LEARNING WITH EXPLANATION CONSTRAINTS

Rattana Pukdee[1]*, **Dylan Sam**[1]*, **J. Zico Kolter**[1,2], **Maria-Florina Balcan**[1], **Pradeep Ravikumar**[1]

[1]Carnegie Mellon University, [2]Bosch Center for AI

ABSTRACT

While supervised learning assumes the presence of labeled data, we may have prior information about how models should behave. In this paper, we formalize this notion as learning from *explanation constraints* and provide a learning theoretic framework to analyze how such explanations can improve the learning of our models. For what models would explanations be helpful? Our first key contribution addresses this question via the definition of what we call EPAC models (models that satisfy these constraints in expectation over new data), and we analyze this class of models using standard learning theoretic tools. Our second key contribution is to characterize these restrictions (in terms of their Rademacher complexities) for a canonical class of explanations given by gradient information for linear models and 2 layer neural networks. Finally, we provide an algorithmic solution for our framework, via a variational approximation that achieves better performance and satisfies these constraints more frequently, when compared to simpler augmented Lagrangian methods to incorporate these explanations. We demonstrate the benefits of our approach over synthetic and real-world experiments.

1 INTRODUCTION

There has been a considerable recent focus on generating explanations of complex black-box models so that humans may better understand their decisions. But what if humans were able to provide explanations for how these models should behave? We are interested in the question of how to learn models given such apriori explanations. We believe that learning from explanations is a natural characterization for training machine learning models as it matches how humans learn. For example, we learn math much better and more efficiently from a (good) teacher, who can explain the underlying principles and rules. As labeled examples are provided by domain experts, we can also ask them to provide explanations for their decisions. This requires effort from the domain expert but can significantly improve the standard learning process, reducing the required number of labeled data.

In this paper, we provide an analytical framework for learning from explanations. We first provide a mathematical framework for model constraints given explanations. Casting explanations as functionals g that take in a model h and input x, we can represent domain knowledge of how models should behave as constraints on the values of such explanations. We can leverage these to then solve a constrained ERM problem where we additionally constrain the model to satisfy these explanation constraints. From an analysis standpoint, this poses challenges as these constraints are random; the explanations and constraints are provided on randomly sampled inputs. To handle these stochastic constraints, we draw from classical approaches in stochastic programming (Kall et al., 1994; Birge & Louveaux, 2011). In particular, we formalize the class of what we term *EPAC* models, or models that satisfy the explanation constraints (in expectation) up to some slack with high probability. Here, the probability is with respect to the randomness of the models themselves. The high level idea is that any model that satisfies the set of explanation constraints on the finite sample can, via standard statistical learning theoretic arguments (Valiant, 1984), can be shown to satisfy the constraints in expectation up to some slack with high probability. Then, we can capture the benefit of learning with explanation constraints by analyzing the generalization capabilities of this restricted class of EPAC models. This analysis builds off of a framework for semi-supervised learning (Balcan & Blum, 2010).

Another key contribution of our work is concretely analyzing this framework for a canonical class of explanation constraints given by gradient information for linear models and 2 layer neural networks. We focus on gradient constraints as we can represent many different notions of explanations, such

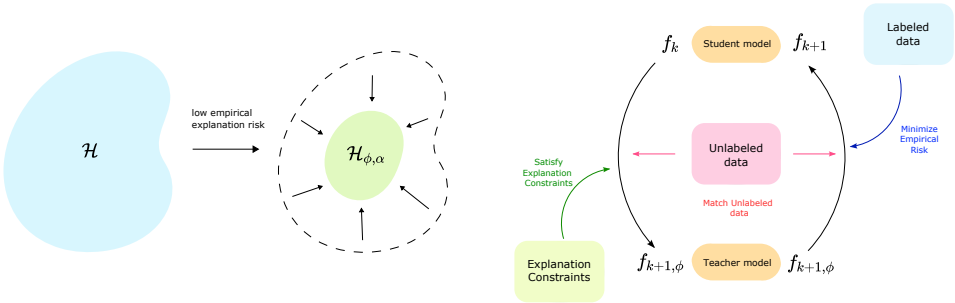

Figure 1: A restricted hypothesis class $\mathcal{H}_{\phi,\tau}$ (left). Our algorithmic solution to solve a proposed variational objective in Section 3 (right).

as feature importance and ignoring background/spurious features as a (noisy) gradient constraint. These corollaries clearly illustrate that restricting the hypothesis class via explanation constraints can lead to fewer required labeled data.

Now that we have provided a learning theoretic framework for these explanation constraints, we next consider the algorithmic question: how do we solve for these explanation-constrained models to begin with? In general, these constraints are not necessarily well-behaved and are difficult to optimize. We draw from seminal work in posterior regularization (Ganchev et al., 2010), which has also been studied in the capacity of model distillation (Hu et al., 2016), to provide a variational objective. Our first algorithmic ingredient is the use of surrogate explanation losses that quantify how well a model satisfies an explanation constraint. Our second algorithmic ingredient relates to the fact that constrained model estimation is much less scalable in general than unconstrained estimation. One can use augmented Lagrangian approaches (Ross et al., 2017; Fioretto et al., 2021), or simply regularized versions of our constrained problems (Rieger et al., 2020) (which however do not in general solve the constrained problems for non-convex parameterizations). However, even these pose challenges for complex models and increasingly complex explanations (where even simple instances of the latter can involve the model's Jacobian). We propose a tractable alternative via a variational objective that iteratively trains a model on the supervised data, and then approximately projects this learnt model onto the set of those hypotheses that satisfy the explanation constraints. Finally, we provide an extensive array of experiments that capture the benefits of learning from explanation constraints in Appendix 4. These experiments clearly illustrate our generalization bounds and also reveal fundamental tradeoffs about the design of explanation constraints.

2 Learning from Explanation Constraints

Let $\mathcal{X}$ be the instance space and $\mathcal{Y}$ be the label space. We focus on binary classification where $\mathcal{Y} = \{-1, 1\}$, but which can be naturally generalized. Let $\mathcal{D}$ be the joint data distribution over $(\mathcal{X}, \mathcal{Y})$ and $\mathcal{D}_{\mathcal{X}}$ the marginal distribution over $\mathcal{X}$. For any classifier $h : \mathcal{X} \to \mathcal{Y}$, we are interested in its classification error $\mathrm{err}(h) := \Pr_{(x,y)\sim D}(h(x) \neq y)$, though one could also use other losses, including surrogate losses to classification error. Our goal is to learn a classifier with small error from a family of functions $\mathcal{H}$. We draw from the explainable machine learning literature, and formalize (local) explanations as functionals that take in a model and test input, and output a vector:

Definition 1 (Explanations)**.** *Given an instance space $\mathcal{X}$, model hypothesis class $\mathcal{H}$, and an explanation functional $g : \mathcal{H} \times \mathcal{X} \to \mathbb{R}^r$, we say $g(h, x)$ is an explanation of h on point x induced by g.*

In our notation, g represents a functional of interest on a classifier $h \in \mathcal{H}$. For simplicity, we consider the setting when g takes a single data point and model as input, but this can be naturally extended to multiple data points and models. In practice, we need to combine these explanations with additional human knowledge on how explanations at particular sample points should look like. This can naturally be expressed in the form of explanation constraints.

Definition 2 (Explanation Constraint Set)**.** *For any instance space $\mathcal{X}$, hypothesis class $\mathcal{H}$, an explanation functional $g : \mathcal{H} \times \mathcal{X} \to \mathbb{R}^r$, and a family of constraint sets $\{C(x) \subseteq \mathbb{R}^r \mid x \in \mathcal{X}\}$, we say that $h \in \mathcal{H}$ satisfies the explanation constraints with respect to C iff:*

$$g(h, x) \in C(x), \ \forall x \in \mathcal{X}.$$

In our definition, $C(x)$ represents values that we believe our explanations should take at a point x. For example, "an input gradient of a feature 1 must be larger than feature 2" can be represented by $g(h, x) = \nabla_x h(x)$ and $C(x) = \{(x_1, \ldots, x_d) \in \mathbb{R}^d \mid x_1 > x_2\}$.

Note that we might not have access to such constraint set $C(x)$ for each of the inputs $x \in \mathcal{X}$. In practice, human annotators will be able to provide such explanation constraints for a random sample of say k data points $S_E = \{x'_1, \ldots, x'_k\}$, which we assume will be drawn i.i.d. from $\mathcal{D}_{\mathcal{X}}$. We then say that any $h \in \mathcal{H}$ S_E-satisfies the explanation constraints with respect to C iff:

$$g(h, x) \in C(x), \ \forall x \in S_E.$$

It can be seen that the constraints *are random* since samples $x'_i \in S_E$ are drawn i.i.d. from $\mathcal{D}_{\mathcal{X}}$. Thus, even if h S_E-satisfies the explanation constraints with respect to C, it likely does not hold that h fully satisfies the explanation constraints with respect to C i.e. for all inputs $x \in \mathcal{X}$. Here, we can draw from stochastic programming as well as standard learning theoretic arguments to reason about probably approximately satisfying the constraints in expectation. Before doing so we wish to first consider the notion of explanation surrogate losses, which will allow us to generalize the setup above to a form that is amenable to practical estimators.

Definition 3. *(Explanation surrogate loss) An explanation surrogate loss $\phi : \mathcal{H} \times \mathcal{X} \to \mathbb{R}$ upper bounds the indicator function for whether (and in general quantifies how well) a model h satisfies the explanation constraint $g(h, x) \in C(x)$, so that for any $h \in \mathcal{H}, x \in \mathcal{X}$:*

$$\phi(h, x) \geq 0, \quad \textit{If } g(h, x) \in C(x) \textit{ then } \phi(h, x) = 0$$

For example, we could define $\phi(h, x) = 1\{g(h, x) \in C(x)\}$. Given such a surrogate loss, we can substitute the explanation constraint that $g(h, x) \in C(x)$ with the surrogate $\phi(h, x) \leq 0$. We now have the machinery to formalize how to reason about the random explanation constraints given a random set of inputs. Consider the class of models that satisfy the explanation constraints with respect to C, as mediated by the explanation surrogate loss ϕ:

$$\mathcal{H}_\phi = \{h \in \mathcal{H} \ : \ \phi(h, x) \leq 0, \ \forall x \in \mathcal{X}\}.$$

And those models that only satisfy the explanation constraints on S_E:

$$\mathcal{H}_{\phi, S_E} = \{h \in \mathcal{H} \ : \ \phi(h, x) \leq 0, \ \forall x \in S_E\}.$$

How do we compare $\mathcal{H}_{\phi, S_E}$ to $\mathcal{H}_\phi$? Towards addressing this question, consider the expected explanation loss $\phi(h, \mathcal{D}) := \mathbb{E}_{x \sim \mathcal{D}}[\phi(h, x)]$. We can then define the class:

$$\mathcal{H}_{\phi, \mathcal{D}, \tau} = \{h \in \mathcal{H} \ : \ \phi(h, \mathcal{D}) \leq \tau\}.$$

It can be seen that this consists of models that satisfy the explanation constraints upto some slack τ (i.e. approximately) and also in expectation. At times, we may suppress the dependence on the data distribution in the notation above, and simply use $\mathcal{H}_{\phi, \tau}$ to denote the class. We can see than for any $\tau_1 < \tau_2$ we have $\mathcal{H}_{\phi, \tau_1} \subseteq \mathcal{H}_{\phi, \tau_2}$ and $\mathcal{H}_{\phi, \infty} = \mathcal{H}$ is the original concept class. We further refine this to the class of what we term EPAC models (Explanation constraints Probably Approximately Correct).

Definition 4 (EPAC model)**.** *We say that h is a τ - EPAC model w.r.t. data distribution $\mathcal{D}$ and a surrogate loss ϕ if $\phi(h, \mathcal{D}) \leq \tau$.*

Definition 5 (EPAC learnability)**.** *For any $\delta \in (0, 1), \tau > 0$, the sample complexity of (δ, τ) - EPAC learning of $\mathcal{H}$ with respect to a surrogate loss ϕ, denoted $\mathcal{M}(\tau, \delta; \mathcal{H}, \phi)$ is defined as the smallest $m \in \mathbb{N}$ for which there exists a learning rule $\mathcal{A}$ such that every data distribution $\mathcal{D}_{\mathcal{X}}$ over $\mathcal{X}$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$,*

$$\phi(\mathcal{A}(S), \mathcal{D}) \leq \inf_{h \in \mathcal{H}} \phi(h, \mathcal{D}) + \tau.$$

If no such m exists, define $\mathcal{M}(\tau, \delta; \mathcal{H}, \phi) = \infty$. We say that $\mathcal{H}$ is EPAC learnable in the agnostic setting with respect to a surrogate loss ϕ if $\forall \delta \in (0, 1), \tau > 0$, $\mathcal{M}(\tau, \delta; \mathcal{H}, \phi)$ is finite.

One might wonder if a model that satisfies the random constraints in S_E might also be an EPAC model. In the proposition below, we use natural statistical learning theoretic arguments (see Appendix C) to show that that indeed is the case. We denote $R_k(\cdot)$ as a Rademacher complexity.

Proposition 2.1. *Suppose a model h S_E-satisfies the explanation constraints so that $h \in \mathcal{H}_{\phi,S_E,\tau}$. Then, with probability at least $1-\delta$:*

$$\phi(h, \mathcal{D}_{\mathcal{X}}) \leq \tau + 2R_k(\mathcal{G}) + \sqrt{\frac{\ln(4/\delta)}{2k}},$$

when $k = |S_E|$ and $\mathcal{G} = \{\phi(h, \cdot) \mid h \in \mathcal{H}\}$.

The class $\mathcal{G}$ contains all surrogate losses of any $h \in \mathcal{H}$. Depending on the explanation constraints, $\mathcal{G}$ can be extremely large. The question of which types of explanation constraints is EPAC learnable might be of independent interest, and we further discuss this in Appendix E and give concrete cases when explanations are learnable, even without knowing the exact value of $C(x)$.

EPAC-ERM Objective. Let us next discuss *combining* the two sources of information: the explanation constraints that we set up in the previous section, together with the usual set of labeled training samples $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d. from $\mathcal{D}$ that informs the empirical risk. Combining these, we get what we call EPAC-ERM objective:

$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h, x_i, y_i) \;\text{ s.t. }\; \frac{1}{k} \sum_{i=1}^{k} \phi(h, x'_i) \leq \tau. \tag{1}$$

2.1 Generalization Bound

We assume that we are in a doubly agnostic setting when there is no classifier in the hypothesis class $\mathcal{H}$ that perfectly labels (x, y). Instead, we hope to achieve the best error rate in the hypothesis class, $h^* = \arg\min_{h \in \mathcal{H}} \mathrm{err}_{\mathcal{D}}(h)$. We also assume that h^* may have $\phi(h^*, D) > 0$. In our analysis, we start by selecting classifiers that has lower empirical explanation risk than a threshold τ, then perform a standard supervised learning with the remaining set of classifiers.

Theorem 2.2 (Generalization Bound for Agnostic Setting)**.** *Consider a hypothesis class $\mathcal{H}$, distribution $\mathcal{D}$, and explanation loss ϕ. Let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be drawn i.i.d. from $\mathcal{D}$ and $S_E = \{x'_1, \ldots, x'_k\}$ drawn i.i.d. from $\mathcal{D}_{\mathcal{X}}$. With probability at least $1-\delta$, for $h \in \mathcal{H}$ that minimizes empirical risk $\mathrm{err}_S(h)$ and has $\phi(h, S_E) \leq \tau$, we have*

$$\mathrm{err}_D(h) \leq \mathrm{err}_D(h^*_{\tau-\varepsilon_k}) + 2R_n(\mathcal{H}_{\phi,\tau+\varepsilon_k}) + 2\sqrt{\frac{\ln(4/\delta)}{2n}}, \qquad \varepsilon_k = 2R_k(\mathcal{G}) + \sqrt{\frac{\ln(4/\delta)}{2k}},$$

when $\mathcal{G} = \{\phi(h, x) \mid h \in \mathcal{H}, x \in \mathcal{X}\}$ and $h^*_\tau = \arg\min_{h \in \mathcal{H}_{\phi,\tau}} \mathrm{err}_{\mathcal{D}}(h)$.

Proof. The largely follows Balcan & Blum (2010), using Rademacher complexity instead of VC-entropy. We defer the full proof to Appendix H. □

Our bound suggests that these constraints help with our learning by a reduction of the hypothesis class $\mathcal{H}$ to $\mathcal{H}_{\phi,\tau+\varepsilon_k}$, reducing the required sample complexity. We can see that a smaller threshold τ leads to a more restricted hypothesis class and a larger reduction. However, there is also a trade-off between reduction and accuracy. In our bound, we compare against the best classifier $h^*_{\tau-\varepsilon_k} \in \mathcal{H}_{\phi,\tau-\varepsilon_k}$ instead of h^*. Since we may have $\phi(h^*, \mathcal{D}) > 0$, if τ is too small, we could reduce $\mathcal{H}$ to a hypothesis class that does not contain any good classifiers. Recall that the generalization bound for standard supervised learning — in the absence of explanation constraints — is given by

$$\mathrm{err}_D(h) \leq \mathrm{err}_D(h^*) + 2R_n(\mathcal{H}) + 2\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

We provide the bounds for gradient explanations constraint in Appendix B

3 Algorithms for Learning from Explanation Constraints

Although we have analyzed learning with explanation constraints, algorithms to solve this constrained optimization problem are non-trivial. In this setting, we assume that we have access to n labeled data $\{(x_i, y_i)\}_{i=1}^n$, m unlabeled data $\{x_{n+1}, \ldots, x_{n+m}\}$, and k data with explanations

$\{(x_i, \phi(\cdot, x_i))\}_{i=n+m+1}^{n+m+k}$. We argue that in many cases, n labeled data are the most expensive to annotate. The k data points with explanations also have non-trivial cost; they require an expert to provide the annotated explanation or provide a surrogate loss ϕ. If the surrogate loss is specified then we can evaluate it on any unlabeled data, otherwise these data points with explanations could be expensive. However, the m data points can cheaply be obtained as they are completely unlabeled. We now consider approaches to incorporate explanation information into a machine learning pipeline.

EPAC-ERM: Recall our EPAC-ERM objective from equation 1:

$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} 1\{h(x_i) \neq y_i\} \text{ s.t. } \frac{1}{k} \sum_{j=n+m+1}^{n+m+k} \phi(h, x_j) \leq t$$

for some constant t. This constraint in general requires more complex optimization techniques (e.g., running multiple iterations and comparing values of t) to solve algorithmically.

Augmented Lagrangian objectives:

$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} 1[h(x_i) \neq y_i] + \frac{\lambda}{k} \sum_{j=n+m+1}^{n+m+k} \phi(h, x_j)$$

As is done in prior work (Rieger et al., 2020), we can consider an augmented Lagrangian objective. However, this does not exactly fit into our analytical framework.

