AN ADVERSARIAL APPROACH TO STRUCTURAL ESTIMATION
Online Appendix

Tetsuya Kaji[1], Elena Manresa[2], and Guillaume Pouliot[1]

[1]University of Chicago
[2]New York University

August 3, 2023

S.1 EQUIVALENCE TO SMM WHEN D IS LOGISTIC

We show that the adversarial estimator with a logistic discriminator is asymptotically equivalent to SMM. Importantly, we do not assume that the logistic discriminator is "correctly specified" so the oracle discriminator D_θ may not take the form of a logistic classifier. In turn, we assume that the moments are correctly specified, $\mathbb{E}[X_i] = \mathbb{E}[X_{i,\theta_0}]$; however, the structural model may still be misspecified. As in Section 3, let $D(x;\lambda) = \Lambda(x^\top\lambda)$ be the logistic discriminator and λ_θ and $\hat{\lambda}_\theta$ be the population parameter and its estimator for each θ, respectively. We employ the same notation as Section 4.2.1.

In particular, we show that the adversarial estimator $\hat{\theta}$ with this discriminator is asymptotically equivalent to the following SMM estimator,

$$\tilde{\theta} \coloneqq \arg\min_{\theta\in\Theta} (\mathbb{E}_n[X] - \mathbb{E}_m[X_\theta])^\top \Omega (\mathbb{E}_n[X] - \mathbb{E}_m[X_\theta]) \quad \text{for} \quad \Omega \coloneqq \Big(\frac{\mathbb{E}[XX^\top]+\mathbb{E}[X_{\theta_0}X_{\theta_0}^\top]}{2}\Big)^{-1}.$$

This is optimally weighted when X and X_θ contain a constant term and the second-order moments are also correctly specified (viz. $\mathbb{E}[XX^\top] = \mathbb{E}[X_{\theta_0}X_{\theta_0}^\top]$), in which case Ω reduces to $\mathbb{E}[XX^\top]^{-1}$. For simplicity, we ignore estimation of Ω. To show their equivalence, we assume the following.

1. (Growing synthetic sample size) n/m converges.
2. (Smooth model) T_θ is twice continuously differentiable in θ for every $x \in \tilde{\mathcal{X}}$.
3. (Finite moments) $\mathbb{E}[XX^\top]$ is positive definite; $\mathbb{E}[\|\dot{X}_\theta\|^2]$ and $\mathbb{E}[\|\ddot{X}_\theta\|]$ are bounded uniformly in θ; $\mathbb{E}_m[\|X_\theta\|^2]$ and $\mathbb{E}_m[\|\dot{X}_\theta\|^2]$ converge uniformly in θ.
4. (Correctly specified moments) $\mathbb{E}[X] = \mathbb{E}[X_{\theta_0}]$.
5. (Identification of λ_{θ_0}) λ_{θ_0} is unique.
6. (Smooth discriminator) λ_θ is continuously differentiable in θ.

7. (Exact maximizer) $\hat{\lambda}_\theta$ is the exact maximizer of $\mathbb{M}_\theta(D(\cdot;\lambda))$ in that the FOC for $\hat{\lambda}_\theta$ is exactly zero for every $\theta \in \Theta$.
8. (Uniform convergence rate of discriminator) $\sup_\theta \|\hat{\lambda}_\theta - \lambda_\theta\| = O_P(n^{-1/2})$.
9. (Identification of θ_0) $\mathbb{E}[\dot{X}_{\theta_0}]$ is of full row rank.
10. (Exact minimizer) $\hat{\theta}$ is the exact minimizer of $\mathbb{M}_\theta(D(\cdot;\hat{\lambda}_\theta))$ in that the FOC for $\hat{\theta}$ is exactly zero.
11. (Consistency) $\hat{\theta}$ and $\tilde{\theta}$ are consistent for θ_0.

