Robust Generalization of Quadratic Neural Networks via Function Identification

Kan Xu [1] **Hamsa Bastani** [1] **Osbert Bastani** [1]

Abstract

A key challenge facing deep learning is that neural networks are often not robust to shifts in the underlying data distribution. We study this problem from the perspective of the statistical concept of parameter identification. Generalization bounds from learning theory often assume that the test distribution is close to the training distribution. In contrast, if we can identify the "true" parameters, then the model generalizes to arbitrary distribution shifts. However, neural networks are typically overparameterized, making parameter identification impossible. We show that for quadratic neural networks, we can identify the function represented by the model even though we cannot identify its parameters. Thus, we can obtain robust generalization bounds even in the overparameterized setting. We leverage this result to obtain new bounds for contextual bandits and transfer learning with quadratic neural networks. Overall, our results suggest that we can improve robustness of neural networks by designing models that can represent the true data generating process. In practice, the true data generating process is often very complex; thus, we study how our framework might connect to neural module networks, which are designed to break down complex tasks into compositions of simpler ones. We prove robust generalization bounds when individual neural modules are identifiable.

1. Introduction

Recent work has demonstrated that neural networks are not robust to shifts in the underlying data, including both distribution shifts (i.e., where the data comes from a new distribution independent of the neural network parameters) (Hendrycks & Dietterich, 2019; Taori et al., 2020) as well as adversarial shifts (i.e., where the shift can depend on the neural network parameters) (Szegedy et al., 2013). Accordingly, there has been a great deal of interest in better understanding why neural networks fail to be robust (Tsipras et al., 2018; Ilyas et al., 2019) and on improving robustness (Goodfellow et al., 2014; Raghunathan et al., 2018; Cohen et al., 2019).

From the perspective of learning theory, there is little reason to expect neural networks to be robust, since generalization bounds typically assume that the test examples are from the same distribution as the training examples. PAC-Bayesian generalization bounds allow for a limited amount of robustness, but only if the support of the target distribution q is contained in that of the source distribution p, since it requires that the KL divergence $D_{\mathrm{KL}}(q \parallel p)$ is finite. Yet, the distribution shifts considered (Hendrycks & Dietterich, 2019) often shift probability mass to inputs that are completely outside of the source distribution.

Instead, the reason we might expect neural networks to be robust to these shifts is that humans are robust to them; for instance, small pixel-level shifts considered in adversarial examples are typically unnoticeable to humans, yet these shifts can move the image completely off of the distribution of natural images. This fact indicates a gap in our theoretical understanding of neural networks. In particular, the key question is understanding settings under which we may expect neural networks to be robust to distribution shifts that are "large" (e.g., in terms of KL divergence).

We study a strategy for closing this gap based on the statistical concept of *identifiability* (Hsu et al., 2012). At a high level, this concept assumes that the ground truth model belongs to the model family; then, in the limit of infinite training data, the learning algorithm can exactly recover the parameters of the ground truth model. For instance, in linear regression, the data is generated according to the model

$$y = \langle \theta^*, x \rangle + \xi,$$

where ξ is σ-subgaussian noise. Then, under mild assumptions on the training data $Z = (X, Y)$, the ordinary least squares (OLS) estimator $\hat{\theta}(Z)$ recovers the true parameter—

[*]Equal contribution [1]University of Pennsylvania, Pennsylvania, USA. Correspondence to: Kan Xu <kanxu@sas.upenn.edu>, Hamsa Bastani <hamsab@wharton.upenn.edu>, Osbert Bastani <obastani@seas.upenn.edu>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.,

i.e., in the limit of infinite data, $\hat{\theta}(Z) = \theta^*$. With finite samples, OLS satisfies high-probability convergence rates of the form

$$\|\hat{\theta}(Z) - \theta^*\|_2 \leq \epsilon. \tag{1}$$

The connection to robustness is that if (1) holds, then for *any* input x such that $\|x\|_2 \leq x_{\max}$, we have

$$|\langle \hat{\theta}(Z), x\rangle - \langle \theta^*, x\rangle| \leq \|\hat{\theta}(Z) - \theta^*\|_2 \cdot \|x\|_2 \leq \epsilon \cdot x_{\max}. \tag{2}$$

Thus, for *any* distribution $q(x)$ with support on $B_2(0, x_{\max}) = \{x \in \mathcal{X} \mid \|x\|_2 \leq x_{\max}\}$, $\hat{\theta}(Z)$ obtains bounded error—i.e., we have $\mathbb{E}_{q(x)}[(\langle \hat{\theta}(Z), x\rangle - \langle \theta^*, x\rangle)^2] \leq \epsilon^2 x_{\max}^2$ with high probability.

Thus, a natural question is whether we can obtain similar kinds of parameter identification bounds for neural networks. A key complication is that neural networks are typically *overparameterized*, possibly to facilitate optimization (Du & Lee, 2018; Jacot et al., 2018). In this setting, identification is no longer possible, since multiple parameters can yield the same model; at best, identification is possible if a finite set of parameters yield the same model, but neural network parameters may have continuous symmetries. Nevertheless, it may be possible to obtain bounds of the form (2)—in particular, even if we do not recover the true parameters θ^*, we can in principle still recover the function $f_{\theta^*}(x) = \langle \theta^*, x\rangle$. We refer to this notion as *function identification*. Furthermore, we show that quadratic neural networks satisfy function identification bounds under mild conditions.

To demonstrate the utility of this result, we show how function identification can be straightforwardly leveraged to obtain regret guarantees for a linear bandit (Rusmevichientong & Tsitsiklis, 2010) where each arm is a quadratic neural network. Linear bandits fundamentally involve covariate shift since their "covariates" are arms, which are adaptively chosen through the learning process as a function of past observations; thus, existing approaches have all operated in the setting where there is a *unique* and identifiable global minimizer. Similarly, we build on recent work proving bounds on transfer learning in the setting of bounded label shift and unbounded covariate shift (Bastani, 2020; Xu et al., 2021); again, we show that we can leverage function identification to easily transfer learn quadratic neural networks.

We also study how our results can connect to neural module networks (Andreas et al., 2016), which are designed to break down complex tasks into smaller ones that can each be solved by a neural network. We study a simplified form of neural module networks, where modules are quadratic neural networks composed in sequence according to a given input; for simplicity, we assume they can be trained in a supervised way, ensuring robust generalization. Then, we show that under certain conditions, compositions of these models are also robust, including the case where there are shifts in the distribution over compositions.

In summary, our contributions include formalizations of function identification and robust generalization (Section 2), a proof that quadratic neural networks satisfy function identification (Section 3), an application of this result to bandits (Section 4) and transfer learning (Section 5), and an analysis of robust generalization for neural module networks (Appendix B). Finally, we provide an expanded discussion of related work in Appendix A.

2. Problem Formulation

We consider a model $f_\theta : \mathcal{X} \to \mathcal{Y}$, with covariates $\mathcal{X} \subseteq \mathbb{R}^d$, labels $\mathcal{Y} \subseteq \mathbb{R}$, and parameters $\theta \in \Theta \subseteq \mathbb{R}^m$. A typical generalization bound from learning theory has form

$$\begin{aligned} &\mathbb{P}_{p(Z)}[L_p(\hat{\theta}(Z)) \leq \epsilon] \geq 1 - \delta \\ &\text{where } L_p(\theta) = \mathbb{E}_{p(x)}[(f_\theta(x) - f_{\theta^*}(x))^2], \end{aligned} \tag{3}$$

where $\epsilon, \delta \in \mathbb{R}_{>0}$, $Z = \{(x_1, y_1), ..., (x_n, y_n)\} \subseteq \mathcal{X} \times \mathcal{Y}$ with $y_i = f_{\theta^*}(x_i) + \xi_i$ a training set of i.i.d. observations from a distribution p (i.e., $p(Z) = p(x_1, y_1) \cdot ... \cdot p(x_n, y_n)$), ξ_i is bounded random noise independent of x_i with $\|\xi\| \leq \xi_{\max}$, $\theta^* \in \Theta$ are the true parameters, and

$$\begin{aligned} &\hat{\theta}(Z) = \arg\min_{\theta \in \Theta} \hat{L}(\theta; Z) \\ &\text{where } \hat{L}(\theta; Z) = \frac{1}{n}\sum_{i=1}^{n}(f_\theta(x_i) - y_i)^2 \end{aligned} \tag{4}$$

is an estimator based on the training data Z.[1] In particular, they assume that the training inputs $x_i \sim p$ are i.i.d. samples from the same distribution as the test example $x \sim p$.

Definition 2.1. The model f_θ and distribution p satisfy *function identification* if for any $\epsilon, \delta \in \mathbb{R}_{>0}$, we have $\mathbb{P}_{p(Z)}[\forall x \in \mathcal{X} \,.\, (f_{\hat{\theta}(Z)}(x) - f_{\theta^*}(x))^2 \leq \epsilon] \geq 1 - \delta$ for $n = |Z|$ sufficiently large.

Function identification implies generalization bounds even when the test data comes from a different distribution q. In particular, we say f_θ *robustly generalizes* if for any q with support on $\mathcal{X}$, we have

$$\mathbb{P}_{p(Z)}[L_q(\hat{\theta}(Z)) \leq \epsilon] \geq 1 - \delta, \tag{5}$$

where the difference from (3) has been highlighted in red. It is easy to see that function identification implies (5).

[1] In (3), the loss L_p omits the label errors ξ; including it would result in an additive constant to L_p. This choice ensures that the optimal parameters have zero loss—i.e., $L_q(\theta^*) = 0$ for any q.

3. Function Identification of Quadratic Neural Networks

Traditional statistical bounds on parameter identification can provide guarantees for arbitrary covariate shift. In particular, suppose we have a bound of the form

$$\mathbb{P}_{p(Z)}\left[\|\hat{\theta}(Z) - \theta^*\|_2 \leq \epsilon\right] \geq 1 - \delta, \tag{6}$$

and assume that the model family f_θ is K-Lipschitz continuous in θ; then, we have

$$L_q(\hat{\theta}(Z)) \leq K^2 \cdot \|\hat{\theta}(Z) - \theta^*\|_2^2 \leq K^2\epsilon^2 \tag{7}$$

with probability at least $1 - \delta$ according to $p(Z)$. In particular, this bound holds for any covariate distribution q. Our goal is to extend these techniques to quadratic neural networks, which are overparameterized so we cannot identify the true parameters θ^*—i.e., (6) does not hold.

3.1. Quadratic Neural Networks

We consider a neural network f_θ, where $\theta \in \mathbb{R}^{d\times k}$, with a single hidden layer with k neurons—i.e., $f_\theta(x) = \sum_{j=1}^k a_j \cdot \sigma(\langle \theta_j, x\rangle)$. We consider the over-parameterization case where k can be much larger than d. Following prior work (Du & Lee, 2018), we assume that f_θ has quadratic activations and output weights equal to one—i.e., $\sigma(z) = z^2$ and $a_j = 1$ for each $j \in [k]$, so we have

$$f_\theta(x) = \sum_{j=1}^k \langle \theta_j, x\rangle^2.$$

We assume the true (training) loss is the mean-squared error $L_p(\theta) = \mathbb{E}_{p(x)}[(f_\theta(x) - f_{\theta^*}(x))^2]$, and we consider a model trained using an empirical estimate of this loss on the training dataset as in Problem (4). Now, our goal is to obtain a bound of the form (7); to this end, we assume the following.