While these approaches are viable, they do not necessarily scale well to larger (deep) models. As a consequence, we propose a *new* variational objective

$$\min_{h \in \mathcal{H}} (1 - \tau) \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}} [\ell(h(x), y)] + \tau \inf_{f \in \mathcal{H}_{\phi,\tau}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [\ell(h(x), f(x))],$$

where ℓ is some loss function and $\tau \geq 0$ is some threshold. We can see that there are two models in the objective: the main "student" model h, and a "teacher" model f which serves as a projection of h onto a set of EPAC models $\mathcal{H}_{\phi,t}$. We approximate this objective with the following iterative technique, drawing inspiration from prior work in posterior regularization (Ganchev et al., 2010; Hu et al., 2016). We remark that we can replace $\mathcal{H}_{\phi,t}$ with a simpler class of teacher models for greater efficiency. Specifically, we iteratively: fix the current student model iterate $h_{t,\phi}$, and learn the explanation-regularized teacher function $f_{t+1,\phi}$ (that aims to project $h_{t,\phi}$ onto the set of explanation constrained models); and then fix that to obtain the next iterate $h_{t+1,\phi}$ of the student model that aims to match the outputs of $f_{t+1,\phi}$ on *unlabeled data* in addition to the labeled samples:

$$f_{t+1,\phi} = \arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=n+1}^{n+m} \ell(h(x_i), h_t(x_i)) + \lambda \max\left(0, \frac{1}{k} \sum_{i=n+m+1}^{n+m+k} \phi(h, x_i) - \tau\right)$$

$$h_{t+1,\phi} = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i) + \frac{1}{m} \sum_{i=n+1}^{n+m} \ell(h(x_i), f_{t+1,\phi}(x_i)),$$

given some initialization h_0. This is related to self-training, although it can be seen that we also use the predictions of our student model to pseudolabel m unlabeled data to train our teacher model.

4 EXPERIMENTS

We provide both synthetic and real-world experiments to support our theoretical results and clearly illustrate interesting tradeoffs of incorporating explanations. In our experiments, we compare our method against 3 baselines: (1) a standard supervised learning approach, (2) a simple Lagrangian regularized method (that directly penalizes the surrogate loss ϕ), and (3) a self-training approach that propagates its own predictions over a set of unlabeled data. These experiments also demonstrate that our new variational approach is preferable to simple Lagrangian methods and other supervised methods in many cases. More extensive ablations are deferred to Appendix O, and code to replicate our experiments will be released publicly with the full paper.

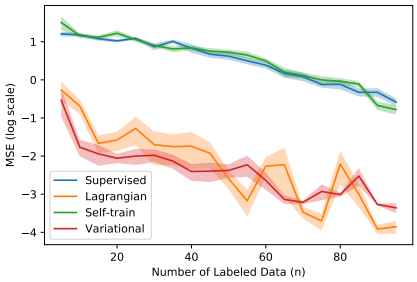

Figure 2: Comparison of MSE on regressing a linear model. Results are averaged over 5 seeds. $m = 1000, k = 20$.

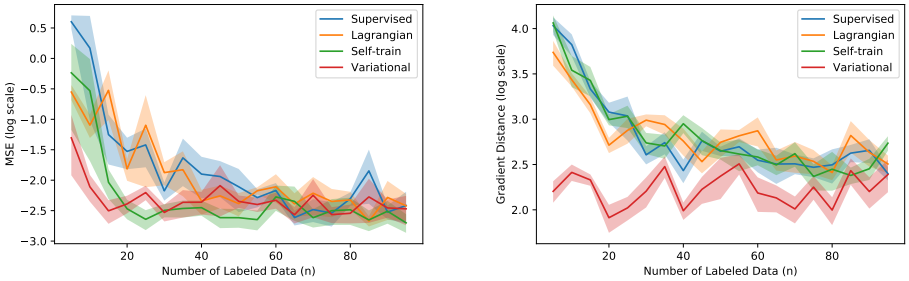

Figure 3: Comparison of MSE on regressing a 2 layer neural network (left). Comparison of ℓ_2 distance over input gradients on the test data as we vary the amount of labeled data n (right). Results are averaged over 5 seeds. $m = 1000, k = 20$.

4.1 Regression Task with Exact Gradient Information

In our synthetic experiments, we focus on a regression task where we try to learn some underlying model in our hypothesis class. Our data is given by $\mathcal{X} = \mathbb{R}^d$, and we try to learn a target function $h^* : \mathcal{X} \to \mathbb{R}$. Our data distribution is given by $X \sim \mathcal{N}(0, \sigma^2 I)$, where I is a $d \times d$ identity matrix. We generate h^* by randomly initializing a model in the specific hypothesis class $\mathcal{H}$. We assume that we have n labeled data, m unlabeled data, and k data with explanations.

We first present a synthetic experiment for learning with a perfect explanation, meaning that $\phi(h^*, S) = 0$. We consider the case where we have the *exact* gradient of h^*. Here, let $\mathcal{H}$ be a linear classifier and note that the exact gradient gives us the slope of the linear model, and we only need to learn the bias term. Incorporating these explanation indeed helps as both methods that include explanation constraints (Lagrangian and ours) perform much better (Figure 2).

We also demonstrate incorporating this information for 2 layer neural networks. We observe a clear difference between the simpler Lagrangian approach and our variational objective (Figure 3 - left). Our method is clearly the best in the setting with limited labeled data and matches the performance of the strong self-training baseline with sufficient labeled data. We note that this is somewhat expected, as these constraints primarily help in the setting with limited labeled data; with enough labeled data, standard PAC bounds suffice for strong performance.

We also analyze how strongly the approaches enforce these explanation constraints on new data points that are seen at test time (Figure 3 - right) for 2 layer NNs. We observe that our variational objective approaches have input gradients that more closely match the ground-truth target network's input gradients. This demonstrates that, in the case of 2 layer NNs with gradient explanations, our approach best achieves both good performance and satisfying the constraints. Standard self-training achieves similar performance in terms of MSE but has no notion of satisfying the explanation constraints. The Lagrangian method does not achieve the same level of satisfying these explanations as it is unable to generalize and satisfy these constraints on new data.

4.2 Tasks with Imperfect Explanations

Assuming access to perfect explanations may be unrealistic in practice, so we present experiments when our explanations are imperfect. We present classification tasks (Figure 4) from a weak supervision benchmark (Zhang et al., 2021). In this setting, we obtain explanations through the approxi-

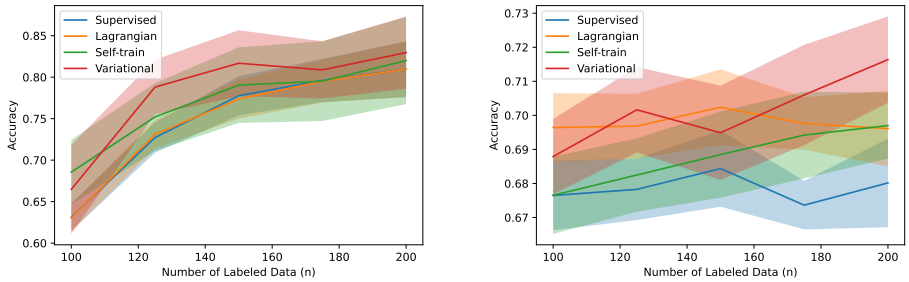

Figure 4: Comparison of accuracy on the YouTube (left) and the Yelp (right) datasets. Here, we let $m = 1500, k = 150, T = 2, t = 1$. Results are averaged over 5 seeds.

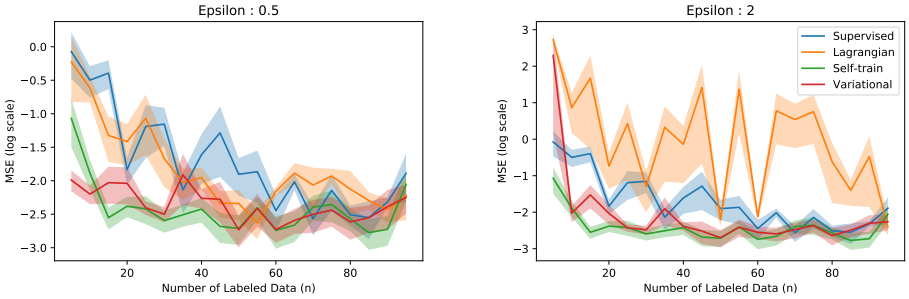

Figure 5: Comparison of MSE on regressing a 2 Layer Neural Network with explanations of noisy gradients. $m = 1000, k = 20, T = 10$. Results are averaged over 5 seeds.

mate gradients of a single weak labeler, as is done in (Sam & Kolter, 2022). We note that this differs from the standard setting of the benchmark, as we assume access to some labeled data and only use gradient information as our explanations (and not the noisy classifiers).

We observe that our variational objective achieves better performance than all other baseline approaches, across varying amounts of labeled data. We remark that the explanation in this dataset is a noisy gradient explanation along two feature dimensions, yet this still improves upon methods that do not incorporate this explanation constraint. Indeed, our method outperforms the Lagrangian approach, showing the benefits of iterative rounds of self-training over the unlabeled data.

In addition to our real-world experiments, we present synthetic experiments in the same regression setting as above. To generate an imperfect explanation, we use $\nabla_x h^*(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In our experiments, we add fixed, randomly sampled noise for each of our k data annotated with explanations. We provide results for a regression task on 2-layer neural networks in Figure 5 and under additional levels of noise in Appendix M.1. This reveals that our method tolerates noisy explanations far better than the Lagrangian approach. Our method also performs comparably to the methods that do not use noisy explanations.

5 Discussion

Our work proposes a new learning theoretic framework that provides insight into how apriori explanations of desired model behavior can benefit the standard machine learning pipeline. We provide instantiations of our analysis for the canonical class of gradient explanations, which captures many explanations of feature importance and not using spurious correlations. It would be of interest to provide corollaries for other types of explanations in future work. As mentioned before, the generality of our framework has larger implications towards incorporating constraints that are not considered as "standard" explanations. For example, this work can be leveraged to incorporate more general notions of side information and inductive biases. As a whole, our paper supports using further information (e.g., explanation constraints) in the standard learning setting.

REFERENCES

Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)*, 57(3):1–46, 2010.

John R Birge and Francois Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

Ferdinando Fioretto, Pascal Van Hentenryck, Terrence WK Mak, Cuong Tran, Federico Baldo, and Michele Lombardi. Lagrangian duality for constrained deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 118–135. Springer, 2021.

Spencer Frei, Difan Zou, Zixiang Chen, and Quanquan Gu. Self-training converts weak learners to strong learners in mixture models. In *International Conference on Artificial Intelligence and Statistics*, pp. 8003–8021. PMLR, 2022.

Kuzman Ganchev, Joao Gracca, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049, 2010.

Siddhant Garg and Yingyu Liang. Functional regularization for representation learning: A unified theoretical perspective. *Advances in Neural Information Processing Systems*, 33:17187–17199, 2020.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pp. 2376–2384. PMLR, 2019.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2410–2420, 2016.

Peter Kall, Stein W Wallace, and Peter Kall. *Stochastic programming*, volume 6. Springer, 1994.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Jeffrey Li, Vaishnavh Nagarajan, Gregory Plumb, and Ameet Talwalkar. A learning theoretic perspective on local explainability. In *International Conference on Learning Representations*, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Tengyu Ma. Lecture notes from machine learning theory, 2022. URL `http://web.stanford.edu/class/stats214/`.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 607–617, 2020.

W James Murdoch, Peter J Liu, and Bin Yu. Beyond word importance: Contextual decomposition to extract interactions from lstms. In *International Conference on Learning Representations*, 2018.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

Rattana Pukdee, Dylan Sam, Maria-Florina Balcan, and Pradeep Ravikumar. Label propagation with weak supervision. *arXiv preprint arXiv:2210.03594*, 2022.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, pp. 269. NIH Public Access, 2017.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pp. 8116–8126. PMLR, 2020.

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 2662–2670, 2017.

Dylan Sam and J Zico Kolter. Losses over labels: Weakly supervised learning via direct loss construction. *arXiv preprint arXiv:2212.06921*, 2022.

Sungyong Seo, Sercan Arik, Jinsung Yoon, Xiang Zhang, Kihyuk Sohn, and Tomas Pfister. Controlling neural networks with rule representations. *Advances in Neural Information Processing Systems*, 34:11196–11207, 2021.

Han Shao, Omar Montasser, and Avrim Blum. A theory of PAC learnability under transformation invariances. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=l1WlfNaRkKw`.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *ICML Workshop on Visualization for Deep Learning, 2017*, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 2020.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698, 2020.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. *Advances in neural information processing systems*, 31, 2018.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565, 2020.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, et al. Concept embedding models. In *NeurIPS 2022-36th Conference on Neural Information Processing Systems*, 2022.

Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. Wrench: A comprehensive benchmark for weak supervision. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

A Related Work

Recent advances in deep learning have led to models that achieve high performance but which are also highly complex (LeCun et al., 2015; Goodfellow et al., 2016). Understanding these complex models is crucial for safe and reliable deployments of these systems in the real-world. One approach to improve our understanding of a model is through explanations. This can take many forms such as feature importance (Ribeiro et al., 2016; Smilkov et al., 2017; Lundberg & Lee, 2017; Sundararajan et al., 2017), high level concepts (Kim et al., 2018; Yeh et al., 2020), counterfactual examples (Wachter et al., 2017; Goyal et al., 2019; Mothilal et al., 2020), or influential training samples (Koh & Liang, 2017; Yeh et al., 2018).

In contrast to generating post-hoc explanations of a given model, we aim to learn models given apriori explanations. There has been some recent work along such lines. Koh et al. (2020); Zarlenga et al. (2022) incorporates explanations within the model architecture by requiring a conceptual bottleneck layer. Ross et al. (2017); Rieger et al. (2020) use explanations to modify the learning procedure for any class of models: they incorporate explanations as a regularizer, penalizing models that do not exhibit apriori given explanations; Ross et al. (2017) penalize input gradients, while Rieger et al. (2020) penalize a Contextual Decomposition score (Murdoch et al., 2018). Some of these suggest that constraining models via explanations leads to higher accuracies and more robustness to spurious correlation, but do not provide analytical guarantees. On the theoretical front, Li et al. (2020) show that models that are easier to explain locally also generalize well.

Our contribution is to provide an analytical framework for learning from explanations that quantifies the benefits of explanation constraints. Our analysis is closely related to the framework of learning with side information. Balcan & Blum (2010) shows how unlabeled data can help in semi-supervised learning through a notion of compatibility between the data and the target model. This seminal work studies classical notions of side information (e.g., margin, smoothness, and co-training). Subsequent papers have adapted this learning theoretic framework to study the benefits of representation learning (Garg & Liang, 2020) and transformation invariance (Shao et al., 2022). On the contrary, our paper focuses on the more recent notion of explanations. Rather than focus on the benefits of unlabeled data, we characterize the quality of different explanations.

There is also prior work proposing learning objectives that incorporate rules into deep neural networks (Hu et al., 2016; Fioretto et al., 2021; Seo et al., 2021). While (Hu et al., 2016) also leverages variational objectives, their method specifically concerns itself with logic rules and over probability distributions using KL divergence projections. On the contrary, our approach handles more general forms of explanations and that naturally conforms to our theoretical framework. Our work can also be connected to the self-training literature (Chapelle et al., 2009; Xie et al., 2020; Wei et al., 2020; Frei et al., 2022), where we could view our variational objective as comprising a regularized (potentially simpler) teacher model that encodes these explanation constraints into a student model.

B Gradient Explanations for Particular Hypothesis Classes

In this section, we further quantify the usefulness of explanation constraints on different concrete examples and characterize the Rademacher complexity of the restricted hypothesis classes. In particular, we consider an explanation constraint of a constraint on the input gradient. For example, we may want our model's gradient to be close to that of some $h' \in \mathcal{H}$. This translates to $g(h, x) = \nabla_x h(x)$ and $C(x) = \{x \in \mathbb{R}^d \mid \|x - \nabla_x h'(x)\| \leq \tau\}$ for some $\tau > 0$.

B.1 Gradient Explanations for Linear Models

Theorem B.1 (Rademacher complexity of linear models with gradient constraint, uniform distribution on a sphere)**.** *Let $\mathcal{D}_{\mathcal{X}}$ be a uniform distribution on a unit sphere in $\mathbb{R}^d$, let $\mathcal{H} = \{h : x \mapsto \langle w_h, x\rangle \mid w_h \in \mathbb{R}^d, ||w_h||_2 \leq B\}$ be a class of linear models with weights bounded by a constant B. Let $\phi(h, x) = \theta(w_h, w_{h'})$ be a surrogate loss where $\theta(u, v)$ is an angle between u, v. We have*

$$R_n(\mathcal{H}_{\phi,\tau}) \leq \frac{B}{\sqrt{n}} \left(\sin(\tau) \cdot p + \frac{1-p}{2} \right),$$

where $p = \operatorname{erf}\left(\frac{\sqrt{d}\sin(\tau)}{\sqrt{2}}\right)$ and $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ is the standard error function.

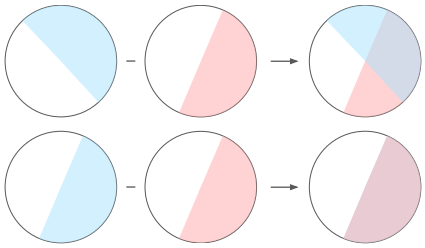

Figure 6: Visualization of the piecewise constant function of $\nabla_x h(x) - \nabla_x h'(x)$ when h is a 2 Layer NNs with 1 node. Background colors represent regions with non-zero value.

Proof. We defer the proof to Appendix J. □

The standard upper bound on the Rademacher complexity of linear models is $\frac{B}{\sqrt{n}}$. Our bound has a nice interpretation; we shrink our bound by a factor of $(\frac{1-p}{2} + \sin(\tau)p)$. We remark that d increases, we observe that $p \to 1$, so the term $\sin(\tau)p$ dominates this factor. As a consequence, we get that our bound is now scaled by $\sin(\tau) \approx \tau$ and the the Rademacher complexity scales down by a factor of τ. This implies that given n labeled data, to achieve a fast rate $\mathcal{O}(\frac{1}{n})$, we need τ to be $O(\frac{1}{\sqrt{n}})$.

B.2 Gradient Explanations for 2 Layer NNs

Theorem B.2 (Rademacher complexity of 2 layer neural networks (m hidden nodes) with gradient constraint)**.** *Let $\mathcal{X}$ be an instance space and $\mathcal{D}_{\mathcal{X}}$ be a distribution over $\mathcal{X}$ with a large enough support. Let $\mathcal{H} = \{h : x \mapsto \sum_{j=1}^{m} w_j \sigma(u_j^\top x) | w_j \in \mathbb{R}, u_j \in \mathbb{R}^d, \sum_{j=1}^{m} |w_j| \|u_j\|_2 \leq B\}$ be a class of two layer neural networks with a ReLU activation function and bounded weight. Assume that there exists some constant $C > 0$ such that $\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}[\|x\|_2^2] \leq C^2$. Consider explanation loss*

$$\phi(h, x) = \|\nabla_x h(x) - \nabla_x h'(x)\|_2 + \infty \cdot 1\{\|\nabla_x h(x) - \nabla_x h'(x)\| > \tau\}$$

for some $\tau > 0$. Then, we have that

$$R_n(\mathcal{H}_{\phi,\tau}) \leq \frac{3\tau m C}{\sqrt{n}}$$

Proof. (Sketch) The key ingredient is to identify the impact of the gradient constraint and the form of class $\mathcal{H}_{\phi,\tau}$. We provide an idea when we have $m = 1$ node. We write $h(x) = w\sigma(u^\top x)$ and $h'(x) = w'\sigma(u'^\top x)$ when $\|u\| = \|u'\| = 1$. We have

$$\nabla_x h(x) - \nabla_x h'(x) = wu1\{u^\top x > 0\} - w'u'1\{(u')^\top x > 0\}$$

is a piecewise constant function (Figure 6). Assume that the probability mass of each region is non-negative, our gradient constraint implies that the norm of each region cannot be larger than τ.