The FOC for λ_{θ_0} gives $\mathbb{E}[(1-\Lambda(X^\top\lambda_{\theta_0}))X] = \mathbb{E}[\Lambda(X_{\theta_0}^\top\lambda_{\theta_0})X_{\theta_0}]$. Conditions 4 and 5 imply $\lambda_{\theta_0} = 0$. The Taylor expansion of the FOC for $\hat{\lambda}_{\theta_0}$ yields $\sqrt{n}(\hat{\lambda}_{\theta_0} - 0) = \Omega\sqrt{n}(\mathbb{E}_n[X] - \mathbb{E}_m[X_\theta]) + o_P(1) \rightsquigarrow N(0, V_\lambda)$ for $V_\lambda \coloneqq \Omega[\mathrm{Var}(X) + \lim \frac{n}{m} \mathrm{Var}(X_{\theta_0})]\Omega$. Also, by the same reasoning as Section 4.2.1, $\sup_\theta \|\dot{\hat{\lambda}}_\theta - \dot{\lambda}_\theta\| = O_P(n^{-1/2})$.

Next, the envelope theorem simplifies the FOC for $\hat{\theta}$ to $\mathbb{E}_m[\Lambda(X_{\hat{\theta}}^\top\hat{\lambda}_{\hat{\theta}})\dot{X}_{\hat{\theta}}^\top\hat{\lambda}_{\hat{\theta}}] = 0$, whose Taylor expansion gives

$$0 = \mathbb{E}_m[\Lambda(X_{\theta_0}^\top\hat{\lambda}_{\theta_0})\dot{X}_{\theta_0}^\top\hat{\lambda}_{\theta_0}] + \mathbb{E}_m[\Lambda(X_{\theta_0}^\top\hat{\lambda}_{\theta_0})[(1-\Lambda(X_{\theta_0}^\top\hat{\lambda}_{\theta_0}))\dot{X}_{\theta_0}^\top\hat{\lambda}_{\theta_0}\hat{\lambda}_{\theta_0}^\top\dot{X}_{\theta_0} + A + \dot{X}_{\theta_0}^\top\dot{\hat{\lambda}}_{\theta_0}]](\hat{\theta}-\theta_0) + o_P(n^{-1/2})$$

where $A = [(\frac{\partial}{\partial\theta_1}\dot{X}_{\theta_0})^\top\hat{\lambda}_{\theta_0}, \cdots, (\frac{\partial}{\partial\theta_d}\dot{X}_{\theta_0})^\top\hat{\lambda}_{\theta_0}]$. As $\hat{\lambda}_{\theta_0} \to 0$ and $\dot{\hat{\lambda}}_{\theta_0} \to \dot{\lambda}_{\theta_0}$, this becomes

$$\sqrt{n}(\hat{\theta}-\theta_0) = -\mathbb{E}[\dot{X}_{\theta_0}^\top\dot{\lambda}_{\theta_0}]^{-1}\mathbb{E}[\dot{X}_{\theta_0}^\top]\sqrt{n}(\hat{\lambda}_{\theta_0} - 0) + o_P(1).$$

As in Section 4.2.1, we have $\dot{\lambda}_{\theta_0} = -\Omega\mathbb{E}[\dot{X}_{\theta_0}]$, which yields $\sqrt{n}(\hat{\theta}-\theta_0) \rightsquigarrow N(0, V_\theta)$ for $V_\theta \coloneqq (\mathbb{E}[\dot{X}_{\theta_0}^\top]\Omega\mathbb{E}[\dot{X}_{\theta_0}])^{-1}\mathbb{E}[\dot{X}_{\theta_0}^\top]V_\lambda\mathbb{E}[\dot{X}_{\theta_0}](\mathbb{E}[\dot{X}_{\theta_0}^\top]\Omega\mathbb{E}[\dot{X}_{\theta_0}])^{-1}$.