Assumption 3.1. $\|x\|_2 \leq x_{\max}$ and $\|\theta\|_F \leq \theta_{\max}$.

Assumption 3.2. There exists $\alpha \in \mathbb{R}_{>0}$ such that $\mathbb{E}_{p(x)}[(x^\top \Delta x)^2] \geq \alpha \|\Delta\|_F^2$ for any symmetric Δ.

A discussion of Assumption 3.2 is given in Appendix C.3.

3.2. Robust Generalization

Our approach leverages the fact that $f_\theta(x) = x^\top(\theta\theta^\top)x$; thus, f_θ resembles a matrix factorization model. Recent work has leveraged this connection to translate matrix factorization theory to quadratic neural networks (Du & Lee, 2018). We let $g(\theta) = \theta\theta^\top$ and $\tilde{f}_\phi(x) = x^\top \phi x$, where $\phi \in \Phi \subseteq \mathbb{R}^{d\times d}$, in which case $f_\theta(x) = \tilde{f}_{g(\theta)}(x)$; in addition, we define $\tilde{L}_p(\phi) = \mathbb{E}_{p(x)}[(\tilde{f}_\phi(x) - \tilde{f}_{\phi^*}(x))^2]$, where $\phi^* = g(\theta^*)$, and $\hat{\tilde{L}}(\phi; Z) = n^{-1}\sum_{i=1}^n (\tilde{f}_\phi(x_i) - y_i)^2$, so $L_p(\theta) = \tilde{L}_p(g(\theta))$ and $\hat{L}(\theta; Z) = \hat{\tilde{L}}(g(\theta); Z)$. Finally, we assume that $\|\phi\|_F \leq \phi_{\max}$; in general, we have $\phi_{\max} \leq \theta_{\max}^2$ by Assumption 3.1.

We state our main result, which says that quadratic neural networks can be functionally identified.

Theorem 3.3. *For any distribution $q(x)$ with support on $B_2(0, x_{max})$, we have*

$$\mathbb{P}_{p(Z)}\left[\forall x \in \mathcal{X} \,.\, (f_{\hat{\theta}(Z)}(x) - f_{\theta^*}(x))^2 \leq \frac{2K^2\epsilon}{\alpha}\right] \geq 1 - \delta$$

The proof is given in Appendix C.1. We first show that our estimate of the loss function is a uniformly good approximation of the true loss. Then, by the strong convexity of the loss $\tilde{L}_p(\phi)$ in ϕ, the estimation error of $g(\theta^*)$ can be upper bounded by that of the loss function. Combined with the Lipschitz continuity of the quadratic neural network, we can finally bound the generalization error by the estimation error of $g(\theta^*)$, regardless of the covariate distribution.

Finally, we also prove that gradient descent can find the global minima of $\hat{L}(\theta; Z)$, which ensures that gradient descent can perform function identification; we give a proof in Appendix C.2.

Theorem 3.4. *All local minima of $\hat{L}(\theta; Z)$ are also global minima.*

4. Quadratic Neural Bandits

A key application of robust generalization bounds is to parametric bandits; this is because, in bandit learning, the distribution of inputs x used to estimate $\hat{\theta} \approx \theta^*$ can differ from the distribution under which $f_{\hat{\theta}}$ is used. Thus, generalization bounds based on notions such as Rademacher complexity cannot be used. Unlike prior literature in bandits, we consider an overparameterized function that does *not* admit a unique solution; in contrast, recent work on neural tangent kernel bandits (Zhou et al., 2020) assumes that there is a unique, identifiable solution.

We consider a standard linear bandit (Rusmevichientong & Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011) with a fixed horizon $T \in \mathbb{N}$, but where the expected reward is parameterized by a quadratic neural network instead of a linear function. At each time step t, the algorithm chooses among a continuum of actions $x_t \in \mathcal{X}$, and receives a reward

$$y_t = f_{\theta^*}(x_t) + \xi_t = \sum_{j=1}^k \langle \theta_j^*, x_t\rangle^2 + \xi_t, \tag{8}$$

where $\theta^* \in \mathbb{R}^{d\times k}$ is an unknown parameter matrix, and ξ_t are bounded i.i.d. random variables. For simplicity, we

assume that $\mathcal{X} = B_2(0,1)$ is the unit ball. Then, our goal is to bound the *regret*

$$R(T) = \sum_{t=1}^{T} (\mathbb{E}_{p(\xi_t)}[y_t] - y^*), \text{ where } y^* = \max_{x \in \mathcal{X}} f_{\theta^*}(x).$$

Next, we consider an explore-then-commit strategy for simplicity (Summarized in Algorithm 1), since it already achieves the asymptotically optimal regret rate (Rusmevichientong & Tsitsiklis, 2010); our approach can similarly be applied to more sophisticated algorithms such as UCB (Abbasi-Yadkori et al., 2011) and Thompson sampling (Agrawal & Goyal, 2013). Our algorithm proceeds in two stages: (i) the *exploration stage* (for $t \in \{1, ..., m\}$), and (ii) the *exploitation stage* (for $t \in \{m+1, ..., T\}$), for some m to be optimized. In the exploration stage, we randomly choose actions $x_t \sim p$, where

$$p(x) = \prod_{i=1}^{d} \text{Uniform}\left(x_i; \left[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\right]\right). \tag{9}$$

Next, we compute an estimate $\hat{\theta}$ of θ^* based on the data Z collected so far, and compute the optimal action $\hat{x}$ assuming $\hat{\theta}$ are the true parameters. Then, in the exploitation stage, we always use action $\hat{x}$. Note that traditional bounds do not provide any guarantees about the accuracy of $f_{\hat{\theta}}(\hat{x})$ compared to $f_{\theta^*}(\hat{x})$, since $\hat{x}$ is not sampled from the distribution p. In contrast, Theorem 3.3 provides a uniform guarantee, and can therefore be used to bound the regret.

Theorem 4.1. *The expected regret of Algorithm 1 is*

$$R(T) \leq C_0 + C_1 \cdot T^{2/3} \left(\log\left(3 + \frac{8\phi_{max} K T}{\ell_{max}^2}\right)\right)^{1/3},$$

where C_0 and C_1 do not depend on T (see Appendix D).

In particular, we have $R(T) = \tilde{O}(T^{2/3})$; we give a proof in Appendix D.

5. Transfer Learning of Quadratic Neural Networks

So far, we have considered shifts in the covariate distribution but not in the label distribution. Now, we consider a transfer learning problem where there is additionally a small shift in the labels. In particular, we assume we have *proxy data* $Z_p \subseteq \mathcal{X} \times \mathcal{Y}$ from the source domain of the form $y_{p,i} = f_{\theta_p^*}(x_{p,i}) + \xi_{p,i}$ (for $i \in [n_p]$), where $\theta_p^* \in \Theta$ are the proxy parameters and $p(x_p)$ is the source covariate distribution, along with *gold data* $Z_g \subseteq \mathcal{X} \times \mathcal{Y}$ from the target domain of the form $y_{g,i} = f_{\theta_g^*}(x_{g,i}) + \xi_{g,i}$ (for $i \in [n_g]$), where $\theta_g^* \in \Theta$ are the gold parameters and $q(x_g)$ is the target covariate distribution. We are interested in the setting $n_p \gg n_g$, and where $\|\theta_g^* - \theta_p^*\|_F \leq B$ is small.

We consider a two-stage estimator (Bastani, 2020) that first computes an estimate of the proxy parameters $\hat{\theta}_p = \arg\min_{\theta \in \Theta} \hat{L}(\theta; Z_p)$, and then computes an estimate of the gold parameters in a way that is constrained towards the proxy parameters. Note that by applying Theorem 3.3 directly to proxy data, we have for $\epsilon_p = \mathcal{O}(\sqrt{\frac{d^2}{n_p}})$ (See Appendix E) that

$$\mathbb{P}_{p(Z)}\left[L_q(\hat{\theta}_p) \leq \frac{2K^2\epsilon_p}{\alpha}\right] \geq 1 - \frac{\delta}{2}.$$

Then, letting $\hat{B} = B + \frac{1}{\sigma_0}\sqrt{\frac{2\epsilon_p}{\alpha}}$ (σ_0 defined in Assumption E.1) be an expanded radius to account for error in our estimate of $\hat{\theta}_p$, we use the estimator

$$\hat{\theta}_g = \underset{\theta \in B_2(\hat{\theta}_p, \hat{B})}{\arg\min} \ \hat{L}(\theta; Z_g)$$
$$\text{where } B_2(\hat{\theta}_p, \hat{B}) = \{\theta \in \Theta \mid \|\theta - \hat{\theta}_p\|_F \leq \hat{B}\}.$$

Then, we show that if B is small and n_p is large, $f_{\hat{\theta}_g}$ is accurate even if n_g is small; we give a proof in Appendix E.

Theorem 5.1. *For any distribution $q(x)$ with support on $B_2(0, x_{max})$, we have*

$$\mathbb{P}_{p(Z)}\left[L_q(\hat{\theta}_g) \leq \frac{2K^2\epsilon_g}{\alpha}\right] \geq 1 - \delta,$$

where $\epsilon_g = \mathcal{O}(\hat{B}\sqrt{\frac{d^2}{n_g}})$ (See Appendix E).

6. Conclusion

We have presented a number of results demonstrating that overparameterization does not fundamentally harm learning models that are robust to arbitrary distribution shifts. In particular, even though we can no longer identify the true parameters for quadratic neural networks, we show that we can identify the true function, thereby enabling us to prove new results in bandits and transfer learning. Finally, we provide preliminary analysis extending these results to neural module networks to handle complex data generating processes. A limitation of our work is that it is specialized to quadratic neural networks; a key direction for future work is exploring how these results extend to other architectures.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pp. 2312–2320, 2011.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135. PMLR, 2013.

Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 39–48, 2016.

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pp. 254–263. PMLR, 2018.

Bach, F., Mairal, J., and Ponce, J. Convex sparse matrix factorizations. *arXiv preprint arXiv:0812.1869*, 2008.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Bastani, H. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 2020.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pp. 137–144, 2007.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Wortman, J. Learning bounds for domain adaptation. In *Advances in neural information processing systems*, pp. 129–136, 2008.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

Du, S. and Lee, J. On the power of over-parametrization in neural networks with quadratic activation. In *International Conference on Machine Learning*, pp. 1329–1338. PMLR, 2018.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685. PMLR, 2019.

Duchi, J. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Foster, D. and Rakhlin, A. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 3199–3210. PMLR, 2020.

Gao, R., Cai, T., Li, H., Hsieh, C.-J., Wang, L., and Lee, J. D. Convergence of adversarial training in overparametrized neural networks. *Advances in Neural Information Processing Systems*, 32:13029–13040, 2019.

Ge, R., Lee, J. D., and Ma, T. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Hsu, D., Kakade, S. M., and Liang, P. Identifiability and unmixing of latent parse trees. *arXiv preprint arXiv:1206.3137*, 2012.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Kearns, M. J., Vazirani, U. V., and Vazirani, U. *An introduction to computational learning theory*. MIT press, 1994.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020.