1. If u, u' have different directions, we have 4 regions in $\nabla_x h(x) - \nabla_x h'(x)$ and can conclude that $|w| < \tau, |w'| < \tau$.
2. If $u = u'$ have the same direction, we only have 2 regions in $\nabla_x h(x) - \nabla_x h'(x)$ and can conclude that $\|wu - w'u'\| = |w - w'| < \tau$.

The gradient constraint enforces a model to have the same node boundary ($u = u'$) with a small weight difference $|w - w'| < \tau$ or that node would have a small weight $|w| < \tau$. This finding allows us to determine the restricted class $\mathcal{H}_{\phi,\tau}$, and we can use this to bound the Rademacher complexity accordingly. For full details, see Appendix K. □

We compare this with the standard Rademacher complexity of a two layer neural network (Ma, 2022)

$$R_n(\mathcal{H}) \leq \frac{2BC}{\sqrt{n}}.$$

We can do better than this standard bound if $\tau < \frac{2B}{3m}$. One interpretation for this is that we have a budget at most τ to change the weight of each node and for total m nodes, we can change the weight by at most τm. We compare this to B which is an upper bound on the total weight $\sum_{j=1}^{m} |w_j| \|u_j\| \leq B$. Therefore, we can do better than a standard bound when we can change the weight by at most two thirds of the average weight $\frac{2B}{3m}$ for each node. We note that our bound does not depend on $\mathcal{D}$ as we choose a strong explanation loss that guarantees that the gradient constraint holds almost everywhere. Extending to a weaker loss such as $\phi(h, x) = \|\nabla_x h(x) - \nabla_x h'(x)\|$ is a future research direction. We also assume that there exists x with a positive probability density at any partition created by $\nabla_x h(x)$. In contrast, our result for linear models uses a weaker explanation loss and depends on $\mathcal{D}$ (see Theorem J.1).

C Uniform Convergence via Rademacher Complexity

A standard tool for providing performance guarantees of supervised learning problems is a generalization bound via uniform convergence. We will first define the Rademacher complexity and its corresponding generalization bound.

Definition 6. *Let $\mathcal{F}$ be a family of functions mapping $\mathcal{X} \to \mathbb{R}$. Let $S = \{x_1, \ldots, x_m\}$ be a set of examples drawn i.i.d. from a distribution $D_{\mathcal{X}}$. Then, the empirical Rademacher complexity of $\mathcal{F}$ is defined as*

$$R_S(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f \in \mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m} \sigma_i f(x_i)\right)\right]$$

where $\sigma_1, \ldots, \sigma_m$ are independent random variables uniformly chosen from $\{-1, 1\}$.

Definition 7. *Let $\mathcal{F}$ be a family of functions mapping $\mathcal{X} \to \mathbb{R}$. Then, the Rademacher complexity of $\mathcal{F}$ is defined as*

$$R_n(\mathcal{F}) = \mathop{\mathbb{E}}_{S \sim \mathcal{D}_{\mathcal{X}}^n}\left[R_S(\mathcal{F})\right].$$

The Rademacher complexity is the expectation of the empirical Rademacher complexity, over n samples drawn i.i.d. from the distribution $\mathcal{D}_{\mathcal{X}}$.

Theorem C.1 (Rademacher-based uniform convergence)**.** *Let $D_{\mathcal{X}}$ be a distribution over $\mathcal{X}$, and $\mathcal{F}$ a family of functions mapping $\mathcal{X} \to [0, 1]$. Let $S = \{x_1, \ldots, x_n\}$ be a set of samples drawn i.i.d. from $D_{\mathcal{X}}$, then with probability at least $1 - \delta$ over our draw S,*

$$|\mathbb{E}_{\mathcal{D}}[f(x)] - \hat{\mathbb{E}}_S[f(x)]| \leq 2R_n(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

This holds for every function $f \in \mathcal{F}$, and $\hat{\mathbb{E}}_S[f(x)]$ is expectation over a uniform distribution over S.

This bound on the empirical Rademacher complexity leads to the standard generalization bound for supervised learning.

Theorem C.2. *For a binary classification setting when $y \in \{\pm 1\}$ with a zero-one loss, for $\mathcal{H} \subset \{h : \mathcal{X} \to \{-1, 1\}\}$ be a family of binary classifiers, let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is drawn i.i.d. from D then with probability at least $1 - \delta$, we have*

$$|\mathrm{err}_{\mathcal{D}}(h) - \widehat{\mathrm{err}}_S(h)| \leq R_n(\mathcal{H}) + \sqrt{\frac{\ln(2/\delta)}{2n}},$$

for every $h \in \mathcal{H}$ when

$$\mathrm{err}_{\mathcal{D}}(h) = \Pr_{(x,y) \sim \mathcal{D}}(h(x) \neq y)$$

and

$$\widehat{\mathrm{err}}_S(h) = \frac{1}{n}\sum_{i=1}^{n} 1[h(x_i) \neq y_i]$$

is the empirical error on S.

For a linear model with a bounded weights in ℓ_2 norm, the Rademacher complexity is $\mathcal{O}(\frac{1}{\sqrt{n}})$. We refer to the proof from Ma (2022) for this result.

Theorem C.3 (Rademacher complexity of a linear model ((Ma, 2022))). *Let $\mathcal{X}$ be an instance space in $\mathbb{R}^d$, let $\mathcal{D}_\mathcal{X}$ be a distribution on $\mathcal{X}$, let $\mathcal{H} = \{h : x \to \langle w_h, x\rangle \mid w_h \in \mathbb{R}^d, ||w_h||_2 \leq B\}$ be a class of linear model with weights bounded by some constant $B > 0$ in ℓ_2 norm. Assume that there exists a constant $C > 0$ such that $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[||x||_2^2] \leq C^2$. For any $S = \{x_1, \ldots, x_n\}$ is drawn i.i.d. from $\mathcal{D}_\mathcal{X}$, we have*

$$R_S(\mathcal{H}) \leq \frac{B}{n}\sqrt{\sum_{i=1}^{n} ||x_i||_2^2}$$

and

$$R_n(\mathcal{H}) \leq \frac{BC}{\sqrt{n}}.$$

Many of our proofs require the usage of Talgrand's lemma, which we now present.

Lemma C.4. *[Talgrand's Lemma (Ledoux & Talagrand, 1991)] Let $\phi : \mathbb{R} \to \mathbb{R}$ be a k-Lipschitz function. Then for a hypothesis class $\mathcal{H} = \{h : \mathbb{R}^d \to \mathbb{R}\}$, we have that*

$$R_S(\phi \circ \mathcal{H}) \leq kR_s(\mathcal{H})$$

where $\phi \circ \mathcal{H} = \{f : z \mapsto \phi(h(z)) | h \in \mathcal{H}\}$.

D EPAC Learnable Constraints

We know that constraints $C(x)$ capture human knowledge about how explanations at a point x should behave. For any constraints $C(x)$ that are known apriori for all $x \in \mathcal{X}$, we can evaluate whether a model satisfies the constraints at a point $x \in \mathcal{X}$. This motivates us to discuss the ability of models to generalize from any finite samples S_E to satisfy these constraints over $\mathcal{X}$ with high probability. Having access to $C(x)$ is equivalent to knowing how models should behave over *all* possible data points in terms of explanations, which may be too strong of an assumption. Nevertheless, many forms of human knowledge can be represented by a closed-form function $C(x)$. For example,

1. An explanation has to take value in a fixed range can be represented by $C(x) = \Pi_{i=1}^{r}[a_i, b_i], \forall x \in \mathcal{X}$.
2. An explanation has to stay in a ball around x can be represented by $C(x) = \{u \in \mathbb{R}^d \mid ||u - x||_2 \leq r\}$.
3. An explanation has to stay in a rectangle around $\frac{x}{3}$ can be represented by $C(x) = \{u \in \mathbb{R}^d \mid \frac{x_i}{3} - a_i \leq u_i \leq \frac{x_i}{3} + b_i, i = 1, \ldots, d\}$.

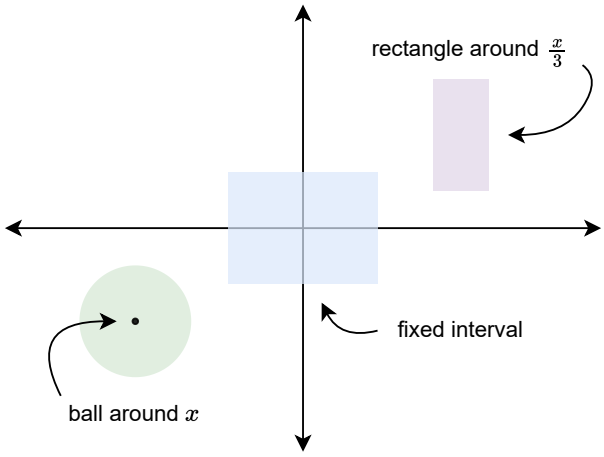

Figure 7: Illustration of examples of explanation constraints, given from some learnable class $C(x)$.

In this case, there always exists a surrogate loss that represents the explanation constraints $C(x)$; for example, we can set $\phi(h, x) = 1\{g(h, x) \in C(x)\}$. On the other hand, directly specifying explanation constraints through a surrogate loss would also imply that $C(x)$ is known apriori for all $x \in \mathcal{X}$. The task of generalization to satisfy the constraint on unseen data is well-defined in this setting. Furthermore, if a surrogate loss ϕ is specified, then we can evaluate $\phi(h, x)$ on any unlabeled data point without the need for human annotators which is a desirable property.

On the other hand, we usually do not have knowledge over all data points $x \in \mathcal{X}$; rather, we may only know these explanation constraints over a random sample of k data points $S_E = \{x'_1, \ldots, x'_k\}$. If we do not know the constraint set $C(x)$, it is unclear what satisfying the constraint at an unseen data point x means. Indeed, without additional assumptions, it may not make sense to think about generalization. For example, if there is no relationship between $C(x)$ for different values of x, then it is not possible to infer about $C(x)$ from $C(x'_i)$ for $i = 1, \ldots, k$. In this case, we could define

$$\phi(h, x) = 1\{g(h, x) \in C(x)\}1\{x \in S_E\},$$

where we are only interested in satisfying these explanation constraints over the finite sample S_E. For other data points, we have $\phi(h, x) = 0$. This guarantees that any model with low empirical explanation loss would also achieve loss expected explanation loss, although this does not have any particular implication on any notion of generalization to new constraints. Regardless, we note that our explanation constraints still reduce the size of the hypothesis class from $\mathcal{H}$ to $\mathcal{H}_{\phi,\tau}$, leading to an improvement in sample complexity.

The more interesting setting, however, is when we make an additional assumption that the true (unknown) surrogate loss ϕ exists and, during training, we only have access to instances of this surrogate loss evaluated on the sample $\phi(\cdot, x'_i)$. We can apply a uniform convergence argument to achieve

$$\phi(h, \mathcal{D}_{\mathcal{X}}) \leq \phi(h, S_E) + 2R_k(\mathcal{G}) + \sqrt{\frac{\ln(4/\delta)}{2k}}$$

with probability at least $1 - \delta$ over S_E, drawn i.i.d. from $\mathcal{D}_{\mathcal{X}}$ and $\mathcal{G} = \{\phi(h, \cdot) | h \in \mathcal{H}\}$, $k = |S_E|$. Although the complexity term $R_k(\mathcal{G})$ is unknown (since ϕ is unknown), we can upper bound this by the complexity of a class of functions Φ (e.g., neural networks) that is large enough to well-approximate any $\phi(h, \cdot) \in \mathcal{G}$, meaning that $R_k(\mathcal{G}) \leq R_k(\Phi)$. Comparing to the former case when $C(x)$ is known for all $x \in \mathcal{X}$ apriori, the generalization bound has a term that increases from $R_k(\mathcal{G})$ to $R_k(\Phi)$, which may require more explanation-annotated data to guarantee generalization to new data points. We note that the simpler constraints lead to a simpler surrogate loss, which in turn implies a less complex upper bound Φ. This means that simpler constraints are easier to learn.

Nonetheless, this is a more realistic setting when explanation constraints are hard to acquire and we do not have the constraints for all data points in $\mathcal{X}$. For example, Ross et al. (2017) considers an image classification task on MNIST, and imposes an explanation constraint in terms of penalizing the input gradient of the background of images. In essence, the idea is that the background should be less important than the foreground for the classification task. In general, this constraint does not have a closed-form expression, and we do not even have access to the constraint for unseen data points. However, if we assume that a surrogate loss $\phi(h, \cdot)$ can be well-approximated by two layer neural networks, then our generalization bound allows us to reason about the ability of model to generalize and ignore background features on new data.

E EXAMPLES FOR EPAC LEARNABLE CONSTRAINTS

In this section, we look at the Rademacher complexity of $\mathcal{G}$ for different explanation constraints to characterize how many samples with explanation constraints are required in order to generalize to satisfying the explanation constraints on unseen data. We remark that this is a different notion of sample complexity; these unlabeled data require annotations of explanation constraints, not standard labels. In practice, this can be easier and less expertise might be necessary if define the surrogate loss ϕ directly. First, we analyze the case where our explanation is given by the gradient of a linear model.

Proposition E.1 (Learning a gradient constraint for linear models)**.** *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$. Let $\mathcal{H} = \{h : x \mapsto \langle w_h, x\rangle \mid w_h \in \mathbb{R}^d, \|w_h\|_2 \leq B\}$ be a class of linear models that pass through the origin. Let $\phi(h, x) = \theta(w_h, w_{h'})$ be a surrogate explanation loss. Let $\mathcal{G} = \{\phi(h, \cdot) \mid h \in \mathcal{H}\}$,*

then we have

$$R_n(\mathcal{G}) \leq \frac{\pi}{2\sqrt{m}}.$$

Proof. For a linear separator, $\phi(h, \cdot)$ is a constant function over $\mathcal{X}$. The Rademacher complexity is given by

$$\begin{aligned}
R_n(\mathcal{G}) &= \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\phi(h,\cdot)\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\phi(h,x_i)\right)\right]\right] \\
&= \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\right)\theta(w_h, w_{h'})\right]\right] \\
&= \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\right)\sup_{h\in\mathcal{H}}\theta(w_h, w_{h'})\right]\right] \\
&= \frac{\pi}{2}\mathop{\mathbb{E}}_{\sigma}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\right|\right] \\
&\leq \frac{\pi}{2\sqrt{m}}.
\end{aligned}$$

□

We compare this with the Rademacher complexity of linear models which is given by $R_m(\mathcal{H}) \leq \frac{B}{\sqrt{m}}$. The upper bound does not depend on the upper bound on the weight B. In practice, we know that the gradient of a linear model is constant for any data point. This implies that knowing a gradient of a single point is enough to identify the gradient of the linear model.

We consider another type of explanation constraint that is given by a noisy model. Here, we could observe either a noisy classifier and noisy regressor, and the constraint could be given by having similar outputs to this noisy model. This is reminiscent of learning with noisy labels (Natarajan et al., 2013) or weak supervision (Ratner et al., 2016; 2017; Pukdee et al., 2022). In this case, our explanation g is simply the hypothesis element h itself, and our constraint is on the values that $h(x)$ can take. We first analyze this in the classification setting.

Proposition E.2 (Learning a constraint given by a noisy classifier)**.** *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$. Consider a binary classification task with $\mathcal{Y} = \{-1, 1\}$. Let $\mathcal{H}$ be a hypothesis class. Let $\phi(h, x) = 1[h(x) \neq h'(x)]$ be a surrogate explanation loss. Let $\mathcal{G} = \{\phi(h, \cdot) \mid h \in \mathcal{H}\}$, then we have*

$$R_n(\mathcal{G}) = \frac{1}{2}R_n(\mathcal{H}).$$

Proof.

$$\begin{aligned}
R_n(\mathcal{G}) &= \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\phi(h,\cdot)\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\phi(h,x_i)\right)\right]\right] \\
&= \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i(\frac{1-h(x)h'(x)}{2})\right)\right]\right] \\
&= \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i(\frac{h(x)h'(x)}{2})\right)\right]\right] \\
&= \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i(\frac{h(x)}{2})\right)\right]\right] \\
&= \frac{1}{2}R_n(\mathcal{H}).
\end{aligned}$$

□

Here, to learn the restriction of $\mathcal{G}$ is on the same order of $R_n(\mathcal{H})$. For a given noisy regressor, we observe slightly different upper bound.

Proposition E.3 (Learning a constraint given by a noisy regressor)**.** *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$. Consider a regression task with $\mathcal{Y} = \mathbb{R}$. Let $\mathcal{H}$ be a hypothesis class that $\forall h \in \mathcal{H}, -h \in \mathcal{H}$. Let $\phi(h, x) = |h(x) - h'(x)|$ be a surrogate explanation loss. Let $\mathcal{G} = \{\phi(h, \cdot) \mid h \in \mathcal{H}\}$, then we have*

$$R_n(\mathcal{G}) \leq 2R_n(\mathcal{H}).$$

Proof.

$$\begin{aligned}
R_n(\mathcal{G}) &= \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\phi(h,\cdot)\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\phi(h, x_i)\right)\right]\right] \\
&= \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i|h(x_i) - h'(x_i)|\right)\right]\right] \\
&= \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\max(0, h(x_i) - h'(x_i)) + \frac{1}{m}\sum_{i=1}^{m}\sigma_i\max(0, h'(x_i) - h(x_i))\right)\right]\right] \\
&\leq \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\max(0, h(x_i) - h'(x_i))\right)\right]\right] + \\
&\qquad \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\max(0, h'(x_i) - h(x_i))\right)\right]\right] \\
&\leq \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h(x_i) - h'(x_i))\right)\right]\right] + \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h'(x_i) - h(x_i))\right)\right]\right],
\end{aligned}$$

where in the last line, we apply Talgrand's lemma C.4 and note that the max function $\max(0, h(x))$ is 1-Lipschitz; in the third line, we note that we break up the supremum as both terms by definition of the $\max$ function are non-negative. Then, noting that we do not optimize over $h'(x)$, we further simplify this as

$$\begin{aligned}
R_n(\mathcal{G}) &\leq \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)\right)\right]\right] + \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i(-h(x_i))\right)\right]\right] \\
&\leq 2R_n(\mathcal{H}).
\end{aligned}$$

□

As mentioned before, knowing apriori surrogate loss ϕ might be too strong. In practice, we may only have access to the instances $\phi(\cdot, x_i)$ on a set of samples $S = \{x_1, \ldots, x_k\}$. We also consider the case when $\phi(h, x) = |h(x) - h'(x)|$ when h' is unknown and h' belongs to a learnable class $\mathcal{C}$.

