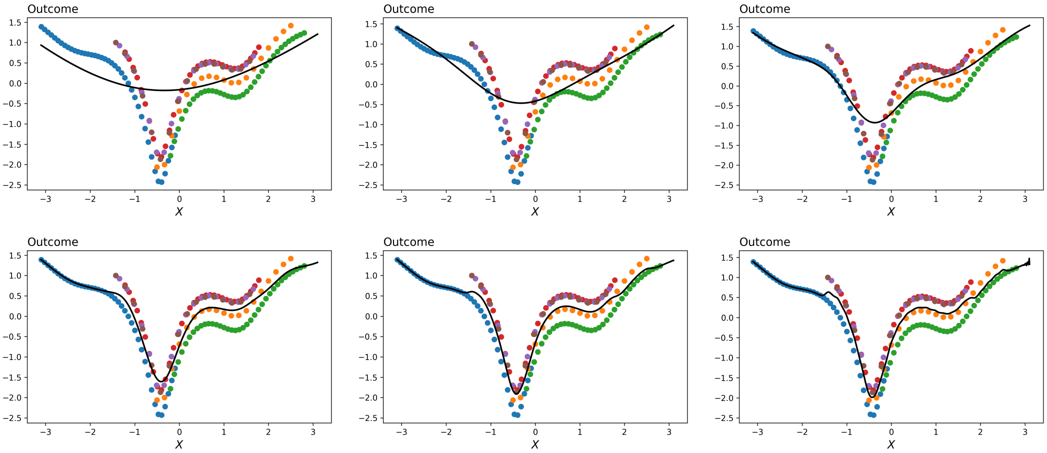

Low Dimensions
Variation in Density of Treatment
Local Variation of Treatment

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

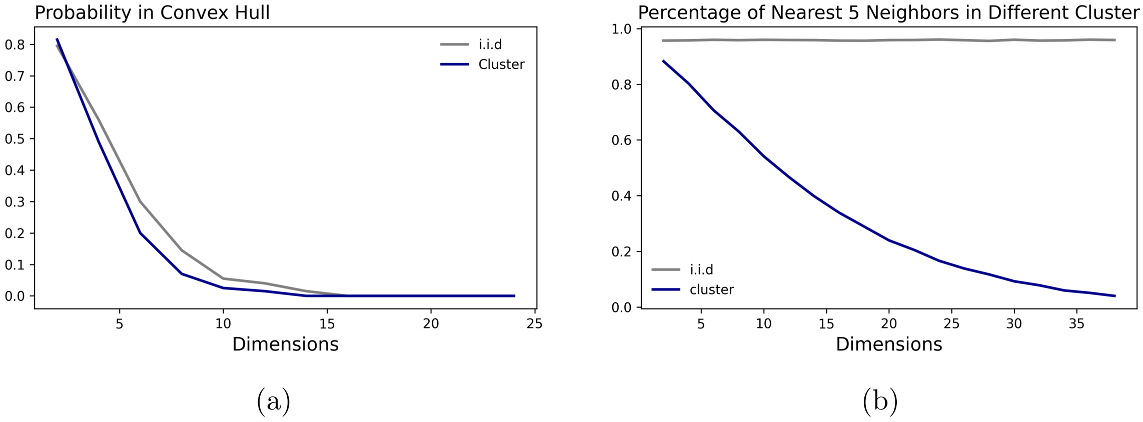
The typical way to balance these tradeoff is, again as referenced bfore, to fit models wich control for cluster level features. Figure 5 captures the behavior of a typical smoothing estimator in this context. A standard nonparametric model (black line) with a "smoothing" hyperparameter struggles: in 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: 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, described fully in Section 4 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.