Long, P. M. and Sedghi, H. Generalization bounds for deep convolutional neural networks. *arXiv preprint arXiv:1905.12600*, 2019.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. *arXiv preprint arXiv:1706.08947*, 2017.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*, 2020.

Ruis, L., Andreas, J., Baroni, M., Bouchacourt, D., and Lake, B. M. A benchmark for systematic generalization in grounded language understanding. *arXiv preprint arXiv:2003.05161*, 2020.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35 (2):395–411, 2010.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Steinhardt, J., Koh, P. W., and Liang, P. Certified defenses for data poisoning attacks. *arXiv preprint arXiv:1706.03691*, 2017.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *arXiv preprint arXiv:2007.00644*, 2020.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Wainwright, M. *High-dimensional statistics: A non-asymptotic viewpoint*. Book Draft (Working Publication), 2016. URL https://www.stat.berkeley.edu/~wainwrig/nachdiplom/Chap2_Sep10_2015.pdf.

Wen, J., Yu, C.-N., and Greiner, R. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning*, pp. 631–639. PMLR, 2014.

Xu, K., Zhao, X., Bastani, H., and Bastani, O. Group-sparse matrix factorization for transfer learning of word embeddings. In *International Conference on Machine Learning*, 2021.

Yu, Y., Wang, T., and Samworth, R. J. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Zhou, D., Li, L., and Gu, Q. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pp. 11492–11502. PMLR, 2020.

A. Related Work

Prior work has connected misspecification (i.e., the true model is in the model family) and robustness to covariate shift (Shimodaira, 2000; Wen et al., 2014); however, having a correctly specified model is insufficient if the true parameters are not identifiable—e.g., in linear regression, if the covariance matrix $\Sigma = \mathbb{E}_{p(x)}[xx^\top]$ is singular, then θ are not identifiable; in this case, the estimated model may not be robust. Quadratic neural networks cannot be identified even if the model is correctly specified since the parameters have a continuous symmetry (i.e., orthogonal transformations).

Recent work has studied learning under adversarial examples (Goodfellow et al., 2014; Raghunathan et al., 2018; Cohen et al., 2019) and corrupted training data (Steinhardt et al., 2017; Diakonikolas et al., 2019). In contrast, we are interested in robustness to covariate shift; there has been recent work empirically showing that neural networks are sensitive to distribution shift (Hendrycks & Dietterich, 2019; Taori et al., 2020; Ruis et al., 2020; Ribeiro et al., 2020; Koh et al., 2020). Distributionally robust optimization enables training of models robust to small shifts (Duchi & Namkoong, 2018), but we are interested in potentially large shifts. Unsupervised domain adaptation (Ben-David et al., 2007; Blitzer et al., 2008) learns a model on a covariate shifted target distribution; however, they rely on unlabeled examples from the target domain, whereas we do not. There has been recent theory on robustness to adversarial perturbations—e.g., showing there may be a tradeoff between robustness and on-distribution generalization (Tsipras et al., 2018), and that non-robust algorithms tend to learn predictive but brittle representations compared to adversarially robust ones (Ilyas et al., 2019). In contrast, we show that these tradeoffs are mitigated when the true model function can be identified despite overparameterization. Furthermore, adversarial shifts are typically bounded (e.g., small L_∞ norm), whereas the shifts we consider may be large.

There has been a great deal of recent work on deep learning theory, including on quadratic neural networks; however, it has largely focused on optimization (Ge et al., 2017; Jacot et al., 2018; Du et al., 2019; Gao et al., 2019), on-distribution generalization (Neyshabur et al., 2017; Du & Lee, 2018; Jacot et al., 2018; Arora et al., 2018; Long & Sedghi, 2019). In contrast, we are interested in out-of-distribution generalization.

Prior literature on parameterized bandits has considered a number of functional forms, ranging from linear (Abbasi-Yadkori et al., 2011; Rusmevichientong & Tsitsiklis, 2010) to neural tangent kernels (Zhou et al., 2020). Most of this work makes a *realizability* assumption that the model family contains the true model;[2] implicitly, they consider model families where there is a *unique*, identifiable true parameter. These assumptions are necessary precisely due to the fact that the test and training distributions are different; thus, much of the bandit literature has focused on proving parameter identification results to enable learning. In contrast, the identifiability assumption does not hold for quadratic neural networks because they are invariant to parameter transformations. To the best of our knowledge, we consider the first overparameterized bandit that considers a model that is not identifiable; we find that similar regret results hold as long as the function represented by the model can be identified. Separately, (Foster & Rakhlin, 2020) makes a general connection between online regression oracles and the regret of a bandit algorithm; however, their approach only provides good guarantees when the regression oracle returns a model that generalizes off-distribution. Finally, recent work on UCB with neural tangent kernels (Zhou et al., 2020) provides general regret bounds, but their bound is only sublinear under conditions such as the true reward function having small RKHS norm (see Remark 4.8 in their paper), which amounts to assuming they can recover the true parameters.

B. Generalization Bounds for Neural Module Networks

While function identification enables robust generalization, many data generating processes are too complex to be identifiable. Neural module networks are designed to break complex prediction problems into smaller tasks that are individually easier to solve. These models take two kinds of input: (i) a sequence of tokens w (e.g., word embeddings) indicating the correct composition of modules, and (ii) the input x to the modules. Then, the model predicts the sequence of modules $j_1 ... j_T$ based on w, and runs the modules in sequence to obtain output $x' = f_{j_T}(...(f_{j_1}(x))...)$.

We study conditions under which neural module networks can robustly generalize. Rather than study arbitrary distribution shifts, we consider two separate shifts:

- **Module inputs:** We assume that the individual modules are identifiable; as a consequence, we assume the shift to the module input x can be arbitrary.
- **Module composition:** We consider shifts to the token sequence w. If the model mapping w to $j_1 ... j_T$ is identifiable, then the entire model is identifiable. Instead, we show that when this model is not identifiable, compositional structure can still

[2] Slightly different from realizability in PAC learning (Kearns et al., 1994), which says there is a model with zero true loss.

aid generalization. Intuitively, we show that while small shifts in the compositional structure can cause large shifts in the distribution $p(w)$, models that leverage the structure of $p(w)$ can still generalize well.

In more detail, consider a simplified neural module network f, which includes (i) a set of neural modules $\{f_j : \mathcal{X} \to \mathcal{X}\}_{j=1}^k$, and (ii) a parser $g : \mathcal{Z}^T \to [k]^T$, where $\mathcal{Z} \subseteq \mathbb{R}^r$. We assume each component of $f_j(x)$ is computed by a separate quadratic neural network; we discuss the architecture of g below. Then, given an input $x \in \mathcal{X} \subseteq \mathbb{R}^d$ and $w \in \mathcal{W} = \mathcal{Z}^T$, the corresponding neural module network $f : \mathcal{X} \times \mathcal{W} \to \mathcal{X}$ is defined by

$$f(x, w) = (f_{j_T} \circ ... \circ f_{j_1})(x) = f_{j_T}(...(f_{j_1}(x))...) \qquad \text{where} \qquad j_1...j_T = g(w).$$

We assume that g has compositional structure—i.e., for some $\tilde{g} : [k] \times \mathcal{Z} \to [k]$, we have

$$g(w) = j_1...j_T \qquad \text{where} \qquad j_t = \begin{cases} 0 & \text{if } t = 0 \\ \tilde{g}(z_t, j_{t-1}) & \text{otherwise,} \end{cases}$$

where $w = z_1...z_T$. Intuitively, w is a sequence of word vectors; then, the current neural module $j_t = \tilde{g}(z_t, j_{t-1})$ depends both on the current word vector z_t and the previous neural module j_{t-1}. First, we assume that the individual modules have been functionally identified.

We assume we have fully labeled data we can use to train the neural modules—i.e., for each input x and sequence w, we have both the desired sequence $j_1...j_T$ of neural modules, as well as the entire execution $x_0, x_1, ..., x_T$, where $x_0 = x$ and $x_{t+1} = f_{j_t}(x_t)$ otherwise. Thus, we can use supervised learning to train the neural modules;[3] in particular, we can construct labeled examples (j_{t-1}, z_t, j_t) used to train the parser $\tilde{g}$, and labeled examples (x_t, x_{t+1}) to train the modules f_{j_t}. For simplicity, we assume we have a uniform lower bound n on the number of training examples for the parser and for each module. Then, we have the following straightforward result:

Lemma B.1. *With probability at least $1 - dk\delta$, for each $j \in [k]$, we have*

$$\|\hat{f}_j(x) - f_j^*(x)\|_2 \leq \sqrt{\frac{2dK^2\epsilon}{\alpha}} =: \epsilon_f \qquad (\forall x \in \mathcal{X}),$$

where $\hat{f}_j$ is the estimated module and f_j^* is the ground truth module.

This result follows straightforwardly from Theorem 3.3 along with a union bound. In contrast, we do not assume the parsing model robustly generalizes, but only on distribution.

Lemma B.2. *With probability at least $1 - \delta$, we have*

$$\mathbb{P}_{\tilde{p}(z,j)}\left[\hat{\tilde{g}}(z, j) \neq \tilde{g}^*(z, j)\right] \leq 4R_n(\mathcal{G}) + \sqrt{\frac{2\log(2/\delta)}{n}} =: \epsilon_g,$$

where $R_n(\mathcal{G})$ is the Rademacher complexity of $\mathcal{G}$ (including its loss function), where $p(z, j) = T^{-1}\sum_{t=1}^T p_t(z, j)$, and where

$$\tilde{p}_t(z, j) = \begin{cases} \mathbb{1}(j = 0) \cdot \tilde{p}(z) & \textit{if } t = 1 \\ \sum_{j'=1}^k \int \mathbb{1}(j = \tilde{g}^*(z', j')) \cdot \tilde{p}(z \mid z') \cdot \tilde{p}_{t-1}(z', j')dz' & \textit{otherwise.} \end{cases}$$

This result is a standard Rademacher generalization bound (Bartlett & Mendelson, 2002). Note that we have also assumed that the distribution over token sequences is structured, which is necessary for our compositional implementation of g to generalize, even on distribution. Intuitively, the distribution over (z, j) consists of both a unigram model over the word vectors:

$$p(z_1, ..., z_T) = \prod_{t=1}^T \tilde{p}(z_t \mid z_{t-1}),$$

[3] Neural modules are often trained with only partial supervision (Andreas et al., 2016); we leave an analysis of this learning strategy to future work since our focus is on understanding generalization rather than learning dynamics.

where we define $\tilde{p}(z_1 \mid z_0) = \tilde{p}(z_1)$, as well as a unigram model over neural modules:

$$p(j_1 ... j_T \mid z_1, ..., z_T) = \prod_{t=1}^{T} \mathbb{1}(j_t = \tilde{g}^*(z_t, j_{t-1})).$$

Next, we consider a shifted distribution $\tilde{q}(z \mid z')$, which is close to $\tilde{p}(z \mid z')$.

Assumption B.3. We have $\|\tilde{q}(\cdot \mid z') - \tilde{p}(\cdot \mid z')\|_{\text{TV}} \leq \alpha$.

Importantly, despite this assumption, the shift between the overall distributions $p(z_1, ..., z_T)$ and $q(z_1, ..., z_T)$ can still be large since it compounds exponentially across the steps $t \in [T]$.