Proposition E.4 (Learning a constraint given by a noisy regressor from some learnable class $\mathcal{C}$)**.** *Assume $\mathcal{D}$ is a distribution over $\mathbb{R}^d$. Let $\mathcal{H}$ and $\mathcal{D}$ be hypothesis classes. Let $\phi_{h'}(h, x) = |h(x) - h'(x)|$ be a surrogate explanation loss of a constraint corresponding to h'. Let $\mathcal{G}_\mathcal{C} = \{\phi_{h'}(h, \cdot)|h \in \mathcal{H}, h' \in \mathcal{C}\}$, then we have*

$$R_n(\mathcal{G}_\mathcal{C}) \leq 2R_n(\mathcal{H}) + 2R_n(\mathcal{C}).$$

Proof.

$$
\begin{aligned}
R_n(\mathcal{G}_\mathcal{C}) &= \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\phi(h,\cdot)\in\mathcal{G}_\mathcal{C}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\phi(h,x_i)\right)\right]\right] \\
&= \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\substack{h\in\mathcal{H},\\ h'\in\mathcal{C}}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i|h(x_i)-h'(x_i)|\right)\right]\right] \\
&\leq \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\substack{h\in\mathcal{H},\\ h'\in\mathcal{C}}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\max(0,h(x_i)-h'(x_i))\right)\right]\right] \quad + \\
&\qquad \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\substack{h\in\mathcal{H},\\ h'\in\mathcal{C}}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\max(0,h'(x_i)-h(x_i))\right)\right]\right] \\
&\leq \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\substack{h\in\mathcal{H},\\ h'\in\mathcal{C}}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h(x_i)-h'(x_i))\right)\right]\right] \quad + \\
&\qquad \mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\substack{h\in\mathcal{H},\\ h'\in\mathcal{C}}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h'(x_i)-h(x_i))\right)\right]\right]
\end{aligned}
$$

where the lasts line again holds by an application of Talgrand's lemma. In this case, we indeed are optimizing over h', so we get that

$$
\begin{aligned}
R_n(\mathcal{G}_\mathcal{C}) &\leq 2\cdot\mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h(x_i))\right)\right]\right] + 2\cdot\mathop{\mathbb{E}}_{x\sim D}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h'\in\mathcal{C}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h'(x_i))\right)\right]\right] \\
&= 2R_n(\mathcal{H}) + 2R_n(\mathcal{C}).
\end{aligned}
$$

□

We remark that while this value is much larger than that of $R_n(\mathcal{H})$, we only need information about $\phi(h,x)$ and *not* the true label. Therefore, in many cases, this is preferable and not as expensive to learn.

F Proof of Theorem 2.2

We consider the agnostic setting of Theorem 2.2. Here, we have two notions of deviations; one is deviation in a model's ability to satisfy explanations, and the other is a model's ability to generalize to correctly produce the target function.

Proof. From Rademacher-based uniform convergence, for any $h \in \mathcal{H}$, with probability at least $1-\delta/2$ over S_E

$$|\phi(h,\mathcal{D}) - \phi(h,S_E)| \leq 2R_k(\mathcal{G}) + \sqrt{\frac{\ln(4/\delta)}{2k}} = \varepsilon_k$$

Therefore, with probability at least $1-\delta/2$, for any $h \in \mathcal{H}_{\phi,t-\varepsilon_k}$ we also have $\phi(h,S_E) \leq t$ and for any h with $\phi(h,S_E) \leq t$, we have $h \in \mathcal{H}_{\phi,t+\varepsilon_k}$. In addition, by a uniform convergence bound, with probability at least $1-\delta/2$, for any $h \in \mathcal{H}_{\phi,t+\varepsilon_k}$

$$|\mathrm{err}_\mathcal{D}(h) - \mathrm{err}_S(h)| \leq R_n(\mathcal{H}_{\phi,t+\varepsilon_k}) + \sqrt{\frac{\ln(4/\delta)}{2n}}.$$

Now, let h' be the minimizer of $\mathrm{err}_S(h)$ given that $\phi(h, S_E) \leq t$. By previous results, with probability $1-\delta$, we have $h' \in \mathcal{H}_{\phi,t+\varepsilon_k}$ and

$$\begin{aligned}\mathrm{err}_{\mathcal{D}}(h') &\leq \mathrm{err}_S(h') + R_n(\mathcal{H}_{\phi,t+\varepsilon_k}) + \sqrt{\frac{\ln(4/\delta)}{2n}} \\ &\leq \mathrm{err}_S(h^*_{t-\varepsilon_k}) + R_n(\mathcal{H}_{\phi,t+\varepsilon_k}) + \sqrt{\frac{\ln(4/\delta)}{2n}} \\ &\leq \mathrm{err}_{\mathcal{D}}(h^*_{t-\varepsilon_k}) + 2R_n(\mathcal{H}_{\phi,t+\varepsilon_k}) + 2\sqrt{\frac{\ln(4/\delta)}{2n}}.\end{aligned}$$

□

G A Generalization Bound in the Realizable Setting

In this section, we assume that we are in the doubly realizable Balcan & Blum (2010) setting where there exists $h^* \in \mathcal{H}$ such that $\mathrm{err}_{\mathcal{D}}(h^*) = 0$ and $\phi(h^*, \mathcal{D}) = 0$. The optimal classifier h^* lies in $\mathcal{H}$ and also achieve zero expected explanation loss. In this case, we want to output a hypothesis h that achieve both zero empirical risk and empirical explanation risk.

Theorem G.1 (Generalization bound for the doubly realizable setting)**.** *For a hypothesis class $\mathcal{H}$, a distribution $\mathcal{D}$ and an explanation loss ϕ. Assume that there exists $h^* \in \mathcal{H}$ that $\mathrm{err}_{\mathcal{D}}(h^*) = 0$ and $\phi(h^*, \mathcal{D}) = 0$. Let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is drawn i.i.d. from $\mathcal{D}$ and $S_E = \{x'_1, \ldots, x'_k\}$ drawn i.i.d. from $\mathcal{D}_{\mathcal{X}}$. With probability at least $1-\delta$, for any $h \in \mathcal{H}$ that $\mathrm{err}_S(h) = 0$ and $\phi(h, S_E) = 0$, we have*

$$\mathrm{err}_D(h) \leq R_n(\mathcal{H}_{\phi,\varepsilon_k}) + \sqrt{\frac{\ln(2/\delta)}{2n}}$$

when

$$\varepsilon_k = 2R_k(\mathcal{G}) + \sqrt{\frac{\ln(2/\delta)}{2k}}$$

when $\mathcal{G} = \{\phi(h, x) \mid h \in \mathcal{H}, x \in \mathcal{X}\}$.

Proof. We first consider only classifiers than has low empirical explanation loss and then perform standard supervised learning. From Rademacher-based uniform convergence, for any $h \in \mathcal{H}$, with probability at least $1 - \delta/2$ over S_E

$$\phi(h, \mathcal{D}) \leq \phi(h, S_E) + 2R_k(\mathcal{G}) + \sqrt{\frac{\ln(2/\delta)}{2k}}$$

when $\mathcal{G} = \{\phi(h, x) \mid h \in \mathcal{H}, x \in \mathcal{X}\}$. Therefore, for any $h \in \mathcal{H}$ with $\phi(h, S_E) = 0$, we have $h \in \mathcal{H}_{\phi,\varepsilon_k}$ with probability at least $1 - \delta/2$. Now, we can apply the uniform convergence on $\mathcal{H}_{\phi,\varepsilon_k}$. For any $h \in \mathcal{H}_{\phi,\varepsilon_k}$ with $\mathrm{err}_S(h) = 0$, with probability at least $1 - \delta/2$, we have

$$\mathrm{err}_{\mathcal{D}}(h) \leq R_n(\mathcal{H}_{\phi,\varepsilon_k}) + \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Therefore, for $h \in \mathcal{H}$ that $\phi(h, S_E) = 0, \mathrm{err}_S(h) = 0$, we have our desired guarantee. □

We remark that, since our result relies on the underlying techniques of the Rademacher complexity, our result is on the order of $O(\frac{1}{\sqrt{n}})$. In the (doubly) realizable setting, this is somewhat loose, and more complicated techniques are required to produce tighter bounds. We leave this as an interesting direction for future work.

H Proof of the Generalization Bound

We consider the agnostic setting of Theorem 2.2. Here, we have two notions of deviations; one is deviation in a model's ability to satisfy explanations, and the other is a model's ability to generalize to correctly produce the target function.

Proof. From Rademacher-based uniform convergence, for any $h \in \mathcal{H}$, with probability at least $1 - \delta/2$ over S_E

$$|\phi(h, \mathcal{D}) - \phi(h, S_E)| \leq 2R_k(\mathcal{G}) + \sqrt{\frac{\ln(4/\delta)}{2k}} = \varepsilon_k$$

Therefore, with probability at least $1 - \delta/2$, for any $h \in \mathcal{H}_{\phi, t-\varepsilon_k}$ we also have $\phi(h, S_E) \leq t$ and for any h with $\phi(h, S_E) \leq t$, we have $h \in \mathcal{H}_{\phi, t+\varepsilon_k}$. In addition, by a uniform convergence bound, with probability at least $1 - \delta/2$, for any $h \in \mathcal{H}_{\phi, t+\varepsilon_k}$

$$|\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)| \leq R_n(\mathcal{H}_{\phi, t+\varepsilon_k}) + \sqrt{\frac{\ln(4/\delta)}{2n}}.$$

Now, let h' be the minimizer of $\mathrm{err}_S(h)$ given that $\phi(h, S_E) \leq t$. By previous results, with probability $1 - \delta$, we have $h' \in \mathcal{H}_{\phi, t+\varepsilon_k}$ and

$$\begin{aligned}
\mathrm{err}_{\mathcal{D}}(h') &\leq \mathrm{err}_S(h') + R_n(\mathcal{H}_{\phi, t+\varepsilon_k}) + \sqrt{\frac{\ln(4/\delta)}{2n}} \\
&\leq \mathrm{err}_S(h^*_{t-\varepsilon_k}) + R_n(\mathcal{H}_{\phi, t+\varepsilon_k}) + \sqrt{\frac{\ln(4/\delta)}{2n}} \\
&\leq \mathrm{err}_{\mathcal{D}}(h^*_{t-\varepsilon_k}) + 2R_n(\mathcal{H}_{\phi, t+\varepsilon_k}) + 2\sqrt{\frac{\ln(4/\delta)}{2n}}.
\end{aligned}$$

□

I REALIZABLE SETTING

In this section, we assume that we are in the doubly realizable Balcan & Blum (2010) setting where there exists $h^* \in \mathcal{H}$ such that $\mathrm{err}_{\mathcal{D}}(h^*) = 0$ and $\phi(h^*, \mathcal{D}) = 0$. The optimal classifier h^* lies in $\mathcal{H}$ and also achieve zero expected explanation loss. In this case, we want to output a hypothesis h that achieve both zero empirical risk and empirical explanation risk.

Theorem I.1 (Generalization bound for the doubly realizable setting)**.** *For a hypothesis class $\mathcal{H}$, a distribution $\mathcal{D}$ and an explanation loss ϕ. Assume that there exists $h^* \in \mathcal{H}$ that $\mathrm{err}_{\mathcal{D}}(h^*) = 0$ and $\phi(h^*, \mathcal{D}) = 0$. Let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is drawn i.i.d. from $\mathcal{D}$ and $S_E = \{x'_1, \ldots, x'_k\}$ drawn i.i.d. from $\mathcal{D}_{\mathcal{X}}$. With probability at least $1 - \delta$, for any $h \in \mathcal{H}$ that $\mathrm{err}_S(h) = 0$ and $\phi(h, S_E) = 0$, we have*

$$\mathrm{err}_D(h) \leq R_n(\mathcal{H}_{\phi, \varepsilon_k}) + \sqrt{\frac{\ln(2/\delta)}{2n}}$$

when

$$\varepsilon_k = 2R_k(\mathcal{G}) + \sqrt{\frac{\ln(2/\delta)}{2k}}$$

when $\mathcal{G} = \{\phi(h, x) \mid h \in \mathcal{H}, x \in \mathcal{X}\}$.

Proof. We first consider only classifiers than has low empirical explanation loss and then perform standard supervised learning. From Rademacher-based uniform convergence, for any $h \in \mathcal{H}$, with probability at least $1 - \delta/2$ over S_E

$$\phi(h, \mathcal{D}) \leq \phi(h, S_E) + 2R_k(\mathcal{G}) + \sqrt{\frac{\ln(2/\delta)}{2k}}$$

when $\mathcal{G} = \{\phi(h, x) \mid h \in \mathcal{H}, x \in \mathcal{X}\}$. Therefore, for any $h \in \mathcal{H}$ with $\phi(h, S_E) = 0$, we have $h \in \mathcal{H}_{\phi, \varepsilon_k}$ with probability at least $1 - \delta/2$. Now, we can apply the uniform convergence on $\mathcal{H}_{\phi, \varepsilon_k}$. For any $h \in \mathcal{H}_{\phi, \varepsilon_k}$ with $\mathrm{err}_S(h) = 0$, with probability at least $1 - \delta/2$, we have

$$\mathrm{err}_{\mathcal{D}}(h) \leq R_n(\mathcal{H}_{\phi, \varepsilon_k}) + \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Therefore, for $h \in \mathcal{H}$ that $\phi(h, S_E) = 0, \mathrm{err}_S(h) = 0$, we have our desired guarantee. □

We remark that, since our result relies on the underlying techniques of the Rademacher complexity, our result is on the order of $O(\frac{1}{\sqrt{n}})$. In the (doubly) realizable setting, this is somewhat loose, and more complicated techniques are required to produce tighter bounds. We leave this as an interesting direction for future work.

J RADEMACHER COMPLEXITY OF LINEAR MODELS WITH A GRADIENT CONSTRAINT

Now, we present the theorem and proof for learning a linear model under a gradient constraint.

Theorem J.1 (Empirical Rademacher complexity of linear models with a gradient constraint)**.** *Let $\mathcal{X}$ be an instance space in $\mathbb{R}^d$, let $\mathcal{D}_\mathcal{X}$ be a distribution on $\mathcal{X}$, let $\mathcal{H} = \{h : x \to \langle w_h, x\rangle \mid w_h \in \mathbb{R}^d, ||w_h||_2 \leq B\}$ be a class of linear model with weights bounded by some constant $B > 0$ in ℓ_2 norm. Assume that there exists a constant $C > 0$ such that $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[||x||_2^2] \leq C^2$. Assume that we have an explanation constraints in term of gradient constraint; we want the gradient of our linear model to be close to the gradient some linear model h'. Let $\phi(h, x) = \theta(w_h, w_{h'})$ be an explanation surrogate loss when $\theta(u, v)$ is an angle between u, v. For any $S = \{x_1, \ldots, x_n\}$ is drawn i.i.d. from $\mathcal{D}_\mathcal{X}$, we have*

$$R_S(\mathcal{H}_{\phi,\tau}) = \frac{B}{n}\mathbb{E}_\sigma\left[\|v\| f(v)\right].$$

when $v = \sum_{i=1}^n x_i\sigma_i$ and

$$f(v) = \begin{cases} 1 & \text{when } \theta(v, w') \leq \tau \\ \cos(\theta(v, w') - \tau) & \text{when } \tau \leq \theta(v, w') \leq \frac{\pi}{2} + \tau \\ 0 & \text{when } \theta(v, w') \geq \frac{\pi}{2} + \tau. \end{cases}$$

Proof. Recall that $\mathcal{H}_{\phi,\tau} = \{h : x \to \langle w_h, x\rangle \mid w_h \in \mathbb{R}^d, ||w_h||_2 \leq B, \theta(w_h, w_{h'}) \leq \tau\}$. For a set of sample S, the empirical Rademacher complexity of $\mathcal{H}_{\phi,\tau}$ is given by

$$\begin{aligned} R_S(\mathcal{H}_{\phi,\tau}) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}_{\phi,\tau}} \sum_{i=1}^n h(x_i)\sigma_i\right] \\ &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\substack{\|w_h\|_2\leq B \\ \theta(w_h,w_{h'})\leq\tau}} \sum_{i=1}^n \langle w_h, x_i\rangle\sigma_i\right] \\ &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\substack{\|w_h\|_2\leq B \\ \theta(w_h,w_{h'})\leq\tau}} \langle w_h, \sum_{i=1}^n x_i\sigma_i\rangle\right]. \end{aligned}$$

For a vector $w' \in \mathbb{R}^d$ with $\|w'\|_2 = 1$, and a vector $v \in \mathbb{R}^d$, we will claim the following,

1. If $\theta(v, w') \leq \tau$, we have
$$\sup_{\substack{\|w\|_2\leq B \\ \theta(w,w')\leq\tau}} \langle w, v\rangle = B\|v\|.$$
2. If $\frac{\pi}{2} + \tau \leq \theta(v, w') \leq \pi$, we have
$$\sup_{\substack{\|w\|_2\leq B \\ \theta(w,w')\leq\tau}} \langle w, v\rangle = 0.$$
3. If $\tau \leq \theta(v, w') \leq \frac{\pi}{2} + \tau$, we have
$$\sup_{\substack{\|w\|_2\leq B \\ \theta(w,w')\leq\tau}} \langle w, v\rangle = B\|v\|\cos(\theta(v, w') - \tau)$$

For the first claim, we can see that if $\theta(v, w') \leq \tau$, we can pick $w = \frac{Bv}{\|v\|}$ and achieve the optimum value. For the second claim, we use the fact that $\theta(\cdot, \cdot)$ satisfies a triangle inequality and for any w

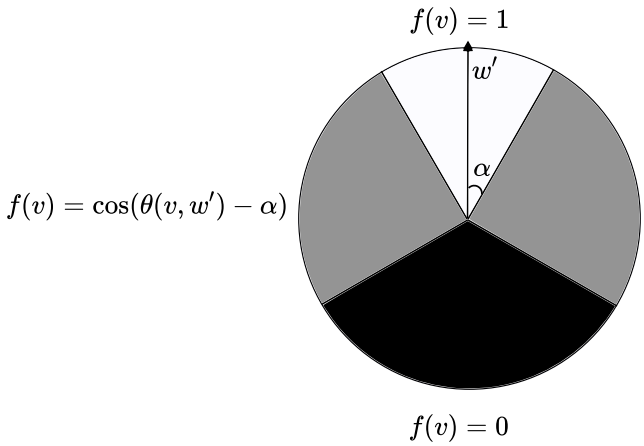

Figure 8: Illustration of different value of a function $f(v)$.

that $\theta(w, w') \leq \tau$, we have

$$\begin{aligned}\theta(v, w) + \theta(w, w') &\geq \theta(v, w') \\ \theta(v, w) &\geq \theta(v, w') - \theta(w, w') \\ \theta(v, w) &\geq \frac{\pi}{2} + \tau - \tau = \frac{\pi}{2}.\end{aligned}$$

This implies that for any w that $\theta(w, w') \leq \tau$, we have $\langle w, v\rangle = \|w\|\|v\|\cos(\theta(v, w)) \leq 0$ and the supremum is given by 0 where we can set $\|w\| = 0$. For the third claim, we know that $\langle w, v\rangle$ is maximum when the angle between v, w is the smallest. From the triangle inequality above, we must have $\theta(w, w') = \tau$ to be the largest possible value so that we have the smallest lower bound $\theta(v, w) \geq \theta(v, w') - \theta(w, w')$. In addition, the inequality holds when v, w', w lie on the same plane. Since we do not have further restriction on w, there exists such w and we have

$$\sup_{\substack{\|w\|_2 \leq B \\ \theta(w, w') \leq \tau}} \langle w, v\rangle = B\|v\|\cos(\theta(v, w') - \tau)$$

as required. One can calculate a closed form formula for w by solving a quadratic equation. Let $w = \frac{B\tilde{w}}{\|\tilde{w}\|}$ when $\tilde{w} = v + \lambda w'$ for some constant $\lambda > 0$ such that $\theta(w, w') = \tau$. With this we have an equation

$$\begin{aligned}\frac{\langle \tilde{w}, w'\rangle}{\|\tilde{w}\|} &= \cos(\tau) \\ \frac{\langle v + \lambda w', w'\rangle}{\|v + \lambda w'\|} &= \cos(\tau)\end{aligned}$$