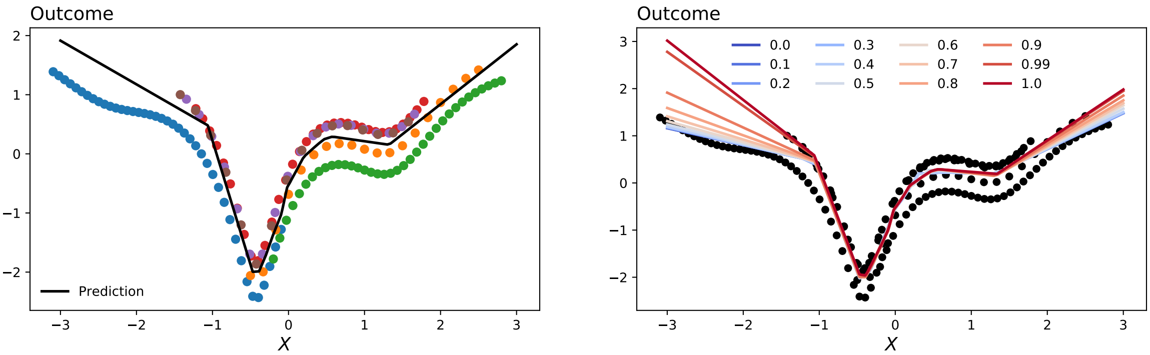


Figure 8

# 4 Method

## 4.1 Model

Our method applies to any parameterized model trained via gradient descent. For instance, Feed-forward neural networks or large-language models would be suitable candidates.

$$f : \Theta \to \mathcal{X} \to \mathcal{R} \tag{17}$$

## 4.2 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 averaging methods don't work in high dimensions. The "success" of neural networks in high dimensions is their ability to learn "feature representations". Our aim, therefore, 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 thought of as a kernel method where the kernel is formed by integrating the inner product of gradient of the neural network evaluated at the corresponding points of the domain along the training path formed by gradient descent. 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 partials out the non-parametric cluster effects by augmenting the parameter gradients.

### 4.3 Proposed Method

Building off the popular meta-learning algorithm MAML (Finn et al. [2017]), we train a neural network via regularized bi-level gradient descent. In the inner level, we allow the parameters of the model 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 fit to 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. An important hyperparameter is $\alpha$ which balances these two loss functions. Setting $\alpha$ equal to zero is equivalent to the MAML algorithm.

$$R_k(\theta) := \frac{1}{N}\sum_i \alpha L_i(\theta) + (1-\alpha)L_i(\theta_c^k(\theta)) \tag{18}$$

The first term captures the typical squared prediction loss. The same parameter of the model is used to generate prediction across all observations.

$$L_i(\theta) := (y_i - f(\theta, x_i))^2 \tag{19}$$

The second term contains the same loss function, but evaluated on the last parameter value of $k$ steps of gradient descent taken only with respect to observations from the same cluster. That is, the parameters used by the model to generate predictions are cluster specific

$$\theta_c^k(\theta) := \theta^{k-1} - \alpha\nabla_\theta \frac{1}{n_c}\sum_{i\in c} L_i(\theta^{k-1}), \quad \theta^0 = \theta \tag{20}$$

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 Section 5.