Proposition B.4. *There exist p and q that satisfy Assumption B.3, but $\|p - q\|_{TV} = 2(1 - (1 - \alpha/2)^T)$.*

That is, even if the single step probabilities $\tilde{p}(z \mid z')$ and $\tilde{q}(z \mid z')$ have total variation (TV) distance bounded as in Assumption B.3, the overall distributions p and q can have TV distance exponentially close to the maximum possible distance of 2 in T; we give a proof in Appendix F.1.

We show that neural module networks generalize since $\hat{g}$ leverages the compositional structure of p. First, we show that under Assumption B.3, the overall shift in the input distribution of $\hat{\tilde{g}}$ is bounded:

Lemma B.5. *We have $\|\tilde{q} - \tilde{p}\|_{TV} \leq T\alpha$.*

That is, while the shift can compound across steps t, it does so only linearly; we give a proof in Appendix F.2. Next, we show that as a consequence, the error of $\hat{g}$ is bounded.

Lemma B.6. *Assuming that $\mathbb{P}_{p(z,j)}[\hat{\tilde{g}}(z, j) \neq \tilde{g}^*(z, j)] \leq \epsilon_g$, then $\mathbb{P}_{p(w)}[\hat{g}(w) \neq g^*(w)] \leq T\epsilon_g$.*

We give a proof in Appendix F.3. Finally, we have our main result.

Theorem B.7. *With probability at least $1 - (dk + 1)\delta$, we have*

$$\mathbb{P}_{q(w)}\left[\|\hat{f}(x, w) - f^*(x, w)\|_2 \leq T\epsilon_f \cdot \max\{K^{T-1}, 1\}\right] \geq 1 - T\epsilon_g - T^2\alpha.$$

We give a proof in Appendix F.4. Intuitively, Theorem B.7 says that the error of the neural module network is linear in T as long as $K \leq 1$. Note that even if there is no distribution shift, its error is

$$\mathbb{P}_{p(w)}\left[\|\hat{f}(x, w) - f^*(x, w)\|_2 \leq T\epsilon_f \cdot \max\{K^{T-1}, 1\}\right] \geq 1 - T\epsilon_g,$$

by the same argument as the proof of Theorem B.7. The exponential dependence on K is unavoidable since $K > 1$ says that the modules f_j can expand the input, which leads to exponential blowup in the magnitude of the output as a function of T, which also makes the estimation error exponential in T. Thus, the only cost to the distribution shift from p to q is the additional error probability $T^2\alpha$.

C. Proofs for Section 3

C.1. Proof of Theorem 3.3

We begin by stating several lemmas needed to prove Theorem 3.3.

Lemma C.1. *The loss $\tilde{L}_p(\phi)$ is 2α-strongly convex in ϕ.*

We give a proof in Appendix C.4.

Lemma C.2. *The model $\tilde{f}_\phi$, loss $\tilde{L}$ is K-Lipschitz in ϕ, where $K = 4\phi_{max}x_{max}^4$.*

We give a proof in Appendix C.5.

Lemma C.3. *For any $\delta \in \mathbb{R}_{>0}$, we have*

$$\mathbb{P}_{p(Z)}\left[\sup_{\theta \in \Theta} |L_p(\theta) - \hat{L}(\theta; Z) - \sigma(Z)| \leq \epsilon\right] \geq 1 - \delta,$$

where $\sigma(Z) = n^{-1}\sum_{i=1}^n \xi_i^2$*, and letting* $\ell_{max} = 2x_{max}^2\phi_{max}$ *be an upper bound on* $|f_\theta(x) - f_{\theta^*}(x)|$*,*

$$\epsilon = \sqrt{\frac{18\ell_{max}^2(\ell_{max}^2 + \xi_{max}^2)}{n}\left(d^2 \max\left\{1, \log\left(1 + \frac{8\phi_{max}Kn}{\ell_{max}^2}\right)\right\} + \log\frac{2}{\delta}\right)}. \tag{10}$$

We give a proof in Appendix C.6. Note that $\epsilon \to 0$ as $n \to \infty$.

Now, we prove Theorem 3.3. First, by Lemma C.1, and since $\nabla_\phi \tilde{L}(g(\theta^*)) = 0$, we have

$$L_p(\hat{\theta}(Z)) - L_p(\theta^*) = \tilde{L}_p(g(\hat{\theta}(Z))) - \tilde{L}_p(g(\theta^*)) \geq \alpha\|g(\hat{\theta}(Z)) - g(\theta^*)\|_F^2. \tag{11}$$

Next, by Lemma C.3 and the fact that $\hat{\theta}$ minimizes $\hat{L}(\theta; Z)$, we have

$$L_p(\hat{\theta}(Z)) \leq \hat{L}(\hat{\theta}; Z) + \epsilon - \sigma(Z) \leq \hat{L}(\theta^*; Z) + \epsilon - \sigma(Z) \leq L_p(\theta^*) + 2\epsilon \tag{12}$$

with probability at least $1 - \delta$. Combining (11) and (12), we have

$$\|g(\hat{\theta}(Z)) - g(\theta^*)\|_F \leq \sqrt{\frac{2\epsilon}{\alpha}}$$

with probability at least $1 - \delta$. Finally, by Lemma C.2, we have

$$(f_{\hat{\theta}(Z)}(x) - f_{\theta^*}(x))^2 = (\tilde{f}_{g(\hat{\theta}(Z))}(x) - \tilde{f}_{g(\theta^*)}(x))^2 \leq K^2\|g(\hat{\theta}(Z)) - g(\theta^*)\|_2^2 \leq \frac{2K^2\epsilon}{\alpha}\ (\forall x \in \mathcal{X})$$

with probability at least $1 - \delta$, as claimed. □

C.2. Proof of Theorem 3.4

$\hat{\tilde{L}}(\phi; Z)$ is twice differentiable and convex in ϕ. Note that the minimization problem of $\hat{L}(\theta; Z)$ is equivalent to that of $\hat{\tilde{L}}(g(\hat{\theta}); Z)$. We consider two cases. First, consider the case where $\hat{\theta}$ has rank d. The first order condition $\nabla\hat{L}(\theta; Z) = 0$ is the same as $\nabla\hat{\tilde{L}}(g(\hat{\theta}); Z) = 0$, which gives

$$\nabla\hat{\tilde{L}}(\hat{\phi}; Z)\hat{\theta} = 0. \tag{13}$$

As $\hat{\theta}$ is of full row rank, there exists a matrix $\hat{\theta}^\dagger \in \mathbb{R}^{k\times d}$ such that $\hat{\theta}\hat{\theta}^\dagger = I$ (e.g. $\hat{\theta}^\dagger = \hat{\theta}^\top(\hat{\theta}\hat{\theta}^\top)^{-1}$). We can right multiply the above equation by $\hat{\theta}^\dagger$ and obtain that

$$\nabla\hat{\tilde{L}}(\hat{\phi}; Z) = 0.$$

As $\hat{\tilde{L}}(\phi; Z)$ is convex in ϕ, the above implies $\hat{\phi} = g(\hat{\theta})$ is a global minimum of $\hat{\tilde{L}}(\phi; Z)$. Therefore, $\hat{\theta}$ is a global minimum of $\hat{L}(\theta; Z)$. Next, consider the case where the rank of $\hat{\theta}$ is smaller than d. In this case, we follow the proof strategy in Proposition 4 in (Bach et al., 2008); we provide here for completeness. In this case, Equation (13) still holds, which implies

$$0 = \nabla\hat{\tilde{L}}(\hat{\phi}; Z)\hat{\theta}\hat{\theta}^\top = \nabla\hat{\tilde{L}}(\hat{\phi}; Z)\hat{\phi}. \tag{14}$$

The Hessian of $\hat{\tilde{L}}(g(\hat{\theta}); Z)$ has

$$\nabla^2\hat{\tilde{L}}(g(\hat{\theta}); Z)(\Delta, \Delta) = 2\langle\nabla\hat{\tilde{L}}(\hat{\phi}; Z), \Delta\Delta^\top\rangle + \nabla^2\hat{\tilde{L}}(\hat{\phi}; Z)(\hat{\theta}\Delta^\top + \Delta\hat{\theta}^\top, \hat{\theta}\Delta^\top + \Delta\hat{\theta}^\top).$$

As $\hat{\theta}R$ is also a local minimum for any orthogonal matrix R (i.e., $RR^\top = R^\top R = I$), we can find a $\hat{\theta}$ with the last column being 0 by right multiplying certain R. Then, consider any Δ with the first $k-1$ columns being 0 and the last column being any $u \in \mathbb{R}^d$. With this choice of Δ and $\hat{\theta}$, $\hat{\theta}\Delta^\top = 0$. Therefore,

$$\nabla^2\hat{\tilde{L}}(g(\hat{\theta}); Z)(\Delta, \Delta) = 2u^\top\nabla\hat{\tilde{L}}(\hat{\phi}; Z)u.$$

Since $\hat{\theta}$ is a local minimum of $\hat{\tilde{L}}(g(\hat{\theta}); Z)$, it holds that $\nabla^2\hat{\tilde{L}}(g(\hat{\theta}); Z)(\Delta, \Delta) \geq 0$, which implies

$$\nabla\hat{\tilde{L}}(\hat{\phi}; Z) \succeq 0. \tag{15}$$

Equation (14) and (15) together comprise the first order conditions of the convex minimization problem $\min_{\phi\succeq 0}\hat{\tilde{L}}(\phi; Z)$. Thus, $\hat{\theta}$ is also a global minimum. □

C.3. Proof of Minimum Eigenvalue for Uniform Distribution

Assumption 3.2 is standard; in particular, it is closely related to the assumption in linear regression that the minimum eigenvalue of the covariance matrix is lower bounded—i.e., $\Sigma = \mathbb{E}_{p(x)}[xx^\top] \succ 0$. As an example, when x is i.i.d. uniform in each component, e.g., $p(x) = \prod_{i=1}^d \text{Uniform}(x_i; [-1/2, 1/2])$, then we can take $\alpha = 1/180$.

Note that

$$\begin{aligned}
\mathbb{E}_{p(x)}[(x^\top \Delta x)^2]) &= \mathbb{E}_{p(x)}\left[\left(\sum_{i,j=1}^d x_i x_j \Delta_{ij}\right)^2\right] \\
&= \mathbb{E}_{p(x)}\left[\sum_i x_i^4 \Delta_{ii}^2 + \sum_{i\neq j} x_i^2 x_j^2 \Delta_{ii}\Delta_{jj} + 2\sum_{i\neq j} x_i^2 x_j^2 \Delta_{ij}^2\right] \\
&= \frac{1}{80}\sum_i \Delta_{ii}^2 + \frac{1}{144}\sum_{i\neq j}\Delta_{ii}\Delta_{jj} + \frac{1}{72}\sum_{i\neq j}\Delta_{ij}^2 \\
&= \left(\frac{1}{80} - \frac{1}{144}\right)\sum_i \Delta_{ii}^2 + \frac{1}{144}\left(\sum_i \Delta_{ii}\right)^2 + \frac{1}{72}\sum_{i\neq j}\Delta_{ij}^2 \\
&\geq \frac{1}{180}\|\Delta\|_F^2,
\end{aligned}$$

as claimed.