Let $\mu = \langle v, w'\rangle$, solving for λ, we have

$$\begin{aligned}\frac{\mu + \lambda}{\sqrt{\|v\|^2 + 2\lambda\mu + \lambda^2}} &= \cos(\tau) \\ \mu^2 + 2\mu\lambda + \lambda^2 &= \cos^2(\tau)(\|v\|^2 + 2\lambda\mu + \lambda^2) \\ \sin^2(\tau)\lambda^2 + 2\sin^2(\tau)\mu\lambda + \mu^2 - \cos^2(\tau)\|v\|^2 &= 0 \\ \lambda^2 + 2\mu\lambda + \frac{\mu^2}{\sin^2(\tau)} - \cot^2(\tau)\|v\|^2 &= 0\end{aligned}$$

Solve this quadratic equation, we have

$$\lambda = -\mu \pm \cot(\tau)\sqrt{\|v\|^2 - \mu^2}.$$

Since $\lambda > 0$, we have $\lambda = -\mu + \cot(\tau)\sqrt{\|v\|^2 - \mu^2}$. We have

$$\begin{aligned}\tilde{w} &= v + \lambda w' \\ &= v + (-\mu + \cot(\tau)\sqrt{\|v\|^2 - \mu^2})w' \\ &= v - \langle v, w'\rangle w' + cot(\tau)w'\sqrt{\|v\|^2 - \mu^2}.\end{aligned}$$

With these claims, we have

$$\begin{aligned}R_S(\mathcal{H}_{\phi,\tau}) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\substack{\|w_h\|_2 \le B \\ \theta(w_h, w_{h'}) \le \tau}} \langle w_h, \sum_{i=1}^n x_i\sigma_i\rangle\right] \\ &= \frac{B}{n}\mathbb{E}_\sigma\left[\|v\|1\{\theta(v,w') \le \tau\} + \|v\|1\{\tau \le \theta(v,w') \le \frac{\pi}{2} + \tau\}\cos(\theta(v,w') - \tau)\right] \\ &= \frac{B}{n}\mathbb{E}_\sigma\left[\|v\|f(v)\right].\end{aligned}$$

□

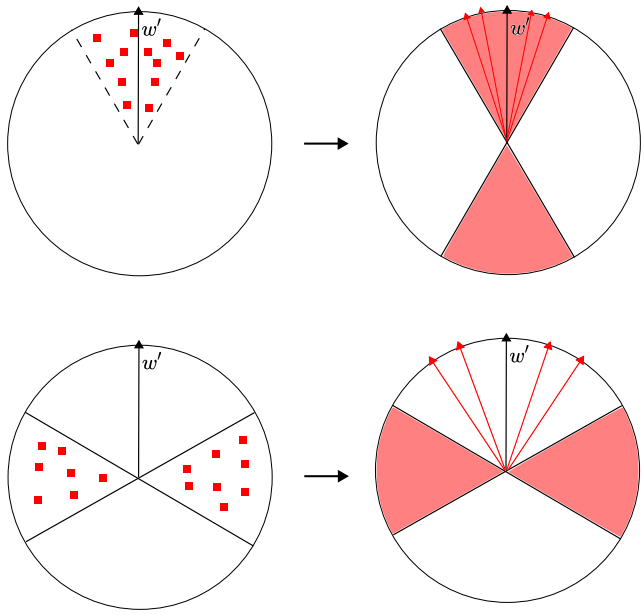

Figure 9: Benefits of an explanation constraints also depends on the data distribution. We represent data points x_i with red squares (Left). The possible regions for $v = \sum_{i=1}^n x_i\sigma_i$ are shaded area (Right). When the data is highly correlated with w', v would lie in a region where $f(v)$ is large (Top). When the data is almost orthogonal to w', v would lie in a region with a small value of $f(v)$ (Bottom)
.

As mentioned before, our restrictions may not be beneficial if the underlying data distribution is already concentrated about this restricted class of hypothesis. The bound above gives us a result that depends on the given linear model w'. In the case when $\theta(v, w') \le \tau$, we have that $v = \sum_{i=1}^n x_i\sigma_i$ is highly correlated with w'. In essence, this means that the data concentrated around the area of w', and the gradient constraint of being close to w' does not actually tell us much information (Figure 9 (Top)). This is illustrated by our bound not changing here, remaining as a factor of $\|v\|$. However, in the case when $\theta(v, w') \ge \tau$, we observe that the data is concentrated in regions other than near w'. This implies that many linear models, including those not close in angle to w' are valid. In this setting, the gradient information indeed restricts the $\mathcal{H}$ effectively (Figure 9 (Bottom)).

This is manifested in our bound, now on the order of $\cos(\theta(v, w') - \tau) \cdot \|v\|$. We remark that $\cos(\theta(v, w') - \tau)$ is the angle between v and a linear model that is within angle τ of w. As this increases (to the value of $\frac{\pi}{2}$), we have a smaller upper bound.

We now consider the case of a uniform distribution on a sphere to make these benefits more concrete. We utilize the symmetry of the uniform distribution over a sphere to derive an upper bound on the Rademacher complexity.

Theorem J.2 (Rademacher complexity of linear models with gradient constraint, uniform distribution on a sphere)**.** *Let $\mathcal{X}$ be an instance space in $\mathbb{R}^d$, let $\mathcal{D}_{\mathcal{X}}$ be a uniform distribution on a unit sphere in $\mathbb{R}^d$, let $\mathcal{H} = \{h : x \to \langle w_h, x\rangle \mid w_h \in \mathbb{R}^d, ||w_h||_2 \leq B\}$ be a class of linear model with weights bounded by some constant $B > 0$ in ℓ_2 norm. Assume that there exists a constant $C > 0$ such that $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[||x||_2^2] \leq C^2$. Assume that we have an explanation constraints in term of gradient constraint; we want the gradient of our linear model to be close to the gradient some linear model h'. Let $\phi(h, x) = \theta(w_h, w_{h'})$ be an explanation surrogate loss when $\theta(u, v)$ is an angle between u, v. We have*

$$R_n(\mathcal{H}_{\phi,\tau}) = \frac{B}{\sqrt{n}}\left(\sin(\tau)\cdot p + \frac{1-p}{2}\right),$$

where

$$p = \operatorname{erf}\left(\frac{\sqrt{d}\sin(\tau)}{\sqrt{2}}\right).$$

Proof. From Theorem J.1, we have that

$$\begin{aligned}
R_n(\mathcal{H}_{\phi,\tau}) &= \mathbb{E}[R_S(\mathcal{H}_{\phi,\tau})] \\
&= \frac{B}{n}\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_\sigma\left[\|v\|1\{\theta(v,w') \leq \tau\} + \|v\|1\{\tau \leq \theta(v,w') \leq \frac{\pi}{2}+\tau\}\cos(\theta(v,w')-\tau)\right]\right] \\
&= \frac{B}{n}\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_\sigma\left[\|v\|1\{\theta(v,w') \leq \frac{\pi}{2}-\tau\} + \|v\|1\{\frac{\pi}{2}-\tau \leq \theta(v,w') \leq \frac{\pi}{2}+\tau\}\cos(\theta(v,w')-\tau)\right]\right]
\end{aligned}$$

when $v = \sum_{i=1}^n x_i\sigma_i$. Because x_i is drawn uniformly from a unit sphere, in expectation $\theta(v, w')$ has a uniform distribution over $[0, \pi]$ and the distribution $\|v\|$ for a fixed value of $\theta(v, w')$ are the same for all $\theta(v, w') \in [0, \pi]$. From Trigonometry, we note that

$$\cos(\frac{\pi}{2} - 2\tau + a) + \cos(\frac{\pi}{2} - a) = \sin(2\tau - a) + \sin(a) \leq 2\sin(\tau).$$

By the symmetry property and the uniformity of the distribution of $\theta(v, w')$ and $\|v\|$.

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_\sigma\left[\|v\|1\{\frac{\pi}{2}-\tau \leq \theta(v,w') \leq \frac{\pi}{2}+\tau\}\cos(\theta(v,w')-\tau)\right]\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_\sigma\left[\|v\|1\{0 \leq \theta(v,w') \leq 2\tau\}\cos(\frac{\pi}{2}+\theta(v,w')-\tau)\right]\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_\sigma\left[\|v\|(1\{0 \leq \theta(v,w') \leq \tau\}\cos(\frac{\pi}{2}+\theta(v,w')-\tau) + 1\{\tau \leq \theta(v,w') \leq 2\tau\}\cos(\frac{\pi}{2}+\theta(v,w')-\tau))\right]\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_\sigma\left[\|v\|(1\{0 \leq \theta(v,w') \leq \tau\}\cos(\frac{\pi}{2}+\theta(v,w')-\tau) + 1\{0 \leq 2\tau-\theta(v,w') \leq \tau\}\cos(\frac{\pi}{2}-(2\tau-\theta(v,w'))))\right]\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_\sigma\left[\|v\|(1\{0 \leq \theta(v,w') \leq \tau\}\cos(\frac{\pi}{2}+\theta(v,w')-\tau) + 1\{0 \leq \tilde{\theta}(v,w') \leq \tau\}\cos(\frac{\pi}{2}-\tilde{\theta}(v,w')))\right]\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_\sigma\left[\|v\|(1\{0 \leq \theta(v,w') \leq \tau\}\cos(\frac{\pi}{2}+\theta(v,w')-\tau) + 1\{0 \leq \theta(v,w') \leq \tau\}\cos(\frac{\pi}{2}-\theta(v,w')))\right]\right] \\
&\leq \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_\sigma\left[\|v\|1\{\frac{\pi}{2}-\tau \leq \theta(v,w') \leq \frac{\pi}{2}+\tau\}\sin(\tau)\right]\right]
\end{aligned}$$

when $\tilde{\theta}(v, w') = \frac{\pi}{2} - \theta(v, w')$. We have

$$\begin{aligned}
R_n(\mathcal{H}_{\phi,\tau}) &\leq \frac{B}{n}\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_\sigma\left[\|v\|1\{\theta(v,w') \leq \frac{\pi}{2}-\tau\} + \|v\|1\{\frac{\pi}{2}-\tau \leq \theta(v,w') \leq \frac{\pi}{2}+\tau\}\sin(\tau)\right]\right] \\
&= \frac{B}{n}\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_\sigma\left[\|v\|\right]\right]\left(\Pr(\theta(v,w') \leq \frac{\pi}{2}-\tau) + \Pr(\frac{\pi}{2}-\tau \leq \theta(v,w') \leq \frac{\pi}{2}+\tau)\sin(\tau)\right)
\end{aligned}$$

The last equation follows from the symmetry and uniformity property. We can bound the first expectation

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\sigma}\|v\|]] &= \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\sigma}\|\sum_{i=1}^{n} x_i\sigma_i\|]] \\
&\leq \mathbb{E}_{\mathcal{D}}[\sqrt{\mathbb{E}_{\sigma}\|\sum_{i=1}^{n} x_i\sigma_i\|^2}]] \\
&= \mathbb{E}_{\mathcal{D}}[\sqrt{\mathbb{E}_{\sigma}\sum_{i=1}^{n}\|x_i\|^2\sigma_i^2}]] \\
&\leq C\sqrt{n}.
\end{aligned}$$

Next, we can simply note that, since our data is distributed over a unit sphere, each data has norm no greater than 1. Therefore, we know that $C = 1$ is indeed an upper bound on $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[||x||_2^2]$. For the probability term, we note that in expectation v has the same distribution as a random vector u drawn uniformly from a unit sphere. We let this be some probability p:

$$p = \Pr\left(\frac{\pi}{2} - \tau \leq \theta(v, w') \leq \frac{\pi}{2} + \tau\right) = \Pr\left(|\langle u, w'\rangle| \leq \sin(\tau)\right).$$

We know that the projection $\langle u, w'\rangle \sim \mathcal{N}(0, \frac{1}{d})$. Then, we have that $|\langle u, w'\rangle|$ is given by a Folded Normal Distribution, which has a CDF given by

$$\begin{aligned}
\Pr\left(|\langle u, w'\rangle| \leq \sin(\tau)\right) &= \frac{1}{2}\left[\operatorname{erf}\left(\frac{\sqrt{d}\sin(\tau)}{\sqrt{2}}\right) + \operatorname{erf}\left(\frac{\sqrt{d}\sin(\tau)}{\sqrt{2}}\right)\right] \\
&= \operatorname{erf}\left(\frac{\sqrt{d}\sin(\tau)}{\sqrt{2}}\right).
\end{aligned}$$

We then observe that

$$\begin{aligned}
\Pr\left(\theta(v, w') \leq \frac{\pi}{2} - \tau\right) &= \frac{1}{2}\left(1 - \Pr\left(\frac{\pi}{2} - \tau \leq \theta(v, w') \leq \frac{\pi}{2} + \tau\right)\right) \\
&= \frac{1-p}{2}
\end{aligned}$$

Plugging this in yields the following bound

$$R_n(\mathcal{H}_{\phi,\tau}) = \frac{B}{\sqrt{n}}\left(\sin(\tau)\cdot p + \frac{1-p}{2}\right),$$

where

$$p = \operatorname{erf}\left(\frac{\sqrt{d}\sin(\tau)}{\sqrt{2}}\right).$$

□

K Rademacher Complexity for Two Layer Neural Networks with a Gradient Constraint

Here, we present the full proof of the generalization bound for two layer neural networks with gradient explanations. In our proof, we use two results from Ma (2022). One result is a technical lemma, and the other is a bound on the Rademacher complexity of two layer neural networks.

Lemma K.1. *Consider a set $S = \{x_1, ..., x_n\}$ and a hypothesis class $\mathcal{F} \subset \{f : \mathbb{R}^d \to \mathbb{R}\}$. If*

$$\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(x_i)\sigma_i \geq 0 \text{ for any } \sigma_i \in \{\pm 1\}, i = 1, ..., n,$$

then, we have that

$$\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}|\sum_{i=1}^n f(x_i)\sigma_i|\right] \leq 2\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(x_i)\sigma_i\right].$$

Theorem K.2 (Rademacher complexity for two layer neural networks Ma (2022))**.** *Let $\mathcal{X}$ be an instance space and $\mathcal{D}_\mathcal{X}$ be a distribution over $\mathcal{X}$. Let $\mathcal{H} = \{h : x \mapsto \sum_{i=1}^m w_i\sigma(u_i^\top x)|w_i \in \mathbb{R}, u_i \in \mathbb{R}^d, \sum_{i=1}^m |w_i|\|u_i\|_2 \leq B\}$ be a class of two layer neural networks with m hidden nodes with a ReLU activation function $\sigma(x) = \max(0, x)$. Assume that there exists some constant $C > 0$ such that $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\|x\|_2^2] \leq C^2$. Then, for any $S = \{x_1, \ldots, x_n\}$ is drawn i.i.d. from $\mathcal{D}_\mathcal{X}$, we have that*

$$R_S(\mathcal{H}) \leq \frac{2B}{n}\sqrt{\sum_{i=1}^n ||x_i||_2^2}$$

and

$$R_n(\mathcal{H}) \leq \frac{2BC}{\sqrt{n}}.$$

We defer interested readers to Ma (2022) for the full proof of this result. Here, the only requirement of the data distribution is that $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\|x\|_2^2] \leq C^2$. We now present our result in the setting of two layer neural networks with one hidden node $m = 1$ to provide clearer intuition for the overall proof.

Theorem K.3 (Rademacher complexity for two layer neural networks ($m = 1$) with gradient constraints)**.** *Let $\mathcal{X}$ be an instance space and $\mathcal{D}_\mathcal{X}$ be a distribution over $\mathcal{X}$. Let $\mathcal{H} = \{h : x \mapsto w\sigma(u^\top x)|w \in \mathbb{R}, u \in \mathbb{R}^d, |w| \leq B, \|u\| = 1\}$. Without loss of generality, we assume that $\|u\| = 1$. Assume that there exists some constant $C > 0$ such that $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\|x\|_2^2] \leq C^2$. Our explanation constraint is given by a constraint on the gradient of our models, where we want the gradient of our learnt model to be close to a particular target function $h' \in \mathcal{H}$. Let this be represented by an explanation loss given by*

$$\phi(h, x) = \|\nabla_x h(x) - \nabla_x h'(x)\|_2 + \infty \cdot 1\{\|\nabla_x h(x) - \nabla_x h'(x)\| > \tau\}$$

for some $\tau > 0$. Let $h'(x) = w'\sigma((u')^\top x)$ the target function, then we have

$$R_n(\mathcal{H}_{\phi,\tau}) \leq \frac{\tau C}{\sqrt{n}} \qquad \text{if } |w'| > \tau,$$

$$R_n(\mathcal{H}_{\phi,\tau}) \leq \frac{3\tau C}{\sqrt{n}} \qquad \text{if } |w'| \leq \tau.$$

Proof. Our choice of $\phi(h, x)$ guarantees that, for any $h \in \mathcal{H}_{\phi,\tau}$, we have that $\|\nabla_x h(x) - \nabla_x h'(x)\| \leq \tau$ almost everywhere. We note that for $h(x) = w\sigma(u^\top x)$, the gradient is given by $\nabla_x h(x) = wu1\{u^\top x > 0\}$, which is a piecewise constant function over two regions (i.e., $u^\top x > 0, u^\top x \leq 0$), captured by Figure K.

We now consider $\nabla_x h(x) - \nabla_x h'(x)$, and we have 3 possible cases.

Case 1: $\theta(u, u') > 0$
This implies that the boundaries of $\nabla_x(h)$ and $\nabla_x h'(x)$ are different. Then, we have that $\nabla_x h(x) - \nabla_x h'(x)$ is a piecewise constant function with 4 regions, taking on values

$$\nabla_x h(x) - \nabla_x h'(x) = \begin{cases} wu - w'u' & \text{when } u^\top x > 0, (u')^\top x > 0 \\ wu & \text{when } u^\top x > 0, (u')^\top x < 0 \\ -w'u' & \text{when } u^\top x < 0, (u')^\top x > 0 \\ 0 & \text{when } u^\top x < 0, (u')^\top x < 0 \end{cases}$$

If we assume that each region has probability mass greater than 0 then our constraint $\|\nabla_x h(x) - \nabla_x h'(x)\|_2 \leq \tau$ implies that $|w| = |w|\|u\| \leq \tau, |w'| = |w'|\|u'\| \leq \tau, \|wu - w'u'\| \leq \tau$.