Meanwhile, the FOC for SMM, $\mathbb{E}_m[\dot{X}_{\tilde{\theta}}^\top]\Omega(\mathbb{E}_n[X] - \mathbb{E}_m[X_{\tilde{\theta}}]) = 0$, expands as

$$0 = \mathbb{E}_m[\dot{X}_{\theta_0}^\top]\Omega(\mathbb{E}_n[X] - \mathbb{E}_m[X_{\theta_0}]) + (B - \mathbb{E}_m[\dot{X}_{\theta_0}^\top]\Omega\mathbb{E}_m[\dot{X}_{\theta_0}])(\tilde{\theta}-\theta_0) + o_P(n^{-1/2})$$

where $B = [\mathbb{E}_m[\frac{\partial\dot{X}_{\theta_0}}{\partial\theta_1}]^\top\Omega(\mathbb{E}_n[X]-\mathbb{E}_m[X_{\theta_0}]), \ldots, \mathbb{E}_m[\frac{\partial\dot{X}_{\theta_0}}{\partial\theta_d}]^\top\Omega(\mathbb{E}_n[X]-\mathbb{E}_m[X_{\theta_0}])]$. Thus, $\sqrt{n}(\tilde{\theta}-\theta_0) = -(\mathbb{E}[\dot{X}_{\theta_0}^\top]\Omega\mathbb{E}[\dot{X}_{\theta_0}])^{-1}\mathbb{E}[\dot{X}_{\theta_0}^\top]\Omega\sqrt{n}(\mathbb{E}_n[X]-\mathbb{E}_m[X_{\theta_0}]) + o_P(1)$, which shows that $\hat{\theta}$ and $\tilde{\theta}$ are asymptotically equivalent in probability as well as in distribution.

Remark. If X and X_θ have a constant term and the second-order moments are correctly specified, V_θ simplifies to $[1 + \lim \frac{n}{m}](\mathbb{E}[\dot{X}_{\theta_0}^\top]\mathbb{E}[XX^\top]^{-1}\mathbb{E}[\dot{X}_{\theta_0}])^{-1}$.

S.2 CONVERGENCE RATES OF THE DISCRIMINATOR

This section establishes the rate of convergence of the discriminator. In addition to results on a general nonparametric discriminator, we present results specific to a neural network discriminator.

The distance of discriminators is measured by a Hellinger-like distance

$$d_\theta(D_1, D_2) \coloneqq \sqrt{h_\theta(D_1, D_2)^2 + h_\theta(1 - D_1, 1 - D_2)^2}$$

where $h_\theta(D_1, D_2) \coloneqq \sqrt{(P_0 + P_\theta)(\sqrt{D_1} - \sqrt{D_2})^2}$.

The size of the neural network sieve is usually measured by the uniform and bracketing entropies. Conceptually, the bracketing entropy gives a stronger bound than the uniform entropy and yields a tighter convergence rate. It also goes nicely with the Bernstein norm that is useful for maximal inequalities for the log likelihood ratio (as well as our discriminators). For this, we go with the bracketing entropy. See van der Vaart and Wellner (2011) for more comparison of the two entropy notions.

Definition (Bracketing number and bracketing entropy integral)**.** The *ε-bracketing number* $N_{[]}(\varepsilon, \mathcal{F}, d)$ of a set $\mathcal{F}$ with respect to a premetric d is the minimal number of ε-brackets in d needed to cover $\mathcal{F}$. The *δ-bracketing entropy integral* of $\mathcal{F}$ with respect to d is $J_{[]}(\delta, \mathcal{F}, d) \coloneqq \int_0^\delta \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{F}, d)} d\varepsilon$.

The results on convergence of the discriminator are stated pointwise in $\theta \in \Theta$, so the discussion is made for fixed θ. Let δ_n be a nonnegative sequence.

S.2.1 General Nonparametric Discriminator

Let $\mathcal{D}_{\theta,\delta} \coloneqq \{D \in \mathcal{D}_n : d_\theta(D, D_\theta) \leq \delta\}$. We first assume that the sieve does not grow too fast.