C.4. Proof of Lemma C.1

We use the notation $U : \nabla^2 f(\phi) : V$ to denote the matrix inner product $\langle U, \nabla^2 f(\phi)(V)\rangle$ for $U, V \in \mathbb{R}^{d\times d}$. The Hessian $\nabla^2 f(\phi)$ can be viewed as a $d^2 \times d^2$ matrix. As everything here is bounded, we can exchange the expectation and differentiation. Therefore, the Hessian of our loss function has for any symmetric matrix Δ

$$\Delta : \nabla^2 \tilde{L}_p(\phi) : \Delta = 2\mathbb{E}_{p(x)}[(x^\top \Delta x)^2] \geq 2\alpha\|\Delta\|_F^2,$$

where the last inequality uses Assumption 3.2. □

C.5. Proof of Lemma C.2

By our definition, for any $\phi, \phi' \in \Phi$,

$$|\tilde{f}_\phi(x) - \tilde{f}_{\phi'}(x)| = |(x^\top(\phi - \phi')x)^2| \leq x_{\text{max}}^2 \|\phi - \phi'\|_F.$$

Given our quadratic loss function, we have

$$\begin{aligned}
&|(\tilde{f}_\phi(x) - \tilde{f}_{\phi^*}(x))^2 - (\tilde{f}_{\phi'}(x) - \tilde{f}_{\phi^*}(x))^2| \\
&\leq |\tilde{f}_\phi(x) - \tilde{f}_{\phi^*}(x) + \tilde{f}_{\phi'}(x) - \tilde{f}_{\phi^*}(x)||\tilde{f}_\phi(x) - \tilde{f}_{\phi'}(x)| \\
&\leq 4\phi_{\text{max}} x_{\text{max}}^4 \|\phi - \phi'\|_F.
\end{aligned}$$

Next, the true loss satisfies

$$|\tilde{L}_p(\phi) - \tilde{L}_p(\phi')| \leq \mathbb{E}_{p(x)}[|(\tilde{f}_\phi(x) - \tilde{f}_{\phi^*}(x))^2 - (\tilde{f}_{\phi'}(x) - \tilde{f}_{\phi^*}(x))^2|] \leq 4\phi_{\text{max}} x_{\text{max}}^4 \|\phi - \phi'\|_F.$$

Finally, the empirical loss satisfies

$$
\begin{aligned}
|\hat{\tilde{L}}(\phi; Z) - \hat{\tilde{L}}(\phi'; Z)| &= \left| \frac{1}{n} \sum_{i=1}^{n} [(\tilde{f}_\phi(x_i) - y_i)^2 - (\tilde{f}_{\phi'}(x_i) - y_i)^2] \right| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} |(\tilde{f}_\phi(x_i) - \tilde{f}_{\phi^*}(x_i))^2 - (\tilde{f}_{\phi'}(x_i) - \tilde{f}_{\phi^*}(x_i))^2| \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} |\xi_i| \cdot |\tilde{f}_\phi(x_i) - \tilde{f}_{\phi'}(x_i)| \\
&\leq (4\phi_{\max} x_{\max}^4 + 2\xi_{\max} x_{\max}^2) \|\phi - \phi'\|_F,
\end{aligned}
$$

as claimed. □

C.6. Proof of Lemma C.3

First, we have the following results:

Lemma C.4 (Covering Number of Euclidean Ball)**.** *For a Euclidean ball in $\mathbb{R}^{n_1 \times n_2}$ with radius R with respect to Frobenius norm, there exists an ϵ-net $\mathcal{E}$ such that*

$$
|\mathcal{E}| \leq \left(1 + \frac{2R}{\epsilon}\right)^{n_1 n_2}.
$$

Proof. This claim follows by a direct application of Proposition 4.2.12 in (Vershynin, 2018). □

Lemma C.5 (Hoeffding's Inequality for Subgaussian Random Variables)**.** *Letting $\{z_i\}_{i=1}^n$ be a set of independent σ-subgaussian random variables, then for all $t \geq 0$, we have*

$$
\Pr\left[\frac{1}{n}\sum_{i=1}^{n} z_i \;\geq\; t\right] \;\leq\; \exp\left(-\frac{2nt^2}{\sigma^2}\right).
$$

Proof. See Proposition 2.1 of (Wainwright, 2016). □

Now, we prove Lemma C.3. Consider an $\epsilon/(4K)$-net $\mathcal{E}$. Then, for any $\phi \in \Phi$, there exists $\phi' \in \mathcal{E}$ such that

$$
|(\hat{\tilde{L}}(\phi; Z) - \tilde{L}_p(\phi)) - (\hat{\tilde{L}}(\phi'; Z) - \tilde{L}_p(\phi'))| \leq 2K\|\phi - \phi'\|_F \leq \frac{\epsilon}{2}.
$$

Therefore, we have

$$
\begin{aligned}
&\mathbb{P}_{p(Z)}\left[\sup_\theta |\hat{L}(g(\theta); Z) - L_p(g(\theta)) - \sigma(Z)| \geq \epsilon\right] \\
&= \mathbb{P}_{p(Z)}\left[\sup_{\phi \in \Phi} |\hat{\tilde{L}}(\phi; Z) - \tilde{L}_p(\phi) - \sigma(Z)| \geq \epsilon\right] \\
&\leq \mathbb{P}_{p(Z)}\left[\max_{\phi \in \mathcal{E}} |\hat{\tilde{L}}(\phi; Z) - \tilde{L}_p(\phi) - \sigma(Z)| \geq \frac{\epsilon}{2}\right] \\
&\leq \sum_{\phi \in \mathcal{E}} \mathbb{P}_{p(Z)}\left[|\hat{\tilde{L}}(\phi; Z) - \tilde{L}_p(\phi) - \sigma(Z)| \geq \frac{\epsilon}{2}\right].
\end{aligned} \tag{16}
$$

Now, defining

$$
\bar{\tilde{L}}(\phi; Z) = \frac{1}{n}\sum_{i=1}^{n} (\tilde{f}_\phi(x_i) - \tilde{f}_{\phi^*}(x_i))^2 \qquad \text{and} \qquad \tilde{\eta}(\phi; Z) = \frac{1}{n}\sum_{i=1}^{n} (\tilde{f}_\phi(x_i) - \tilde{f}_{\phi^*}(x_i))\xi_i,
$$

and recalling that $\sigma(Z) = n^{-1}\sum_{i=1}^n \xi_i^2$, then we have

$$\hat{\tilde{L}}(\phi; Z) = \bar{\tilde{L}}(\phi; Z) + 2\tilde{\eta}(\phi; Z) + \sigma(Z).$$

Thus, continuing from (16), we have

$$\begin{aligned}
&\sum_{\phi\in\mathcal{E}} \mathbb{P}_{p(Z)}\left[|\hat{\tilde{L}}(\phi; Z) - \tilde{L}_p(\phi) - \sigma(Z)| \geq \frac{\epsilon}{2}\right] \\
&\leq \sum_{\phi\in\mathcal{E}} \mathbb{P}_{p(Z)}\left[|\bar{\tilde{L}}(\phi; Z) - \tilde{L}_p(\phi)| + 2|\tilde{\eta}(\phi; Z)| \geq \frac{\epsilon}{2}\right] \\
&\leq \sum_{\phi\in\mathcal{E}} \mathbb{P}_{p(Z)}\left[|\bar{\tilde{L}}(\phi; Z) - \tilde{L}_p(\phi)| \geq \frac{\epsilon}{6}\right] + \mathbb{P}_{p(Z)}\left[|\tilde{\eta}(\phi; Z)| \geq \frac{\epsilon}{6}\right]. \qquad (17)
\end{aligned}$$

For the first term in (17), note that $|(\tilde{f}_\phi(x) - \tilde{f}_{\phi^*}(x))^2| \leq \ell_{\max}^2$, so $(\tilde{f}_\phi(x) - \tilde{f}_{\phi^*}(x))^2$ is $\ell_{\max}^2$-subgaussian; thus, by Lemma C.5, we have

$$\sum_{\phi\in\mathcal{E}} \mathbb{P}_{p(Z)}\left[|\bar{\tilde{L}}(\phi; Z) - \tilde{L}_p(\phi)| \geq \frac{\epsilon}{6}\right] \leq 2|\mathcal{E}| \cdot \exp\left(-\frac{n\epsilon^2}{18\ell_{\max}^4}\right). \qquad (18)$$

Next, for the second term in (17), note that $|(\tilde{f}_\phi(x_i) - \tilde{f}_{\phi^*}(x_i))\xi_i| \leq \ell_{\max}\xi_{\max}$, so $(\tilde{f}_\phi(x_i) - \tilde{f}_{\phi^*}(x_i))\xi_i$ is $\ell_{\max}\xi_{\max}$-subgaussian; thus, by Lemma C.5, we have

$$\sum_{\phi\in\mathcal{E}} \mathbb{P}_{p(Z)}\left[|\tilde{\eta}(\phi; Z)| \geq \frac{\epsilon}{6}\right] \leq 2|\mathcal{E}| \cdot \exp\left(-\frac{n\epsilon^2}{18\ell_{\max}^2\xi_{\max}^2}\right). \qquad (19)$$

Combining (18) & (19), continuing from (17), we have

$$\begin{aligned}
&\sum_{\phi\in\mathcal{E}} \mathbb{P}_{p(Z)}\left[|\bar{\tilde{L}}(\phi; Z) - \tilde{L}_p(\phi)| \geq \frac{\epsilon}{6}\right] + \mathbb{P}_{p(Z)}\left[|\tilde{\eta}(\phi; Z)| \geq \frac{\epsilon}{6}\right] \\
&\leq 4|\mathcal{E}| \cdot \exp\left(-\frac{n\epsilon^2}{18\ell_{\max}^2(\ell_{\max}^2 + \xi_{\max}^2)}\right) \\
&\leq 2\left(1 + \frac{8\phi_{\max}K}{\epsilon}\right)^{d^2} \cdot \exp\left(-\frac{n\epsilon^2}{18\ell_{\max}^2(\ell_{\max}^2 + \xi_{\max}^2)}\right) \\
&= 2\exp\left(-\frac{n\epsilon^2}{18\ell_{\max}^2(\ell_{\max}^2 + \xi_{\max}^2)} + d^2\log\left(1 + \frac{8\phi_{\max}K}{\epsilon}\right)\right), \qquad (20)
\end{aligned}$$

where for the first inequality, we have used $\max\{\ell_{\max}^2, \xi_{\max}^2\} \leq \ell_{\max}^2 + \xi_{\max}^2$, and the second inequality follows since by Lemma C.4, the covering number of the ϵ-net $\mathcal{E}$ of Φ satisfies

$$|\mathcal{E}| \leq \left(1 + \frac{2\phi_{\max}}{\epsilon}\right)^{d^2}.$$

Finally, we choose ϵ so that (20) is smaller than δ—in particular, letting

$$\epsilon = \sqrt{\frac{18\ell_{\max}^2(\ell_{\max}^2 + \xi_{\max}^2)}{n}\left(d^2\max\left\{1, \log\left(1 + \frac{8\phi_{\max}Kn}{\ell_{\max}^2}\right)\right\} + \log\frac{2}{\delta}\right)}.$$

then continuing (20), we have

$$2\exp\left(-\frac{n\epsilon^2}{18\ell_{\max}^2(\ell_{\max}^2 + \xi_{\max}^2)} + d^2\log\left(1 + \frac{8\phi_{\max}K}{\epsilon}\right)\right) \leq \delta,$$

as claimed. □

Algorithm 1 Explore-Then-Commit Algorithm for Quadratic Neural Network Bandit

Initialize $Z \leftarrow \varnothing$
Let m be as in (26)
for $t \in \{1, ..., m\}$ **do**
  Sample i.i.d. action $x_t \sim p$, where p is as in (9)
  Take action x_t and obtain reward y_t as in (8)
  Update $Z \leftarrow Z \cup \{(x_t, y_t)\}$
end for
Compute $\hat{\theta} = \arg\min_\theta \hat{L}(\theta; Z)$, where $\hat{L}(\theta; Z) = m^{-1} \sum_{i=1}^m (f_\theta(x_i) - y_i)^2$
Compute $\hat{x} = \arg\max_{x \in \mathcal{X}} f_{\hat{\theta}}(x)$
for $t \in \{m+1, ..., T\}$ **do**
  Take action $x_t = \hat{x}$ and obtain reward y_t as in (8)
end for

D. Proofs for Section 4

First, we make the following assumption, which says that $\phi^* = \theta^* \theta^{*\top}$ has a gap in its top eigenvalue:

Assumption D.1. Let $\phi^* = \theta^* \theta^{*\top}$, and let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d$ be the eigenvalues of ϕ^*. There exists a constant $M \in \mathbb{R}_{>0}$ such that $\lambda_1 - \lambda_2 \geq 4/M$.