Case 2: $\theta(u, u') = 0$
This implies that the boundary of $\nabla_x h(x)$ and $\nabla_x h'(x)$ are the same. Then, $\nabla_x h(x) - \nabla_x h'(x)$ is a piecewise constant over two regions

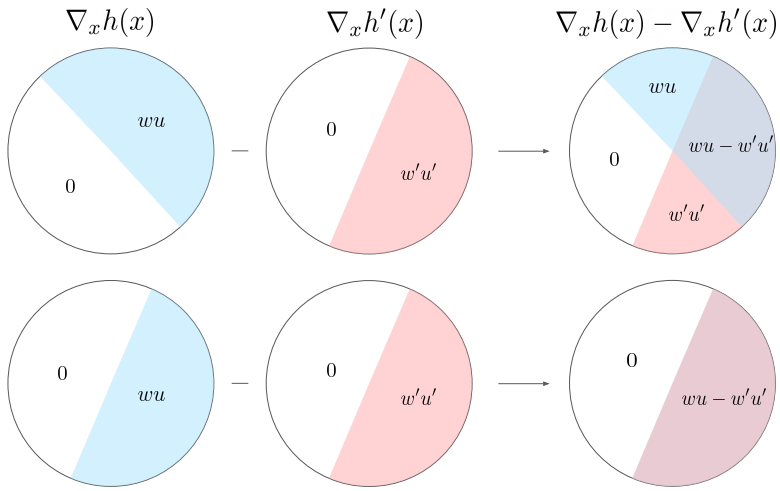

Figure 10: Visualization of the piecewise constant function of $\nabla_x h(x) - \nabla_x h'(x)$ over 4 regions.

$$\nabla_x h(x) - \nabla_x h'(x) = \begin{cases} wu - w'u' & \text{when } u^\top x > 0 \\ 0 & \text{when } u^\top x < 0 \end{cases}$$

This gives us that $|w - w'| = \|wu - w'u'\| \leq \tau$.

Case 3: $\theta(u, u') = \pi$
Here, we have that the decision boundaries of $\nabla_x h(x)$ and $\nabla_x h'(x)$ are the same but the gradients are non-zero on different sides. Then, $\nabla_x h(x) - \nabla_x h'(x)$ is a piecewise constant on two regions

$$\nabla_x h(x) - \nabla_x h'(x) = \begin{cases} wu & \text{when } u^\top x > 0 \\ -w'u' & \text{when } u^\top x < 0 \end{cases}$$

This gives us that $|w| \leq \tau$ and $|w'| \leq \tau$.

These different cases tell us that the constraint $\|\nabla_x h(x) - \nabla_x h'(x)\| \leq \tau$ reduces $\mathcal{H}$ into a class of models follows either

1. $u = u'$ and $|w - w'| < \tau$.
2. $u \neq u'$ and $|w| < \tau$. However, this case only possible when $|w'| < \tau$.

If $|w'| > \tau$, we know that we must only have the first case. Now, we can calculate the Rademacher complexity of our restricted class $\mathcal{H}_{\phi,\tau}$. We will again do this in separate cases.

Case 1: $|w'| > \tau$
For any $h \in \mathcal{H}_{\phi,\tau}$, we have that $u = u'$ and $|w - w'| < \tau$. For a sample $S = \{x_1, ..., x_n\}$,

$$\begin{aligned} R_s(\mathcal{H}_{\phi,\tau}) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{h \in \mathcal{H}_{\phi,\tau}} \sum_{i=1}^n h(x_i)\sigma_i\right] \\ &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{w} \sum_{i=1}^n w\sigma((u')^\top x_i)\sigma_i\right] && (\text{ as } u = u') \\ &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{w} w\left(\sum_{i=1}^n \sigma((u')^\top x_i)\sigma_i\right)\right]. \end{aligned}$$

Since, $|w - w'| < \tau$,

$$w' - \tau < w < w' + \tau$$

Then, we can compute the supremum over w as

$$w = \begin{cases} w' - \tau & \text{if } \left(\sum_{i=1}^{n} \sigma((u')^\top x_i)\sigma_i\right) < 0 \\ w' + \tau & \text{if } \left(\sum_{i=1}^{n} \sigma((u')^\top x_i)\sigma_i\right) \geq 0 \end{cases}$$

Therefore, we have

$$\sup_w w \left(\sum_{i=1}^{n} \sigma((u')^\top x_i)\sigma_i\right) = \left(w' + \tau \operatorname{sign}\left(\sum_{i=1}^{n} \sigma((u')^\top x_i)\sigma_i\right)\right) \cdot \left(\sum_{i=1}^{n} \sigma((u')^\top x_i)\sigma_i\right).$$

Now, we can calculate the Rademacher complexity as

$$\begin{aligned}
R_S(\mathcal{H}_{\phi,\tau}) &= \frac{1}{n}\mathbb{E}_\sigma\left[w'\left(\sum_{i=1}^{n} \sigma((u')^\top x_i)\sigma_i\right) + \tau|\sum_{i=1}^{n} \sigma((u')^\top x_i)\sigma_i|\right] \\
&= \frac{\tau}{n}\mathbb{E}_\sigma\left[|\sum_{i=1}^{n} \sigma((u')^\top x_i)\sigma_i|\right] \\
&\leq \frac{\tau}{n}\sqrt{\mathbb{E}_\sigma\left[\|\sum_{i=1}^{n} \sigma((u')^\top x_i)\sigma_i\|^2\right]} && \text{(Jensen's inequality)} \\
&= \frac{\tau}{n}\sqrt{\mathbb{E}_\sigma\left[\sum_{i=1}^{n} \sigma((u')^\top x_i)^2\sigma_i^2\right]} && \text{(since } \sigma_i, \sigma_j \text{ are independent with mean 0)} \\
&\leq \frac{\tau}{n}\sqrt{\sum_{i=1}^{n}((u')^\top x_i)^2} \\
&\leq \frac{\tau}{n}\sqrt{\sum_{i=1}^{n}\|x_i\|^2}.
\end{aligned}$$

Combining this with the fact that $\mathbb{E}\left[\|x\|^2\right] \leq C^2$, we have

$$\begin{aligned}
R_n(\mathcal{H}_{\phi,\tau}) &= \mathbb{E}[R_S(\mathcal{H}_{\phi,\tau})] \\
&\leq \frac{\tau}{n}\mathbb{E}[\sqrt{\sum_{i=1}^{n}\|x_i\|^2}] \\
&\leq \frac{\tau}{n}\sqrt{\mathbb{E}[\sum_{i=1}^{n}\|x_i\|^2]} && \text{(Jensen's inequality)} \\
&\leq \frac{\tau C}{\sqrt{n}}.
\end{aligned}$$

Case 2: $|w'|\|u'\| < \tau$.
We know that $\mathcal{H}_{\phi,\tau} = \mathcal{H}^{(1)}_{\phi,\tau} \bigcup \mathcal{H}^{(2)}_{\phi,\tau}$ when

$$\begin{aligned}
\mathcal{H}^{(1)}_{\phi,\tau} &= \{h \in \mathcal{H} | h : x \to w\sigma(u^\top x), u = u', |w - w'| < \tau\} \\
\mathcal{H}^{(2)}_{\phi,\tau} &= \{h \in \mathcal{H} | h : x \to w\sigma(u^\top x), \|u\| = 1, u \neq u', |w| < \tau\}
\end{aligned}$$

We have

$$\begin{aligned} R_S(\mathcal{H}_{\phi,\tau}) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}_{\phi,\tau}}\sum_{i=1}^n h(x_i)\sigma_i\right] \\ &\leq \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}^{(1)}_{\phi,\tau}}\sum_{i=1}^n h(x_i)\sigma_i + \sup_{h\in\mathcal{H}^{(2)}_{\phi,\tau}}\sum_{i=1}^n h(x_i)\sigma_i\right] \\ &= R_S(\mathcal{H}^{(1)}_{\phi,\tau}) + R_S(\mathcal{H}^{(2)}_{\phi,\tau}) \end{aligned}$$

The second line holds as $\sup_{x\in A\cup B} f(x) \leq \sup_{x\in A} f(x) + \sup_{x\in B} f(x)$ when $\sup_{x\in A} f(x) \geq 0$ and $\sup_{x\in B} f(x) \geq 0$. We know that both of these supremums be greater than zero, as we can recover the value of 0 with $w = 0$. From **Case 1**, we know that

$$R_n(\mathcal{H}^{(1)}_{\phi,\tau}) \leq \frac{\tau C}{\sqrt{n}}.$$

We also note that $\mathcal{H}^{(2)}_{\phi,\tau}$ is a class of two layer neural networks with weights with norms bounded by τ. From Theorem K.2, we have that

$$R_n(\mathcal{H}^{(2)}_{\phi,\tau}) \leq \frac{2\tau C}{\sqrt{n}}.$$

Therefore, in **Case 2**,

$$R_n(\mathcal{H}_{\phi,\tau}) \leq \frac{3\tau C}{\sqrt{n}}.$$

as required. □

Now, we consider in the general setting (i.e., no restriction on m).

Theorem K.4 (Rademacher complexity for two layer neural networks with gradient constraints). *Let $\mathcal{X}$ be an instance space and $\mathcal{D}_\mathcal{X}$ be a distribution over $\mathcal{X}$ with a large enough support. Let $\mathcal{H} = \{h : x \mapsto \sum_{j=1}^m w_j\sigma(u_j^\top x) | w_j \in \mathbb{R}, u_j \in \mathbb{R}^d, \|u_j\|_2 = 1, \sum_{j=1}^m |w_j| \leq B\}$. Assume that there exists some constant $C > 0$ such that $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\|x\|_2^2] \leq C^2$. Our explanation constraint is given by a constraint on the gradient of our models, where we want the gradient of our learnt model to be close to a particular target function $h' \in \mathcal{H}$. Let this be represented by an explanation loss given by*

$$\phi(h, x) = \|\nabla_x h(x) - \nabla_x h'(x)\|_2 + \infty \cdot 1\{\|\nabla_x h(x) - \nabla_x h'(x)\| > \tau\}$$

for some $\tau > 0$. Then, we have that

$$R_n(\mathcal{H}_{\phi,\tau}) \leq \frac{3\tau m C}{\sqrt{n}}.$$

To be precise,

$$R_n(\mathcal{H}_{\phi,\tau}) \leq \frac{(2m+q)\tau C}{\sqrt{n}}.$$

when q is the number of node j of h' such that $|w'_j| < \tau$.

We note that this result indeed depends on the number of hidden dimensions m; however, we note that in the general case (Theorem K.2), the value of B is $O(m)$ as it is a sum over the values of each hidden node. We now present the proof for the more general version of our theorem.

Proof. For simplicity, we first assume that any $h \in \mathcal{H}$ has that $\|u_j\| = 1, \forall j$. Consider $h \in \mathcal{H}$, we write $h = \sum_{j=1}^m w'_j\sigma((u'_j)^\top x)$ and let $h'(x) = \sum_{j=1}^m w'_j\sigma((u'_j)^\top x)$ be a function for our gradient constraint. The gradient of a hypothesis h is given by

$$\nabla_x h(x) = \sum_{j=1}^m w_j u_j \cdot 1\{u_j^\top x > 0\},$$

which is a piecewise constant function over at most 2^m regions. Then, we consider that

$$\nabla_x h(x) - \nabla_x h'(x) = \sum_{j=1}^{m} w_j u_j \cdot 1\{u_j^\top x > 0\} - \sum_{j=1}^{m} w_j' u_j' \cdot 1\{(u_j')^\top x > 0\},$$

which is a piecewise constant function over at most 2^{2m} regions. We again make an assumption that each of these regions has a non-zero probability mass. Our choice of $\phi(h, x)$ guarantees that the norm of the gradient in each region is less than τ. Similar to the case with $m = 1$, we will show that the gradient constraint leads to a class of functions with the same decision boundary or neural networks that have weights with a small norm.

Assume that among $u_1, ..., u_m$ there are k vectors that have the same direction as $u_1', ..., u_m'$. Without loss of generality, let $u_j = u_j'$ for $j = 1, ..., k$. In this case, we have that $\nabla_x h(x) - \nabla_x h'(x)$ is a piecewise function over 2^{2m-k} regions. As each region has non-zero probability mass, for each $j \in \{1, ..., k\}$, we know that $\exists x$ such that

$$u_j^\top x = (u_j')^\top x > 0, \qquad u_i^\top x < 0 \text{ for } i \neq j, \qquad (u_i')^\top x < 0 \text{ for } i \neq j.$$

In other words, we can observe a data point from each region that uniquely defines the value of a particular w_j, u_j. In this case, we have that

$$\begin{aligned} \nabla_x h(x) - \nabla_x h'(x) &= w_j u_j - w_j' u_j' \\ &= (w_j - w_j') u_j'. \end{aligned}$$

From our gradient constraint, we know that $||\nabla_x h(x) - \nabla_x h'(x)|| \leq \tau, \forall x$, which implies that $|w_j - w_j'| \leq \tau$ for $j = 1, ..., k$.

On the other hand, for the remaining $j = k + 1, ..., m$, we know that there exists x such that

$$u_j^\top x > 0, \qquad u_i^\top x < 0 \text{ for } i \neq j, \qquad (u_i')^\top x < 0 \text{ for } i = 1, ..., m.$$

Then, we have that $\nabla_x h(x) = w_j u_j$, and our constraint implies that $|w_j| \|u_j\| = |w_j| \leq \tau$. Similarly, we have that $|w_j'| \|u_j'\| = |w_j'| < \tau$, for $j = k + 1, ..., m$. We can conclude that $\mathcal{H}_{\phi,\tau}$ is a class of two layer neural networks with m hidden nodes (assuming $\|u_i\|$ = 1) that for each node $w_j \sigma(u_j^\top x)$ satisfies

1. There exists $l \in [m]$ that $u_j = u_l'$ and $|w_j - w_l'| < \tau$.
2. $|w_j| < \tau$

We further note that for a node $w_l' \sigma((u_l')^\top x)$ in $h'(x)$ that has that a high weight $|w_l'| > \tau$, there must be a node $w_j \sigma(u_j^\top x)$ in h with the same boundary $u_j = u_l$. Otherwise, there is a contradiction with $|w_l'| < \tau$ for all nodes in h' without a node in h with the same boundary. We can utilize this characterization of the restricted class $\mathcal{H}_{\phi,\tau}$ to bound the Rademacher complexity of the class. Let

$$\mathcal{H}' = \{h : x \mapsto \sum_{j=1}^{m} w_j' \sigma((u_j')^\top x) a_j \mid a_j \in \{0, 1\} \text{ and for } j \text{ that } |w_j'| > \tau, a_j = 1\}.$$

This is a class of two layer neural networks with at most m nodes such that each node is from h'. We also have a condition that if the weight of the j-th node in h' is greater than τ, the j-th node must be present in any member of this class. Let

$$\mathcal{H}^{(\tau)} = \{h : x \mapsto \sum_{j=1}^{m} w_j \sigma((u_j)^\top x) a_j \mid w_j \in \mathbb{R}, u_j \in \mathbb{R}^d, |w_j| < \tau, \|u_j\| = 1\}.$$

be a class of two layer neural networks with m nodes such that the weight of each node is at most τ. We claim that for any $h \in \mathcal{H}_{\phi,\tau}$ there exists $h_1 \in \mathcal{H}', h_2 \in \mathcal{H}^{(\tau)}$ that $h = h_1 + h_2$. For any $h \in \mathcal{H}_{\phi,\tau}$, let $p_h : [m] \to [m] \cup \{0\}$ be a function that match a node in h with the node with the same boundary in h'. Formally,

$$p_h(j) = \begin{cases} l & \text{when } u_j = u_l' \\ 0 & \text{otherwise.} \end{cases}$$

The function p_h maps j to 0 if there is no node in h' with the same boundary. Let $w'_0 = 0, u'_0 = [0, \dots, 0]$, we can write

$$
\begin{aligned}
h(x) &= \sum_{j=1}^{m} w_j \sigma(u_j^\top x) \\
&= \sum_{j=1}^{m} w_j \sigma(u_j^\top x) - w'_{p_h(j)} \sigma((u')^\top_{p_h(j)} x) + w'_{p_h(j)} \sigma((u')^\top_{p_h(j)} x) \\
&= \underbrace{\sum_{p_h(j) \neq 0} (w_j - w'_{p_h(j)}) \sigma((u')^\top_{p_h(j)} x) + \sum_{p_h(j)=0} w_j \sigma(u_j^\top x)}_{\in \mathcal{H}^{(\tau)}} + \underbrace{\sum_{p(j) \neq 0} w'_{p_h(j)} \sigma((u')^\top_{p_h(j)} x)}_{\in \mathcal{H}'} .
\end{aligned}
$$

The first term is a member of $\mathcal{H}^{(\tau)}$ because we know that $|w_j - w'_{p(j)}| < \tau$ or $|w_j| < \tau$. The second term is also a member of $\mathcal{H}'$ since for any l that $|w'_l| > \tau$, there exists j that $p_h(j) = l$. Therefore, we can write h in terms of a sum between a member of $\mathcal{H}'$ and $\mathcal{H}^{(\tau)}$. This implies that

$$
R_n(\mathcal{H}_{\phi,\tau}) \leq R_n(\mathcal{H}') + R_n(\mathcal{H}^{(\tau)}).
$$

From Theorem K.2, we have that

$$
R_n(\mathcal{H}^{(\tau)}_{\phi,\tau}) \leq \frac{2\tau m C}{\sqrt{n}}.
$$

Now, we will calculate the Rademacher complexity of $\mathcal{H}'$. For $S = \{x_1, \dots, x_n\}$,

$$
\begin{aligned}
R_S(\mathcal{H}') &= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}'} \sum_{i=1}^{n} h(x_i) \sigma_i \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}'} \sum_{i=1}^{n} (\sum_{j=1}^{m} w'_j \sigma((u'_j)^\top x_i) a_j) \sigma_i \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}'} \sum_{i=1}^{n} (\sum_{|w'_j| < \tau} w'_j \sigma((u'_j)^\top x_i) a_j + \sum_{|w'_j| > \tau} w'_j \sigma((u'_j)^\top x_i)) \sigma_i \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{a_j \in \{0,1\}} \sum_{i=1}^{n} \sum_{|w'_j| < \tau} w'_j \sigma((u'_j)^\top x_i) a_j \sigma_i \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{a_j \in \{0,1\}} \sum_{|w'_j| < \tau} a_j (w'_j \sum_{i=1}^{n} \sigma((u'_j)^\top x_i) \sigma_i) \right] .
\end{aligned}
$$

To achieve the supremum, if $w'_j \sum_{i=1}^n \sigma((u'_j)^\top x_i)\sigma_i > 0$ we need to set $a_j = 1$, otherwise, we need to set $a_j = 0$. Therefore,

$$
\begin{aligned}
R_S(\mathcal{H}') &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{a_j\in\{0,1\}} \sum_{|w'_j|<\tau} a_j(w'_j \sum_{i=1}^n \sigma((u'_j)^\top x_i)\sigma_i)\right] \\
&= \frac{1}{n}\mathbb{E}_\sigma\left[\sum_{|w'_j|<\tau} \sigma(w'_j \sum_{i=1}^n \sigma((u'_j)^\top x_i)\sigma_i)\right] \\
&= \frac{1}{2n}\mathbb{E}_\sigma\left[\sum_{|w'_j|<\tau} (w'_j \sum_{i=1}^n \sigma((u'_j)^\top x_i)\sigma_i) + |w'_j \sum_{i=1}^n \sigma((u'_j)^\top x_i)\sigma_i|\right] && (\sigma(x) = \frac{x+|x|}{2}) \\
&= \frac{1}{2n}\mathbb{E}_\sigma\left[\sum_{|w'_j|<\tau} |w'_j \sum_{i=1}^n \sigma((u'_j)^\top x_i)\sigma_i|\right] \\
&\leq \frac{1}{2n}\left(\sum_{|w'_j|<\tau} |w'_j|\right)\mathbb{E}_\sigma\left[\sup_{\|u\|=1} |\sum_{i=1}^n \sigma(u^\top x_i)\sigma_i|\right] \\
&\leq \frac{1}{n}\left(\sum_{|w'_j|<\tau} |w'_j|\right)\mathbb{E}_\sigma\left[\sup_{\|u\|=1} \sum_{i=1}^n \sigma(u^\top x_i)\sigma_i\right] && (\text{Lemma } K.1) \\
&\leq \left(\sum_{|w'_j|<\tau} |w'_j|\right) \underbrace{\mathbb{E}_\sigma\left[\frac{1}{n}\sup_{\|u\|=1} \sum_{i=1}^n u^\top x_i\sigma_i\right]}_{\text{Empirical Rademacher complexity of a linear model}} && (\text{Talagrand's Lemma}).
\end{aligned}
$$

From Theorem B.1, we can conclude that

$$R_n(\mathcal{H}') \leq \sum_{|w'_j|<\tau} |w'_j|\frac{C}{\sqrt{n}} \leq \frac{q\tau C}{\sqrt{n}} \leq \frac{m\tau C}{\sqrt{n}}$$

when q is the number of nodes j of h' such that $|w'_j| < \tau$. Therefore,

$$R_n(\mathcal{H}') \leq \frac{(2m+q)\tau C}{\sqrt{n}} \leq \frac{3m\tau C}{\sqrt{n}}.$$

□

A tighter bound is given by $\frac{(2m+q)\tau C}{\sqrt{n}}$ when q is the number of w'_j that $|w'_j| < \tau$. As $\tau \to 0$, we also have $q \to 0$. This implies that we have an upper bound of $\frac{2m\tau C}{\sqrt{n}}$ if τ is small enough. When comparing this to the original bound $\frac{2BC}{\sqrt{n}}$, we can do much better if $\tau \ll \frac{B}{m}$. We would like to point out that our bound does not depend on the distribution $\mathcal{D}$ because we choose a strong explanation loss

$$\phi(h, x) = \|\nabla_x h(x) - \nabla_x h'(x)\|_2 + \infty \cdot 1\{\|\nabla_x h(x) - \nabla_x h'(x)\| > \tau\}$$

which guarantees that $\|\nabla_x h(x) - \nabla_x h'(x)\|_2 \leq \tau$ almost everywhere. We also assume that we are in a high-dimensional setting $d \gg m$ and there exists x with a positive probability density at any partition created by $\nabla_x h(x)$.