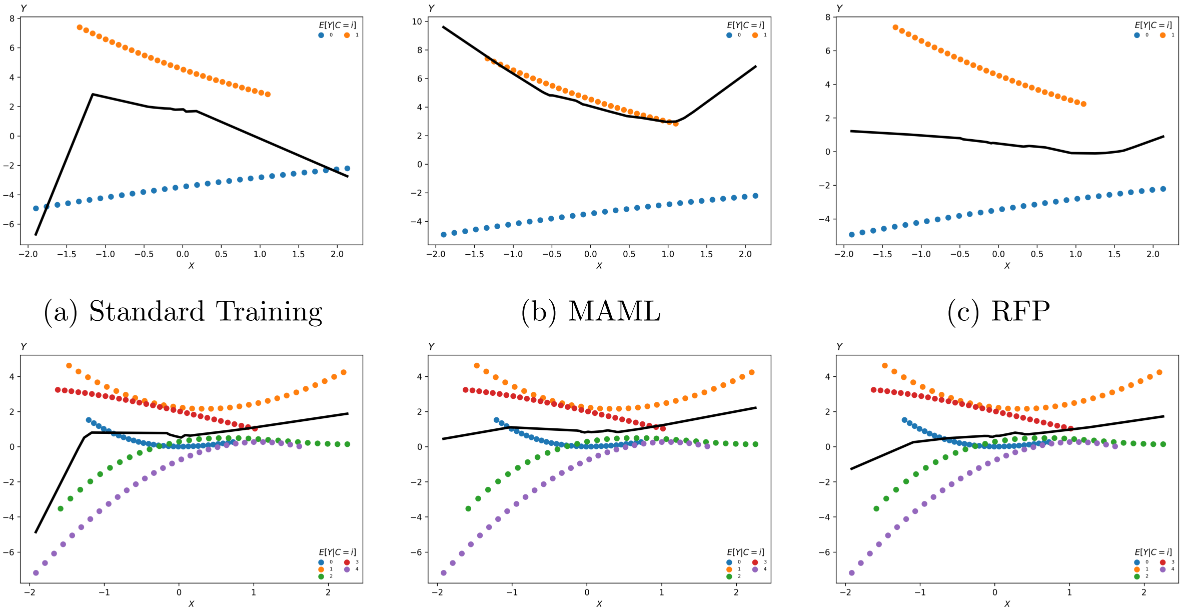
## 5 Results

We compare our method to two standard baselines: the popular meta-learning algorithm MAML and a locally adoptive estimator. In the two simulations, we allow the distribution over covariates and the conditional expectation functions to differ across clusters. Observations from the same cluster have the same color. The first simulation shown in Figures 9a, 9b, and 9c is meant to capture high dimensional space, where there is very little overlap across clusters in the finite sample. Figure 9d, 9e, and 9f show the results from the second simulation which illustrate the performance of these models in low-dimension space when there is more overlap across the clusters.

Across both simulations, it's evident that the locally adaptive method is highly sensitive to smallest discrepancies in overlap across the clusters. In “quasi”-high dimensional space shown in Figure 9a, the model sharply overfits the tails. In “quasi”-low dimensional space with greater overlap across the clusters, Figure 9d, the model's performance is better, but is still sensitive to slight deviations in the overlap. The model inherently lacks an ability to account for the cluster nature of the data.

At a high level, the MAML algorithm selects a parameter values such that after a few updates of gradient descent over only observations within the same cluster, the model has a good cluster specific fit. In low dimensional space, as shown in Figure 9e, this appears to be sufficient. With a high degree of overlap across the clusters and limiting the number of inner gradient steps, the model implicitly learns a reasonable in sample representation. In higher dimensions though, as shown in Figure 9b, MAML doesn't appear to learn a reasonable representation. It seems to find parameter values that allow it to fit one cluster almost perfectly and which after a few steps of gradient descent (not shown) can interpolate the other the observations from the other cluster.

There are many reasons why we use estimators, but a fundamental one is because we can't “draw” in high dimensions. The act of fitting a regression model to data is akin to asking the estimator to draw the conditional expectation function. As Figure 9c and 9f illustrate, this regularized bi-level gradient descent can (subject to hyperparameter tuning!) draw reasonable function across cluster data across in both high and low dimensions. And therefore suggests that it might be a preferred method relative to a locally adoptive model, or MAML when working with cluster data in high dimensions.



(a) Standard Training

(b) MAML

(c) RFP

(d) Standard Training

(e) MAML

(f) RFP

Figure 9: The colorful dots represent data from separate clusters. Each figure corresponds to fitting a neural network to this data under different training algorithms

# 6 Conclusion

Sometimes in the literature one comes across statements labeling certain econometric decisions as valid and others as invalid. In some cases, this is perfectly reasonable. For instance, one shouldn't condition on a bad control.

It is more often the case, though, in empirical settings though that there isn't a clear "right" approach. Working with quasi-experimental variation in attempt to answer an important question, (1) presents a broad set of possibly reasonable approaches and (2) the results likely differ across these approaches.

The best empirical work, therefore, doesn't attempt to convince someone that their approach or set of results are the only correct ones. The best work, involves investing time to understand the context – talking to stakeholders – and making reasonable judgments based on statistical intuition, the nature of the data generating process, and the parameter of interest.

Through a combination mathematically derivations, conceptual diagrams, and visually simplistic simulations, the aim in this paper is to present a conceptual framework so that the reader may better evaluate the tradeoffs when working with cluster data for their specific context.

# 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.