This assumption ensures that the eigenvectors of ϕ^* to be stable under perturbations. The eigenvectors of ϕ^* correspond to the optimal action $x^* = \arg\max_{x \in \mathcal{X}} f_{\theta^*}(x)$ since $f_{\theta^*}(x) = x^\top \phi^* x$; thus, this assumption ensures that if $\hat{\theta} \approx \theta^*$, then the optimal action $\hat{x} = \arg\max_{x \in \mathcal{X}} f_{\hat{\theta}}(x)$ satisfies $\hat{x} \approx x^*$.

Then, we provide the full statement of Theorem 4.1 (including constants).

Theorem D.2. *The expected regret of Algorithm 1 is*

$$R(T) \leq C_0 + C_1 \cdot T^{2/3} \left(\log \left(3 + \frac{8\phi_{max} K T}{\ell_{max}^2} \right) \right)^{1/3},$$

where

$$C_0 = \frac{(\phi_{max})^{\frac{2d^2+2}{2d^2-1}}}{4(15M\ell_{max}^2 d^{7/5})^{\frac{2d^2+2}{2d^2-1}} (8\phi_{max} K/\ell_{max}^2)^{\frac{3d^2}{2d^2-1}}}, \qquad C_1 = 16^{4/3} d^{14/15} (M^2 \ell_{max}^4 \phi_{max})^{1/3}.$$

Before we prove Theorem D.2, we first prove a preliminary result establishing an analog of the *smooth best arm response* property (Rusmevichientong & Tsitsiklis, 2010) to our setting. First, we have the following useful result:

Lemma D.3. *Let $\phi, \phi' \in \mathbb{R}^{d \times d}$ be symmetric matrices, let $x, x' \in \mathbb{R}^d$ be eigenvectors of ϕ, ϕ' corresponding to their top eigenvalue, such that $\|x\|_2 = \|x'\|_2 = 1$, and let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d$ be the eigenvalues of ϕ'. Suppose that $\langle x, x' \rangle \geq 0$. Then, we have*

$$\|x - x'\| \leq \frac{2^{3/2} \|\phi - \phi'\|_{op}}{\lambda_2 - \lambda_1}.$$

Proof. See Corollary 3 of (Yu et al., 2015). □

Next, let $\chi : \mathbb{R}^{d \times d} \to 2^{\mathcal{X}}$ denote the subset of reward-maximizing arms for $g(\theta) = \theta\theta^\top$—i.e.,

$$\chi(\phi) = \arg\max_{x \in \mathcal{X}} x^\top \phi x,$$

where the argmax returns the set of all optimal values. Then, we have the following analog of smooth best arm response:

Lemma D.4. *For any $\phi \in \mathbb{R}^{d \times d}$, there exists $x \in \chi(\phi)$ and $x^* \in \chi(\phi^*)$ such that*

$$\|xx^\top - x^* x^{*\top}\|_F \leq M \|\phi - \phi^*\|_F.$$

Proof. First, note that x, x^* are eigenvectors of ϕ, ϕ^* corresponding to their top eigenvalues, respectively. Next, note that if $x^* \in \chi(\phi^*)$, then we also have $-x^* \in \chi(\phi^*)$; thus, without loss of generality, we can assume that $\langle x^*, x\rangle \geq 0$. Also, note that $\|x\|_F = \|x^*\|_F = 1$ since the optimizer maximizes the magnitude of x. Thus, we have

$$\|x - x^*\|_2 \leq \frac{2^{3/2}\|\phi - \phi^*\|_{\text{op}}}{\lambda_1 - \lambda_2} \leq M\|\phi - \phi^*\|_F,$$

where the first inequality follows by by Lemma D.3, and the second inequality follows by Assumption D.1, as claimed. □

Now, we prove Theorem D.2. The cumulative regret $R(T)$ of a horizon of T has that

$$\begin{aligned} R(T) &= \mathbb{E}\left[\sum_{t=1}^{T}(f_{\theta^*}(x^*) - f_{\theta^*}(x_t))\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{m}(f_{\theta^*}(x^*) - f_{\theta^*}(x_t)) + \sum_{t=m+1}^{T}(f_{\theta^*}(x^*) - f_{\theta^*}(x_t))\right] \\ &\leq 2m\phi_{\max} + \mathbb{E}\left[\sum_{t=m+1}^{T}\langle g(\hat{\theta}) - g(\theta^*), \hat{x}\hat{x}^\top - x^*x^{*\top}\rangle + \sum_{t=m+1}^{T}\langle g(\hat{\theta}), x^*x^{*\top} - \hat{x}\hat{x}^\top\rangle\right], \end{aligned} \tag{21}$$

where $\hat{\theta}$ is an estimator that minimizes the empirical loss of the first m samples, $\hat{x} \in \chi(g(\hat{\theta}))$ is maximizes the estimated expected reward $f_{\hat{\theta}}(x)$, and $\langle \phi, \phi'\rangle = \sum_{i,j=1}^{d}\phi_{ij}\phi'_{ij}$ is the matrix inner product. Since $\hat{x}$ is a maximizer of $f_{\hat{\theta}}(x) = \langle g(\hat{\theta}), xx^\top\rangle$, we have $\langle g(\hat{\theta}), x^*x^{*\top} - \hat{x}\hat{x}^\top\rangle \leq 0$, so the second term in (21) is zero. Thus, continuing from (21), we have

$$\begin{aligned} R(T) &\leq 2m\phi_{\max} + \mathbb{E}\left[\sum_{t=m+1}^{T}\langle g(\hat{\theta}) - g(\theta^*), \hat{x}\hat{x}^\top - x^*x^{*\top}\rangle\right] \\ &\leq 2m\phi_{\max} + (T-m)\mathbb{E}\left[\|g(\hat{\theta}) - g(\theta^*)\|_F\|\hat{x}\hat{x}^\top - x^*x^{*\top}\|_F\right]. \end{aligned} \tag{22}$$

To bound the second term in (22), note that

$$\|\hat{x}\hat{x}^\top - x^*x^{*\top}\|_F \leq \|\hat{x}\hat{x}^\top - \hat{x}x^{*\top}\|_F + \|\hat{x}x^{*\top} - x^*x^{*\top}\|_F \leq 2M\|g(\hat{\theta}) - g(\theta^*)\|_F,$$

where the last step follows by Lemma D.4.

Note that, following the discussion in Section 3, for the choice of p in (9), Assumption 3.2 holds for the dataset Z with $\alpha = d^{-2/5}/15$ for any $d \geq 2$. Then, by Theorem 3.3, we have

$$\|g(\hat{\theta}) - g(\theta^*)\|_F \leq \sqrt{\frac{2\epsilon}{\alpha}} = d^{1/5}\left(\frac{1800\ell_{\max}^4}{m}\left(d^2\max\left\{1, \log\left(1 + \frac{8\phi_{\max}Km}{\ell_{\max}^2}\right)\right\} + \log\frac{2}{\delta}\right)\right)^{1/4}$$

with probability at least $1 - \delta$. Now, defining the event

$$\mathcal{G} = \left\{\|g(\hat{\theta}) - g(\theta^*)\|_F \leq \sqrt{\frac{2\epsilon}{\alpha}}\right\},$$

letting

$$\delta = 2\exp\left(-d^2\max\left\{1, \log\left(1 + \frac{8\phi_{\max}Km}{\ell_{\max}^2}\right)\right\}\right),$$

and continuing from (22), we have

$$\begin{aligned} R(T) &\leq 2m\phi_{\max} + T\cdot\mathbb{E}\left[\|g(\hat{\theta}) - g(\theta^*)\|_F\|\hat{x}\hat{x}^\top - x^*x^{*\top}\|_F\mathbb{1}(\mathcal{G})\right] + 4T\phi_{\max}\cdot\mathbb{P}(\mathcal{G}^c) \\ &\leq 2m\phi_{\max} + 2MT\cdot\mathbb{E}\left[\|g(\hat{\theta}) - g(\theta^*)\|_F^2 \,\middle|\, \mathcal{G}\right] + 4T\phi_{\max}\cdot\mathbb{P}(\mathcal{G}^c) \\ &\leq 2m\phi_{\max} + 120M\ell_{\max}^2 d^{7/5}T\sqrt{\frac{\log(3 + 8\phi_{\max}KT/\ell_{\max}^2)}{m}} + \frac{8T\phi_{\max}}{(8\phi_{\max}Km/\ell_{\max}^2)^{d^2}}. \end{aligned} \tag{23}$$

The third term in inequality (23) is smaller than the second term when

$$m \geq \left(\frac{\phi_{\max}}{15M\ell_{\max}^2 d^{7/5}(8\phi_{\max}K/\ell_{\max}^2)^{d^2}\sqrt{\log(3+8\phi_{\max}KT/\ell_{\max}^2)}} \right)^{1/(d^2-1/2)}. \tag{24}$$

For a choice of m satisfying (24), continuing from (23), we have

$$R(T) \leq 2m\phi_{\max} + 240M\ell_{\max}^2 d^{7/5}T\sqrt{\frac{\log(3+8\phi_{\max}KT/\ell_{\max}^2)}{m}}. \tag{25}$$

Next, we choose m to minimize the upper bound in (24) for sufficiently large T—in particular,

$$m = \left\lceil \left(\frac{60M\ell_{\max}^2 d^{7/5}T\sqrt{\log(3+8\phi_{\max}KT/\ell_{\max}^2)}}{\phi_{\max}} \right)^{\frac{2}{3}} \right\rceil. \tag{26}$$

With this choice of m, we have

$$R(T) \leq 16^{\frac{4}{3}}(M^2\ell_{\max}^4\phi_{\max})^{\frac{1}{3}} d^{14/15}T^{2/3}(\log(3+8\phi_{\max}KT/\ell_{\max}^2))^{\frac{1}{3}}.$$

Finally, note that (24) holds under the choice of m in (25) for T satisfying

$$T\sqrt{\log(3+8\phi_{\max}KT/\ell_{\max}^2)} \geq \frac{(\phi_{\max})^{\frac{2d^2+2}{2d^2-1}}}{4(15M\ell_{\max}^2 d^{7/5})^{\frac{2d^2+2}{2d^2-1}}(8\phi_{\max}K/\ell_{\max}^2)^{\frac{3d^2}{2d^2-1}}}.$$

The claim follows. □

E. Proofs for Section 5

Note that we have for the proxy data

$$\mathbb{P}_{p(Z)}\left[L_q(\hat{\theta}_p) \leq \frac{2K^2\epsilon_p}{\alpha}\right] \geq 1 - \frac{\delta}{2},$$

where

$$\epsilon_p = \sqrt{\frac{18\ell_{\max}^2(\ell_{\max}^2+\xi_{\max}^2)}{n_p}\left(d^2\max\left\{1, \log\left(1+\frac{8\phi_{\max}Kn_p}{\ell_{\max}^2}\right)\right\} + \log\frac{4}{\delta}\right)},$$

where we have highlighted the differences from ϵ in (10) in red. Next, we make a technical assumption

Assumption E.1. For some $\sigma_0 \in \mathbb{R}_{>0}$, $\sigma_{\min}(\theta_p^*) \geq \sigma_0$, where $\sigma_{\min}(\theta)$ is the dth singular value of θ.