L Algorithmic Results for 2 Layer NNs with a Gradient Constraint

Now that we have provided generalization bounds for the restricted class of 2 layer neural networks, we also present an algorithm that can identify the parameters of a 2 layer neural network (up to

a permutation of the weights). In practice, we might solve this via our variational objective or other simpler regularized techniques. However, we also provide a theoretical result for the required amount of data (given some assumptions about the data distribution) and runtime for an algorithm to exactly recover the parameters of these networks under gradient constraints.

We again know that the gradient of 2 layer neural networks with ReLU activations can be written as

$$\nabla_x f_{w,U}(x) = \sum_{i=1}^{m} w_i u_i \cdot 1\{u_i^T x > 0\},$$

where we consider $||u_i|| = 1$. Therefore, an exact gradient constraint given of the form of pairs $(x, \nabla_x f(x))$ produces a system of equations.

Proposition L.1. *If the values of u_i's are known, we can identify the parameters w_i with exactly m fixed samples.*

Proof. We can select m datapoints, which each achieve value 1 for the indicator value in the gradient of the 2 layer neural network. This would give us m equations, which each are of the form

$$\nabla_x f_{w,U}(x_i) = w_i u_i.$$

Therefore, we can easily solve for the values of w_i, given that u_i is known. □

To make this more general, we now consider the case where u_i's are not known but are at least linearly independent.

Proposition L.2. *Let the u_i's be linearly independent. Assume that each region of the data (when partitioned by the values of u_i) has non-trivial support $> p$. Then, with probability $1 - \delta$, we can identify the parameters w_i, u_i with $O\left(2^m + \frac{m+\log(\frac{1}{\delta})}{\log(\frac{1}{1-p})}\right)$ data points and in $O(2^{2m})$ time.*

Proof. Let us partition $\mathcal{X}$ into regions satisfying unique values of the binary vector $(1\{u_1^T x > 0\}, ..., 1\{u_m^T x > 0\})$, which by our assumption each have at least some probability mass p. First, we calculate the probability that we observe one data point with an explanation from each region in this partition. This is equivalent to sampling from a multinomial distribution with probabilities $(p_1, ..., p_{2^m})$, where $p_i \geq p, \forall i$. Then,

$$\begin{aligned}\Pr(\text{observe all regions in } n \text{ draws}) &= 1 - \Pr(\exists i \text{ s.t. we do not observe region } i) \\ &= 1 - 2^m(1-p)^n.\end{aligned}$$

Setting this as no less than $1 - \delta$ leads to that $n \geq \frac{m+\log(\frac{1}{\delta})}{\log(\frac{1}{1-p})}$.

Given $O(2^m + \frac{m+\log(\frac{1}{\delta})}{\log(\frac{1}{1-p})})$ pairs of data and gradients, we will observe at least one pair from each region of the partition. Then, identifying the values of u_i's and w_i's is equivalent to identifying the datapoints that correspond to a value of the binary vector where only one indicator value is 1. These values can be identified in $O(2^{3m})$ time; the algorithm is given in Algorithm L.1. These results demonstrate that we can indeed learn the parameters (up to a permutation) of a 2 layer neural network given exact gradient information. □

L.1 Algorithm for identifying regions

We first note that identifying the parameters u_i's and w_i's of a 2 layer neural network is equivalent to identifying the values $\{x_1, ..., x_m\}$ from the set $\{\sum_{x \in C} x | C \in \mathcal{P}(\{x_1, ..., x_m\})\}$, where $\mathcal{P}$ denotes the power set. We also assume that $x_1, ..., x_m$ are linearly independent, so that we cannot create x_i from any linear combination of x_j's with $j \neq i$. Then, we can identify the set $\{x_1, ..., x_m\}$ as follows:

This algorithm runs in $O(2^{3m})$ time as it iterates through each point in M and computes the overlapping set O and resulting updated sum S, which takes $O(2^{2m})$ time. From the resulting set B, we can exactly compute values u_i and w_i up to a permutation.

Algorithm 1 Algorithm for identifying parameters of a 2 layer neural network, given exact gradient constraints

Input: We are given $M = \{\sum_{x \in C} x | C \in \mathcal{P}(\{x_1, ..., x_m\})\}$, with $\{x_1, ..., x_m\}$ linearly independent
Output: The set of basis elements $\{x_1, ..., x_m\}$
function
  $B = \{\}, S = \{\}$ {Set for basis vectors and set for a current sum of at least 2 elements}
  for $x \in M$ **do**
    if $x \in S$ **then**
      pass
    else
      $B = B \cup \{x\}$
      if $|B| = 2$ **then**
        $S = \{y_1 + y_2\}$, where $B = \{y_1, y_2\}$
      else
        $S = S \cup \{y + x | y \in S\}$ {Updating sums from adding x}
      end if
      $O = B \cap S$ {Computing overlap between current basis and sums}
      $B = B \setminus O$ {Removing elements contained in pairwise span}
      $S = \{y - y_o | y \in S, y_o \in O\}$ {Updating sums S from removing set O}
    end if
  end for
  return B
end function

M ADDITIONAL SYNTHETIC EXPERIMENTS

We now present additional synthetic experiments that demonstrate the performance of our approach under settings with imperfect explanations and compare the benefits of using *different types* of explanations.

M.1 VARIATIONAL OBJECTIVE IS BETTER WITH NOISY GRADIENT EXPLANATIONS

Here, we present the remainder of the results from the synthetic regression task of 5, under more settings of noise ϵ added to the gradient explanation.

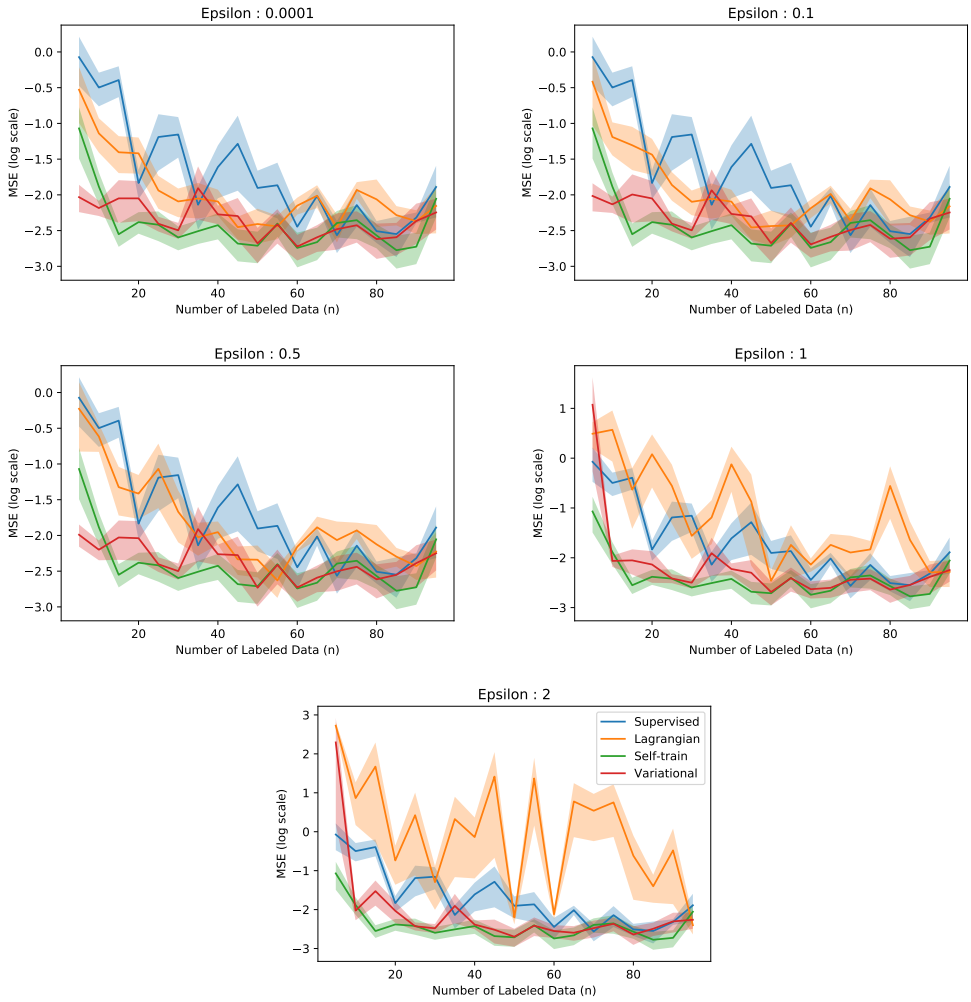

Figure 11: Comparison of MSE on regressing a 2 Layer Neural Network with explanations of noisy gradients. $m = 1000, k = 20, \lambda = 10$. For the iterative methods, $T = 10$. Results are averaged over 5 seeds.

Again, we observe that our method does better than that of the Lagrangian approach and the self-training method. Under high levels of noise, the Lagrangian method does poorly. On the contrary, our method is resistant to this noise and also outperforms self-training significantly in settings with limited labeled data.

M.2 COMPARING DIFFERENT TYPES OF EXPLANATIONS

Here, we present synthetic results to compare using different types of explanation constraints. We focus on comparing noisy gradients as before, as well as noisy classifiers, which are used in the setting of weak supervision (Ratner et al., 2016). Here, we generate our noisy classifiers as $h^*(x)+\epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We omit the results of self-training as it does not use any explanations, and we keep the supervised method as a baseline. Here, $t = 0.25$.

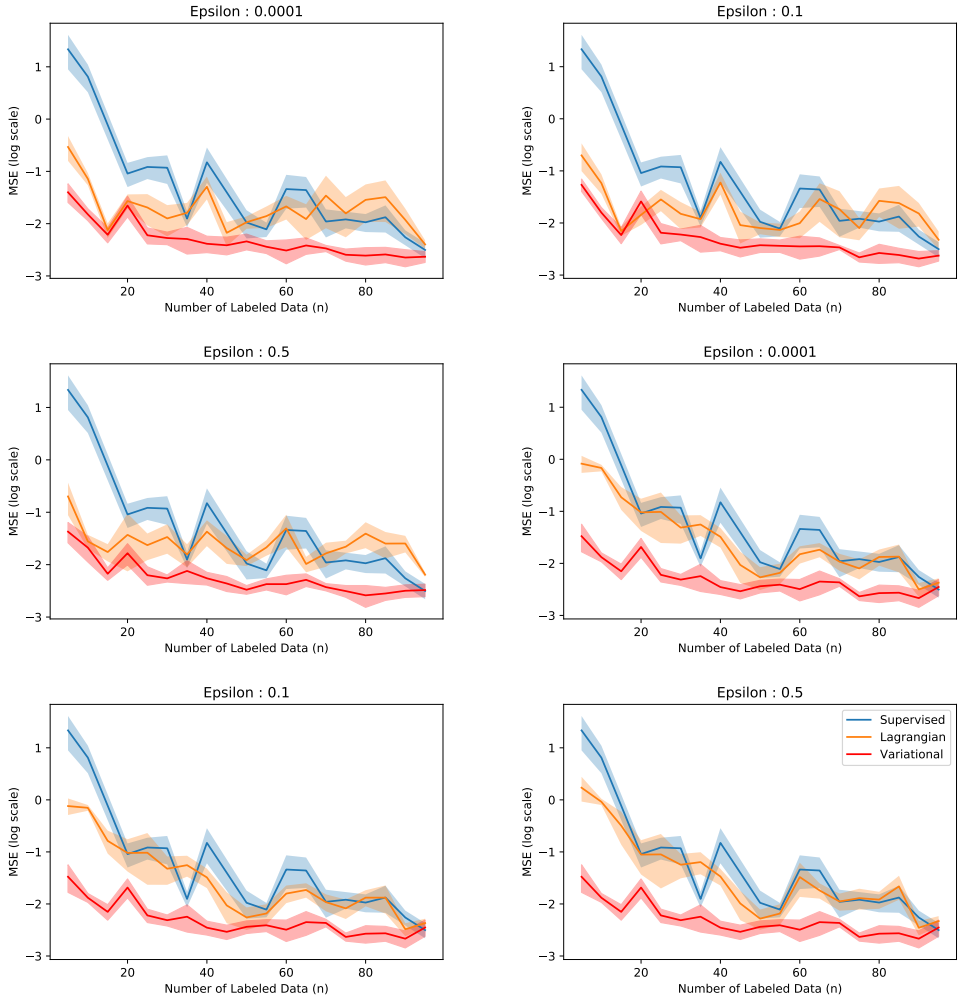

Figure 12: Comparison of MSE on regressing a 2 Layer Neural Network with explanations as a noisy classifier (top) and noisy gradients (bottom). $m = 1000, k = 20$. For the iterative methods, $T = 10$. Results are averaged over 5 seeds. ϵ represents the variance of the noise added to the noisy classifier or noisy gradient.

We observe different trends in performance as we vary the amount of noise in the noisy gradient or noisy classifier explanations. With any amount of noise and sufficient regularization (λ), this influences the overall performance of the methods that incorporate constraints. With few labeled data, using noisy classifiers helps outperform standard supervised learning. With a larger amount of labeled data, this leads to no benefits (if not worse performance of the Lagrangian approach). However, with the noisy gradient, under small amounts of noise, the restricted class of hypothesis will still capture solutions with low error. Therefore, in this case, we observe that the Lagrangian approach outperforms standard supervised learning in the case with few labeled data and matches it with sufficient labeled data. Our method outperforms or matches both methods across all settings.

We consider another noisy setting, where noise has been added to the weights of a copy of the target 2 layer neural network. Here, we compare how this information impacts learning from the direct outputs (noisy classifier) or the gradients (noisy gradients) of that noisy copy (Figure 13).

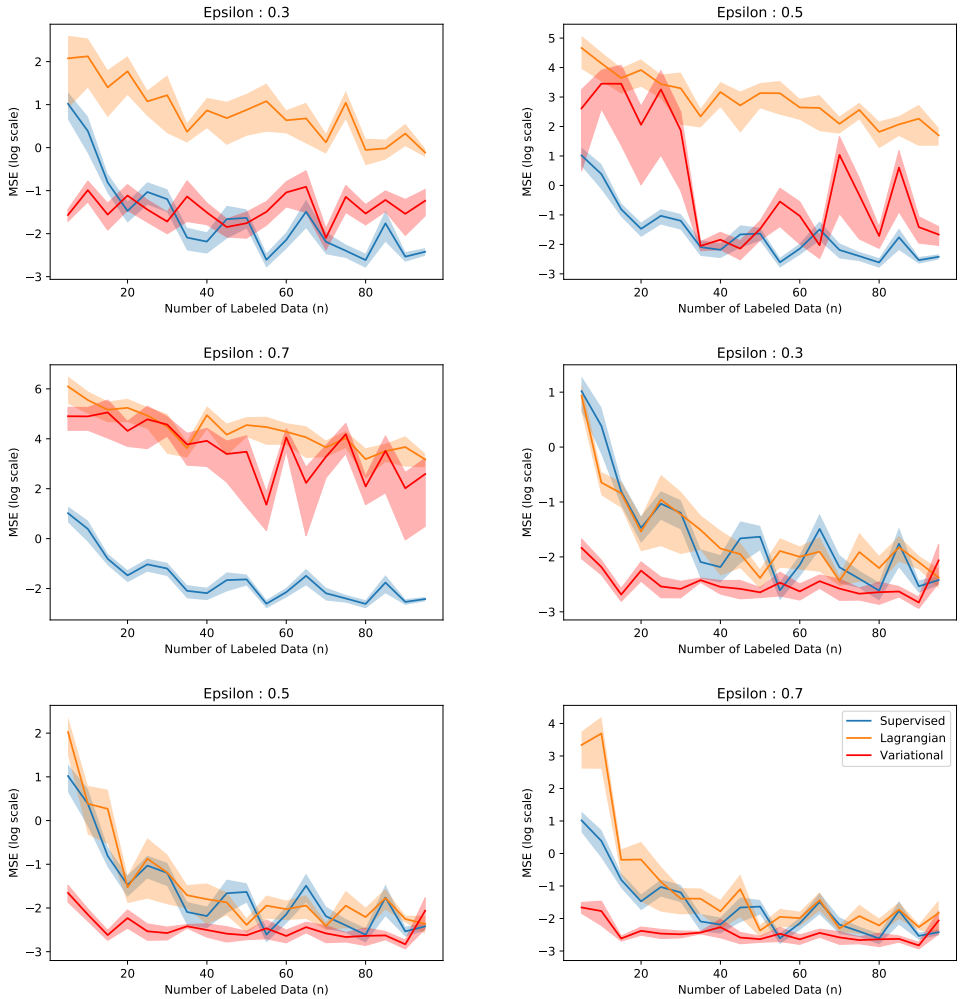

Figure 13: Comparison of MSE on regressing a 2 Layer Neural Network with explanations as a noisy classifier (top) and noisy gradients (bottom). $m = 1000, k = 20$. For the iterative methods, $T = 10$. Results are averaged over 5 seeds. ϵ represents the variance of the noise added to the noisy classifier or noisy gradient.