Equivalently, the minimum eigenvalue of $g(\theta_p^*)$ is positive; intuitively, this assumption ensures a good estimate of $\theta_p^*\theta_p^{*\top}$ implies a good estimate of θ_p^* (up to an orthogonal transformation).

Then, we provide a full statement of Theorem 5.1.

Theorem E.2. *For any distribution $q(x)$ with support on $B_2(0, x_{max})$, we have*

$$\mathbb{P}_{p(Z)}\left[L_q(\hat{\theta}_g) \leq \frac{2K^2\epsilon_g}{\alpha}\right] \geq 1 - \delta,$$

where

$$\epsilon_g = \hat{B}\cdot\sqrt{\frac{18K^2(K^2\hat{B}^2+\xi_{max}^2)}{n_g}\left(d^2\max\left\{1, \log\left(1+\frac{8\phi_{max}Kn_g}{\ell_{max}^2}\right)\right\} + \log\frac{4}{\delta}\right)}.$$

Before we prove Theorem E.2, we state the following key result:

Lemma E.3. *Let $\theta, \theta' \in \mathbb{R}^{d\times k}$, and let $\phi = \theta\theta^\top$ and $\phi' = \theta'\theta'^\top$. Assume that $\|\phi - \phi'\|_F \le \eta$, and that $\sigma_{min}(\theta) \ge \sigma_0 > 0$, where $\sigma_{min}(\theta)$ is the minimum singular value of θ (more precisely, the dth largest singular value). Then, there exist orthogonal matrices $R, R' \in \mathbb{R}^{k\times k}$ such that*

$$\|\theta R - \theta' R'\|_F \le \frac{\eta}{\sigma_0}. \tag{27}$$

Proof. Consider the SVDs $\theta = U\Sigma V^\top$ and $\theta' = U'\Sigma' V'^\top$, where $U, U' \in \mathbb{R}^{d\times d}$, $\Sigma, \Sigma' \in \mathbb{R}^{d\times d}$, and $V, V' \in \mathbb{R}^{k\times d}$; then, we have $\phi = U\Sigma^2 U^\top$ and $\phi' = U'\Sigma'^2 U'^\top$. Then, we claim that the choices $R = VU^\top$ and $R' = V'U'^\top$ satisfy (27). In particular, note that $\theta R = U\Sigma U^\top$ and $\theta' R' = U'\Sigma' U'^\top$, since $V^\top V = V'^\top V' = I_d$ since $k \ge d$, where $I_d \in \mathbb{R}^{d\times d}$ is the d-dimensional identity matrix. Thus, it suffices to show that

$$\sigma_0 \|U\Sigma U^\top - U'\Sigma' U'^\top\|_F \le \eta. \tag{28}$$

To this end, note that

$$\eta \ge \|\phi - \phi'\|_F = \|U\Sigma^2 U^\top - U'\Sigma'^2 U'^\top\|_F = \|U'^\top U\Sigma^2 - \Sigma'^2 U'^\top U\|_F, \tag{29}$$

where in the last step, we have multiplied the expression inside the Frobenius norm by $U'^\top$ on the left and by U on the right, using the fact that the Frobenius norm is invariant under multiplication by orthogonal matrices. Defining $W = U'^\top U$, note that

$$(W\Sigma)_{ij} = \sum_{k=1}^{d} W_{ik}\Sigma_{kj} = W_{ij}\Sigma_{jj} \tag{30}$$

$$(\Sigma W)_{ij} = \sum_{k=1}^{d} \Sigma'_{ik} W_{kj} = W_{ij}\Sigma'_{ii} \tag{31}$$

$$(W\Sigma^2)_{ij} = \sum_{k=1}^{d} W_{ik}(\Sigma^2)_{kj} = W_{ij}\Sigma_{jj}^2 \tag{32}$$

$$(\Sigma^2 W)_{ij} = \sum_{k=1}^{d} (\Sigma'^2)_{ik} W_{kj} = W_{ij}\Sigma_{ii}'^2. \tag{33}$$

Then, continuing from (29), we have

$$\begin{aligned}
\eta^2 \ge \|W\Sigma^2 - \Sigma'^2 W\|_F^2 &= \sum_{i,j=1}^{d} W_{ij}^2 (\Sigma_{jj}^2 - \Sigma_{ii}'^2)^2 \\
&= \sum_{i,j=1}^{d} W_{ij}^2 (\Sigma_{jj} - \Sigma'_{ii})^2 (\Sigma_{jj} + \Sigma'_{ii})^2 \\
&\ge \sum_{i,j=1}^{d} W_{ij}^2 (\Sigma_{jj} - \Sigma'_{ii})^2 \sigma_0^2 \\
&= \sigma_0^2 \|W\Sigma - \Sigma' W\|_F^2 \\
&= \sigma_0^2 \|U'^\top U\Sigma - \Sigma' U'^\top U\|_F^2 \\
&= \sigma_0^2 \|U\Sigma U^\top - U'\Sigma' U'^\top\|_F^2,
\end{aligned}$$

where on the first line, we have used (32) & (33), on the third line we have used $\Sigma_{jj} \ge \sigma_0$, on the fourth line we have used (30) & (31), and on the last line we have multiplied on by U' on the left $U^\top$ on the right, again using the fact that the Frobenius norm is invariant under multiplication by orthogonal matrices. Thus, we have shown (28), so the claim follows. □

Now, we prove Theorem E.2. First, by directly applying the arguments in the proof of Theorem 3.3, we have

$$\|g(\hat{\theta}_p) - g(\theta_p^*)\|_F \leq \sqrt{\frac{2\epsilon_p}{\alpha}}$$

with probability at least $1 - \delta/2$. However, $\hat{\theta}_p$ itself may not be close to θ_p^*. Instead, applying Lemma E.3 with $\theta = \hat{\theta}_p$ and $\theta' = \theta_p^*$, and with $\eta = \sqrt{2\epsilon_p/\alpha}$, there exists a orthogonal matrix $R_p = R'R^\top$ that "aligns" $\hat{\theta}_p$ with θ_p^*, yielding

$$\|\hat{\theta}_p - \theta_p^* R_p\|_F \leq \frac{1}{\sigma_0}\sqrt{\frac{2\epsilon_p}{\alpha}},$$

where σ_0 is the minimum singular value of θ_p^*. Now, let $\tilde{\theta}_g = \theta_g^* R_p$, and note that this is a global minimizer (i.e., $g(\tilde{\theta}_g) = g(\theta_g^*)$), since R_p is orthogonal. Then, we have

$$\begin{aligned}
\|\tilde{\theta}_g - \hat{\theta}_p\|_F &\leq \|\theta_g^* R_p - \theta_p^* R_p\|_F + \|\theta_p^* R_p - \hat{\theta}_p\|_F \\
&\leq \|\theta_g^* - \theta_p^*\|_F + \frac{1}{\sigma_0}\sqrt{\frac{2\epsilon_p}{\alpha}} \\
&\leq B + \frac{1}{\sigma_0}\sqrt{\frac{2\epsilon_p}{\alpha}} \qquad (34)
\end{aligned}$$

with probability at least $1 - \delta/2$. In other words, an alternative global minimizer $\tilde{\theta}_g$ exists within a small Frobenius norm of our proxy estimator $\hat{\theta}_p$, even if $\hat{\theta}_p$ is not close to θ_p^*.

Finally, on the event that (34) holds, note that for $\theta \in B_2(\hat{\theta}_p, \hat{B})$, we have the alternative upper bound

$$|f_\theta(x) - f_{\theta_g^*}(x))| \leq K\|g(\theta) - g(\theta_g^*)\|_F \leq K\hat{B},$$

where the first inequality holds by Lemma C.2; thus, we can take $\ell_{\text{max}} = K\hat{B}$. Thus, on the event that (34) holds, by Theorem 3.3, we have

$$\mathbb{P}_{p(Z)}\left[L_q(\hat{\theta}_g) \leq \frac{2K^2\epsilon_g}{\alpha}\right] \geq 1 - \frac{\delta}{2},$$

so the claim follows by a union bound. □

F. Proofs for Appendix B

F.1. Proof of Proposition B.4

Suppose that $z_t \in \{0, 1\}$ is binary, $z_0 = 0$, and

$$\tilde{p}(z \mid z') = \begin{cases} 1 & \text{if } z = z' \\ 0 & \text{otherwise.} \end{cases}$$

In particular, since $z_0 = 0$, $p(w) = \mathbb{1}(w = w_0)$ places all weight on the zero sequence $w_0 = 0...0$. Next, consider the shifted distribution

$$\tilde{q}(z_t \mid z_{t-1}) = \begin{cases} 1 & \text{if } z = z' = 1 \\ 1 - \alpha/2 & \text{if } z = z' = 0 \\ \alpha/2 & \text{otherwise.} \end{cases}$$

Note that $\|\tilde{p}(\cdot \mid z') - \tilde{q}(\cdot \mid z')\|_{\text{TV}} \leq \alpha$, so Assumption B.3 is satisfied. Note that

$$q(w_0) = \prod_{t=1}^{T} \tilde{q}(0 \mid 0) = (1 - \alpha/2)^T.$$

As a consequence, we have

$$\begin{aligned}\|p-q\|_{\mathrm{TV}} &= \sum_{w\in\mathcal{W}} |p(w)-q(w)| \\ &= |p(w_0)-q(w_0)| + \sum_{w\in\mathcal{W}\setminus\{w_0\}} q(w) \\ &= (1-(1-\alpha/2)^T) + (1-(1-\alpha/2)^T) \\ &= 2(1-(1-\alpha/2)^T),\end{aligned}$$

as claimed. □

F.2. Proof of Lemma B.5

Note that

$$\begin{aligned}&\|\tilde{q}_t-\tilde{p}_t\|_{\mathrm{TV}} \\ &= \sum_{j=1}^{k}\int |\tilde{q}_t(z,j)-\tilde{p}_t(z,j)|dz \\ &= \sum_{j=1}^{k}\sum_{j'=1}^{k}\int \mathbb{1}(j=\tilde{g}^*(z',j'))\cdot|\tilde{q}(z\mid z')\tilde{q}_{t-1}(z',j')-\tilde{p}(z\mid z')\tilde{p}_{t-1}(z',j')|dz'dz \\ &= \sum_{j'=1}^{k}\int |\tilde{q}(z\mid z')\tilde{q}_{t-1}(z',j')-\tilde{p}(z\mid z')\tilde{p}_{t-1}(z',j')|dz'dz \\ &\le \sum_{j'=1}^{k}\int |\tilde{q}(z\mid z')-\tilde{p}(z\mid z')|\cdot\tilde{q}_{t-1}(z',j')+\tilde{p}(z\mid z')\cdot|\tilde{q}_{t-1}(z',j')-\tilde{p}_{t-1}(z',j')|dz'dz \\ &\le \sum_{j'=1}^{k}\int \alpha\cdot\tilde{q}_{t-1}(z',j')+|\tilde{q}_{t-1}(z',j')-\tilde{p}_{t-1}(z',j')|dz' \\ &\le \alpha+\|\tilde{q}_{t-1}-\tilde{p}_{t-1}\|_{\mathrm{TV}}.\end{aligned}$$

Since $q_0(z,j)=p_0(z,j)$ for all $z\in\mathcal{Z}$ and $j\in[k]$, by induction, $\|\tilde{q}_t-\tilde{p}_t\|_{\mathrm{TV}}\le t\alpha$. Thus, we have

$$\|\tilde{q}-\tilde{p}\|_{\mathrm{TV}} \le \frac{1}{T}\sum_{t=1}^{T}\|\tilde{q}_t-\tilde{p}_t\| \le T\alpha,$$

as claimed. □

F.3. Proof of Lemma B.6

First, we prove the following lemma.

Lemma F.1. *We have*

$$\tilde{p}_t(z_t,j_{t-1}) = \sum_{j_1,\ldots,j_{t-2}}^{k}\int\left(\prod_{\tau=1}^{t-1}\mathbb{1}(j_\tau=\tilde{g}^*(z_\tau,j_{\tau-1}))\right)\cdot p(z_1,\ldots,z_t)dz_1\ldots dz_{t-1}.$$

Proof. For the base case, we have

$$\begin{aligned}\tilde{p}_2(z_2, j_1) &= \sum_{j_0=1}^{k} \int \mathbb{1}(j_1 = \tilde{g}^*(z_1, j_0)) \cdot \tilde{p}(z_2 \mid z_1) \cdot \tilde{p}_1(z_1, j_0) dz_1 \\ &= \sum_{j_0=1}^{k} \int \mathbb{1}(j_1 = \tilde{g}^*(z_1, j_0)) \cdot \tilde{p}(z_2 \mid z_1) \cdot \mathbb{1}(j = 0) \cdot \tilde{p}(z_1) dz_1 \\ &= \int \mathbb{1}(j_1 = \tilde{g}^*(z_1, j_0)) \cdot p(z_1, z_2) dz_1,\end{aligned}$$

as claimed. For the inductive case, we have

$$\begin{aligned}\tilde{p}_t(z_t, j_{t-1}) &= \sum_{j_{t-2}=1}^{k} \int \mathbb{1}(j_{t-1} = \tilde{g}^*(z_{t-1}, j_{t-2})) \cdot \tilde{p}(z_t \mid z_{t-1}) \cdot \tilde{p}_{t-1}(z_{t-1}, j_{t-2}) dz_{t-1} \\ &= \sum_{j_1, \ldots, j_{t-2}=1}^{k} \int \left(\prod_{\tau=1}^{t-1} \mathbb{1}(j_{\tau-1} = \tilde{g}^*(z_{\tau-1}, j_{\tau-2})) \right) \cdot p(z_1, ..., z_t) dz_1 ... dz_{t-1}.\end{aligned}$$

as claimed. □

Now, we prove Lemma B.6. First, note that for each $t \in [T]$, we have

$$\begin{aligned}
&\mathbb{P}_{p(w)} \left[(\hat{g}(w)_t \neq g^*(w)_t) \wedge \left(\bigwedge_{\tau=1}^{t-1} \hat{g}(w)_\tau = g^*(w)_\tau \right) \right] \\
&= \int \mathbb{1}(\hat{g}(w)_t \neq g^*(w)_t) \cdot \left(\prod_{\tau=1}^{t-1} \mathbb{1}(\hat{g}(w)_\tau = g^*(w)_\tau) \right) \cdot p(w) dw \\
&= \sum_{j_1, \ldots, j_{t-1}=1}^{k} \int \mathbb{1}(\hat{g}(w)_t \neq g^*(w)_t) \cdot \left(\prod_{\tau=1}^{t-1} \mathbb{1}(\hat{g}(w)_\tau = g^*(w)_\tau) \right) \cdot p(j_1 ... j_{t-1} \mid w) \cdot p(w) dw \\
&= \sum_{j_1, \ldots, j_{t-1}=1}^{k} \int \mathbb{1}(\hat{g}(w)_t \neq g^*(w)_t) \cdot \cdot \left(\prod_{\tau=1}^{t-1} \mathbb{1}(\hat{g}(w)_\tau = g^*(w)_\tau) \right) \cdot \left(\prod_{\tau=1}^{t-1} \mathbb{1}(j_\tau = \tilde{g}^*(z_\tau, j_{\tau-1})) \right) \\
&\qquad \cdot p(w) dw \\
&= \sum_{j_1, \ldots, j_{t-1}=1}^{k} \int \mathbb{1}(\hat{\tilde{g}}(z_\tau, j_{\tau-1}) \neq \tilde{g}^*(z_\tau, j_{\tau-1})) \cdot \left(\prod_{\tau=1}^{t-1} \mathbb{1}(\hat{\tilde{g}}(z_\tau, j_{\tau-1}) = \tilde{g}^*(z_\tau, j_{\tau-1})) \right) \\
&\qquad \cdot \left(\prod_{\tau=1}^{t-1} \mathbb{1}(j_\tau = \tilde{g}^*(z_\tau, j_{\tau-1})) \right) \cdot p(w) dw \\
&\leq \sum_{j_1, \ldots, j_{t-1}=1}^{k} \int \mathbb{1}(\hat{\tilde{g}}(z_\tau, j_{\tau-1}) \neq \tilde{g}^*(z_\tau, j_{\tau-1})) \cdot \left(\prod_{\tau=1}^{t-1} \mathbb{1}(j_\tau = \tilde{g}^*(z_\tau, j_{\tau-1})) \right) \cdot p(w) dw \\
&= \sum_{j_1, \ldots, j_{t-1}=1}^{k} \int \mathbb{1}(\hat{\tilde{g}}(z_\tau, j_{\tau-1}) \neq \tilde{g}^*(z_\tau, j_{\tau-1})) \cdot \left(\prod_{\tau=1}^{t-1} \mathbb{1}(j_\tau = \tilde{g}^*(z_\tau, j_{\tau-1})) \right) \cdot p(z_1, ..., z_t) dz_1 ... dz_t \\
&= \mathbb{P}_{p_t(z,j)} \left[\hat{\tilde{g}}(z_\tau, j_{\tau-1}) \neq \tilde{g}^*(z_\tau, j_{\tau-1}) \right],
\end{aligned}$$

where the last step follows from Lemma F.1. Now, note that

$$\begin{aligned}\mathbb{P}_{p(w)}[\hat{g}(w) \neq g^*(w)] &= \sum_{t=1}^{T} \mathbb{P}_{p(w)}\left[(\hat{g}(w)_t \neq g^*(w)_t) \wedge \left(\bigwedge_{\tau=1}^{t-1} \hat{g}(w)_\tau = g^*(w)_\tau\right)\right] \\ &\leq \sum_{t=1}^{T} \mathbb{P}_{\tilde{p}_t(z,j)}\left[\hat{\tilde{g}}(z,j) \neq \tilde{g}^*(z,j)\right] \\ &\leq T\epsilon_g,\end{aligned}$$

as claimed. □

F.4. Proof of Theorem B.7

First, we show that $\mathbb{P}_{q(z,j)}[\hat{\tilde{g}}(z,j) \neq \tilde{g}^*(z,j)] \leq \epsilon_g + T\alpha$. To this end, note that

$$\begin{aligned}&\mathbb{P}_{q(z,j)}[\hat{\tilde{g}}(z,j) \neq \tilde{g}^*(z,j)] \\ &= \mathbb{P}_{p(z,j)}[\hat{\tilde{g}}(z,j) \neq \tilde{g}^*(z,j)] + \mathbb{P}_{q(z,j)}[\hat{\tilde{g}}(z,j) \neq \tilde{g}^*(z,j)] - \mathbb{P}_{p(z,j)}[\hat{\tilde{g}}(z,j) \neq \tilde{g}^*(z,j)] \\ &\leq \epsilon_g + \sum_{j=1}^{k} \int \mathbb{1}(\hat{\tilde{g}}(z,j) \neq \tilde{g}^*(z,j)) \cdot |\tilde{q}(z,j) - \tilde{p}(z,j)| dz \\ &\leq \epsilon_g + \|\tilde{q} - \tilde{p}\|_{\mathrm{TV}} \\ &\leq \epsilon_g + T\alpha.\end{aligned}$$

Next, by Lemma B.6 with q in place of p and $\epsilon_g + T\alpha$ in place of ϵ_g, we have $\mathbb{P}_{q(w)}[\hat{g}(w) \neq g^*(w)] \leq T\epsilon_g + T^2\alpha$. Then, assuming that $\hat{g}(w) = g^*(w)$, we have

$$\begin{aligned}&\|f^*(x,w) - \hat{f}(x,w)\|_2 \\ &= \|(f^*_{j_T} \circ ... \circ f^*_{j_1})(x) - (\hat{f}_{j_T} \circ ... \circ \hat{f}_{j_1})(x)\|_2 \\ &\leq \sum_{t=1}^{T} \|(f^*_{j_T} \circ ... \circ f^*_{j_{t+1}} \circ f^*_{j_t} \circ \hat{f}_{j_{t-1}} \circ ... \circ \hat{f}_{j_1})(x) - (f^*_{j_T} \circ ... \circ f^*_{j_{t+1}} \circ \hat{f}_{j_t} \circ \hat{f}_{j_{t-1}} \circ ... \circ \hat{f}_{j_1})(x)\|_2 \\ &\leq \sum_{t=1}^{T} K^{T-t} \cdot \|(f^*_{j_t} \circ \hat{f}_{j_{t-1}} \circ ... \circ \hat{f}_{j_1})(x) - (\hat{f}_{j_t} \circ \hat{f}_{j_{t-1}} \circ ... \circ \hat{f}_{j_1})(x)\|_2 \\ &\leq \sum_{t=1}^{T} K^{T-t}\epsilon_f \\ &\leq T\epsilon_f \cdot \max\{K^{T-1}, 1\}.\end{aligned}$$

The claim follows by a union bound. □