N Experimental Details

For all of our synthetic and real-world experiments, we use values of $m = 1000, k = 20, T = 3, t = 0, \lambda = 1$, unless otherwise noted. For our synthetic experiments, we use $d = 100, \sigma^2 = 5$. Our 2 layer neural networks have hidden dimensions of size 10. They are trained with a learning rate of 0.01 for 50 epochs. We evaluate all networks on a (synthetic) test set of size 2000.

For our real-world data, our 2 layer neural networks have a hidden dimension of size 10 and are trained with a learning rate of 0.1 (YouTube) and 0.1 (Yelp) for 10 epochs. $\lambda = 0.01$ and gradient values computed by the smoothed approximation in (Sam & Kolter, 2022) has $c = 1$. Test splits are used as follows from the YouTube and Yelp datasets in the WRENCH benchmark (Zhang et al., 2021).

We choose the initialization of our variational algorithm h_0 as the standard supervised model, trained using gradient descent.

O ABLATIONS

We also perform ablation studies in the same regression setting as Section 4. We vary parameters that determine either the experimental setting or hyperparameters of our algorithms.

O.1 NUMBER OF EXPLANATIONS

First, we vary the value of k to illustrate the benefits of our method over the existing baselines.

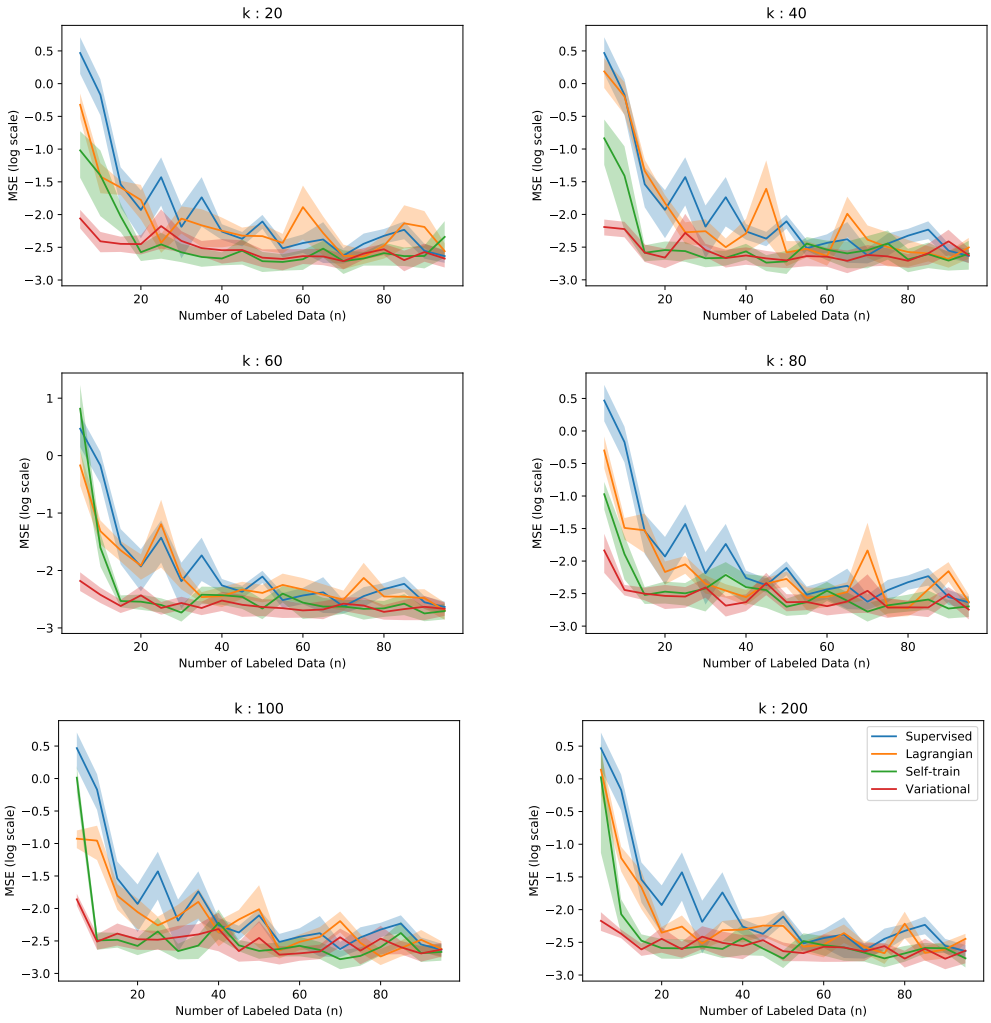

Figure 14: Comparison of MSE on regressing a 2 layer neural networks over different amounts of explanation annotated data k. $m = 1000$. For the iterative methods, $T = 10$. Results are averaged over 5 seeds.

We observe that our variationa approach performs much better than a simple augmented Lagrangian method, which in turn does better than supervised learning with sufficiently large values of k. Our approach is always better than the standard supervised approach.

We also provide results for how well these methods satisfy these explanations over varying values of k.

O.2 SIMPLER TEACHER MODELS CAN MAINTAIN GOOD PERFORMANCE

As noted before, we can use *simpler* teacher models to be regularized into the explanation-constrained subspace. This can lead to overall easier optimization problems, and we synthetically

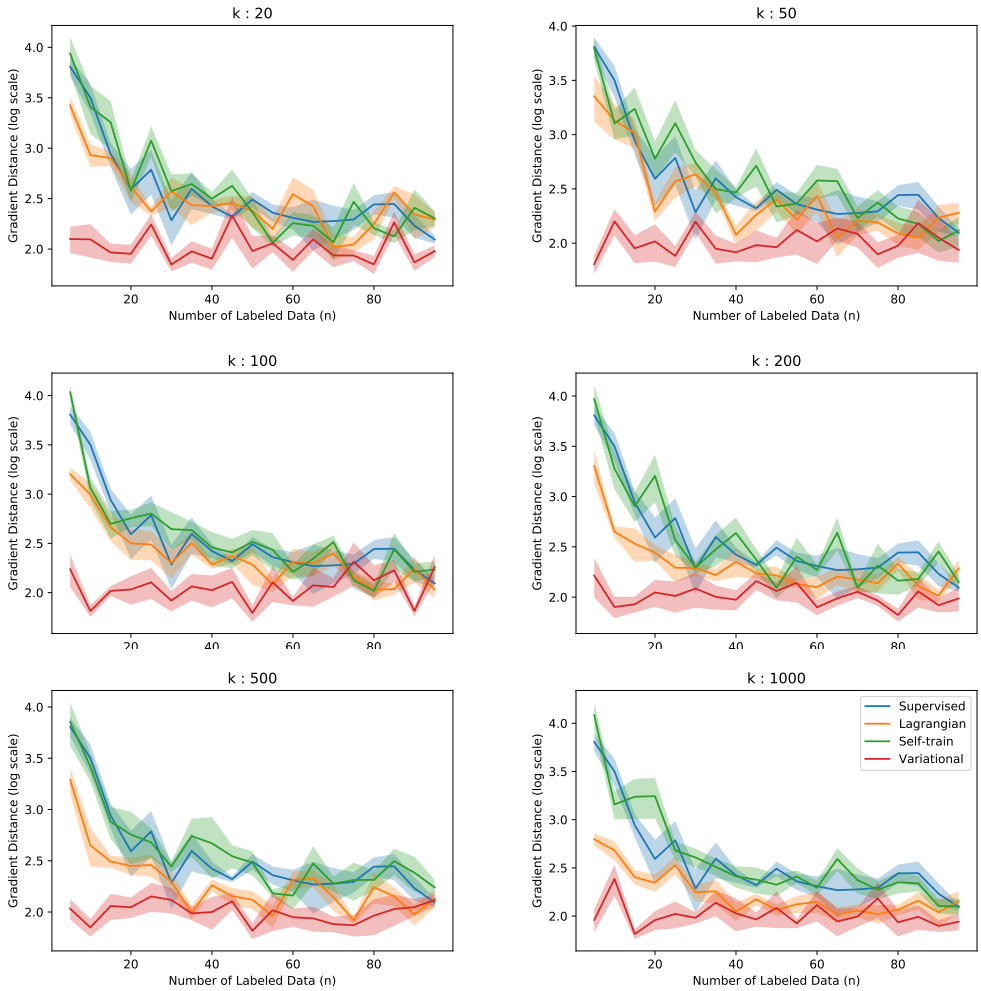

Figure 15: Comparison of Input Gradient Distance when regressing a 2 layer neural network over different values of k. $m = 1000, T = 10$. Results are averaged over 5 seeds.

verify the impacts on the overall performance. In this experimental setup, we are regressing a 2 layer neural network with a hidden dimension size of 100, which is much larger than in our other synthetic experiments. Here, we vary over simpler teacher models by changing their hidden dimension size.

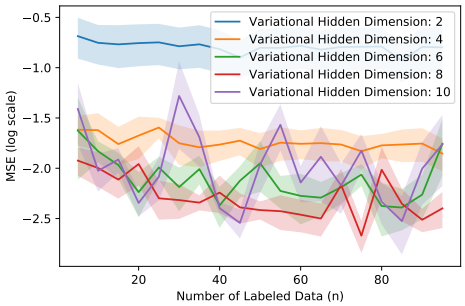

Figure 16: Comparison of MSE on regressing a 2 Layer Neural Network over simpler teacher models (hidden dimension). Here, $k = 20, m = 1000, T = 10$. Results are averaged over 5 seeds.

We observe no major differences as we shrink the hidden dimension size by a small amount. For significantly smaller hidden dimensions (e.g., 2 or 4), we observe a large drop in performance as these simpler teachers can no longer fit the approximate projection onto our class of EPAC models accurately. However, slightly smaller networks (e.g., 6, 8) can fit this projection as well, if not better

in some cases. This is a useful finding, meaning that our teacher can be a *smaller model* and get comparable results, showing that this simpler teacher can help with scalability without much or any drop in performance.

O.3 NUMBER OF UNLABELED DATA

As a main benefit of our approach is the ability to incorporate large amounts of unlabeled data, we provide a study as we vary the amount of unlabeled data m that is available. When varying the amount of unlabeled data, we observe that the performance of self-training and our variational objective improves at similar rates.

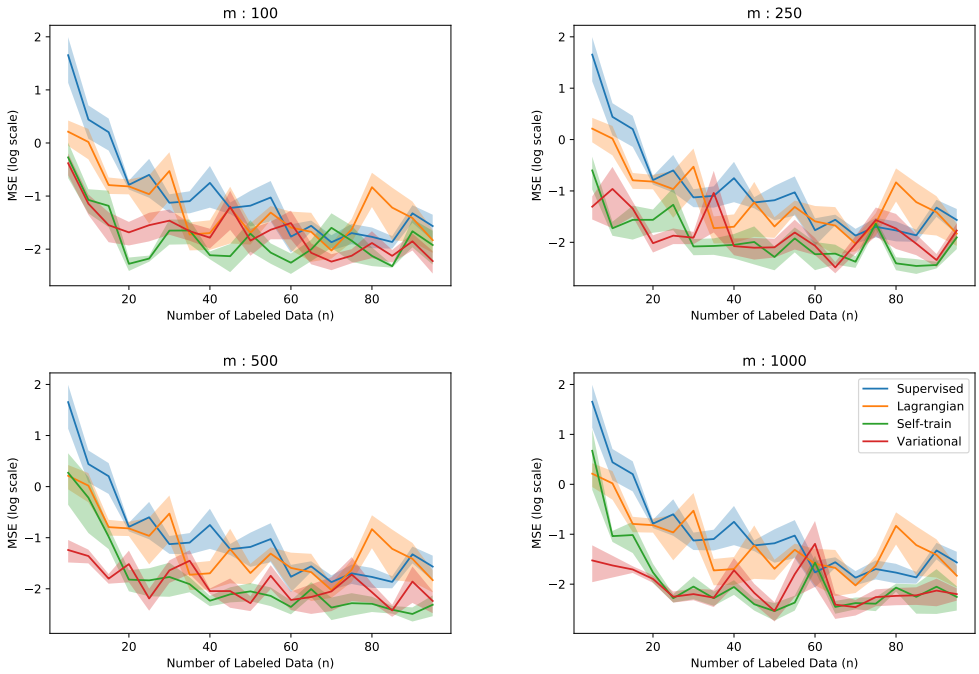

Figure 17: Comparison of MSE on regressing a 2 layer neural network over different values of m. $k = 20, T = 10$. Results are averaged over 5 seeds.

O.4 DATA DIMENSION

We also provide ablations as we vary the underlying data dimension d. As we increase the dimension d, we observe that the methods seem to achieve similar performance, due to the difficulty in modeling the high-dimensional data. Also, here gradient information is much harder to incorporate, as the input gradient itself is d-dimensional, so we do not see as much of a benefit of our approach as d grows.

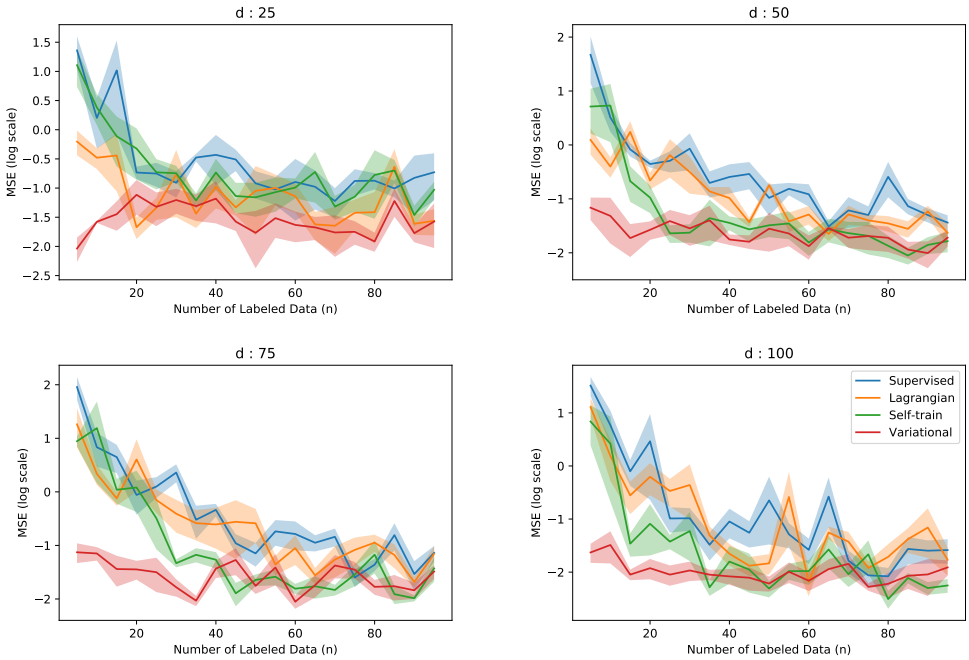

Figure 18: Comparison of MSE on regressing a 2 Layer Neural Network over different underlying data dimensions d. $m = 1000, k = 20$. For the iterative methods, $T = 10$. Results are averaged over 5 seeds.

O.5 HYPERPARAMETERS

First, we compare the different approaches over differnet values of regularization (λ) towards satisfying the explanation constraints. Here, we compare the augmented Lagrangian approach, the self-training approach, the iterative self-training and variational approaches.

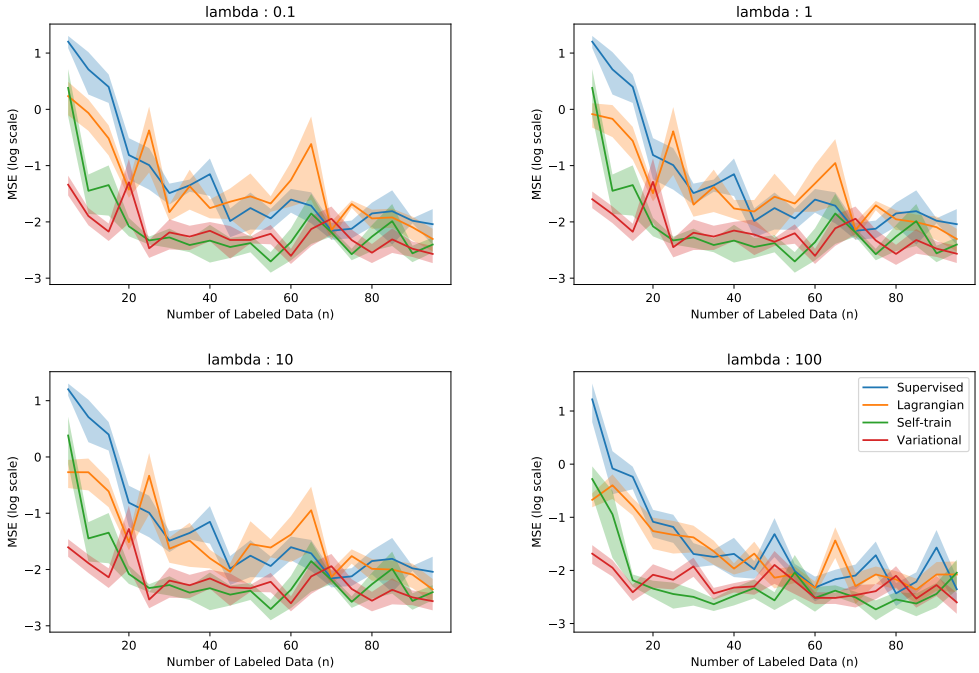

Figure 19: Comparison of MSE on regressing a 2 Layer Neural Network over different values of λ. $m = 1000, k = 20$. For the iterative methods, $T = 10$. Results are averaged over 5 seeds.

We observe that there is not a significant trend as we change the value of λ across the different methods. Since we know that our explanation is perfect (our restricted EPAC class contains the target classifier), increasing the value of λ should help, until this constraint is met.

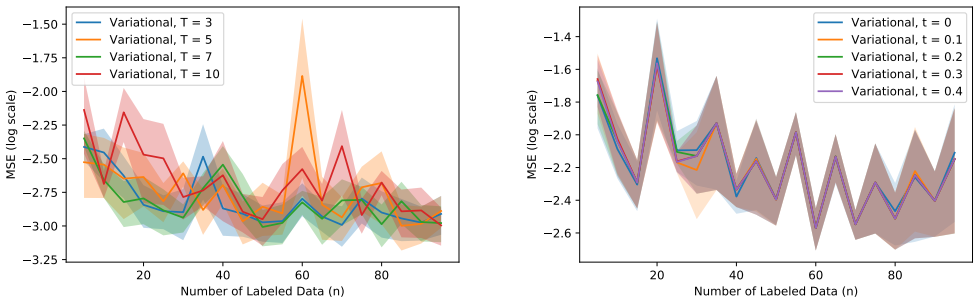

Figure 20: Comparison of MSE on regressing a 2 Layer Neural Network over different values of T (left) and t (right) in our variational approach. $m = 1000, k = 20, t = 10, T = 10$, unless noted otherwise. Results are averaged over 5 seeds.

Next, we compare different hyperparameter settings for our variational approach. Here, we analyze trends as we vary the values of T (number of iterations) and t (threshold before adding hinge penalty). We note that the value of t does not significantly impact the performance of our method while increasing values of T seems to generally benefit performance on this task.