Assumption S.1 (Entropy of sieve)**.** The entropy integral satisfies $J_{[]}(\delta_n, \mathcal{D}_{\theta,\delta_n}, d_\theta) \lesssim \delta_n^2 \sqrt{n}$. Also, there exists $\alpha < 2$ such that $J_{[]}(\delta, \mathcal{D}_{\theta,\delta}, d_\theta)/\delta^\alpha$ has a majorant decreasing in $\delta > 0$. □

The estimated discriminator need not be the exact maximizer of the loss but is required to maximize it up to some rate.

Assumption S.2 (Approximately maximizing discriminator)**.** The trained discriminator $\hat{D}_\theta$ satisfies $\mathbb{M}_\theta(\hat{D}_\theta) \geq \mathbb{M}_\theta(D_\theta) - O_P(\delta_n^2)$. □

In a sense, we can interpret Assumption S.1 as a requirement that the sieve be not too rich and Assumption S.2 that the sieve be rich enough. For example, if $\mathcal{D}_{\theta,\delta_n}$ is an empty set, Assumption S.1 is trivially satisfied, but there is no way to attain Assumption S.2. On the contrary, if $\mathcal{D}_n$ contains every function, there would exist an element in $\mathcal{D}_n$ that satisfies Assumption S.2 but Assumption S.1 will be violated. Both assumptions collectively require that the sieve is small but good enough for D_θ. With these, we obtain the rate of convergence of the discriminator.

Theorem S.1 (Rate of convergence of discriminator)**.** *Under Assumptions 2, S.1, and S.2,* $d_\theta(\hat{D}_\theta, D_\theta) = O_P^*(\delta_n)$.

One interesting observation is that Theorem S.1 does not require convergence of the objective function. This is reminiscent of the nonparametric maximum likelihood literature. To prove it without requiring convergence of the objective function, we think in terms of a pseudo-objective function. Let $m_q^p \coloneqq \log \frac{p+q}{2q}$ and

$$\tilde{M}_\theta(D) \coloneqq P_0 m_{D_\theta}^D + P_\theta m_{1-D_\theta}^{1-D}, \qquad \tilde{\mathbb{M}}_\theta(D) \coloneqq \mathbb{P}_0 m_{D_\theta}^D + \mathbb{P}_\theta m_{1-D_\theta}^{1-D}.$$

Proof. The concavity of the logarithm and Assumption S.2 imply $\tilde{\mathbb{M}}_\theta(\hat{D}_\theta) - \tilde{\mathbb{M}}_\theta(D_\theta) \geq \frac{1}{2}[\mathbb{M}_\theta(\hat{D}_\theta) - \mathbb{M}_\theta(D_\theta)] \geq -O_P(\delta_n^2)$. Then, apply van der Vaart and Wellner (1996, Theorem 3.4.1) with Lemma S.1 and Assumption S.1. ■

The following is a maximal inequality used to prove Theorem S.1. Let $\mathcal{M}_{\theta,\delta}^1 \coloneqq \{m_{D_\theta}^D : D \in \mathcal{D}_{\theta,\delta}\}$ and $\mathcal{M}_{\theta,\delta}^2 \coloneqq \{m_{1-D_\theta}^{1-D} : D \in \mathcal{D}_{\theta,\delta}\}$.

Lemma S.1 (Maximal inequality for pseudo-cross-entropy discriminator)**.** *For every* $D \in \mathcal{D}$, $\tilde{M}_\theta(D) - \tilde{M}_\theta(D_\theta) \leq -d_\theta(D, D_\theta)^2/(1+\sqrt{2})^2$. *For every* $\delta > 0$,

$$\mathbb{E}^* \sup_{D \in \mathcal{D}_{\theta,\delta}} \sqrt{n} \left| (\tilde{\mathbb{M}}_\theta - \tilde{M}_\theta)(D) - (\tilde{\mathbb{M}}_\theta - \tilde{M}_\theta)(D_\theta) \right| \lesssim J_{[]}(\delta, \mathcal{D}_{\theta,\delta}, d_\theta) \left[1 + \sqrt{\tfrac{n}{m}} + (1 + \tfrac{n}{m}) \tfrac{J_{[]}(\delta, \mathcal{D}_{\theta,\delta}, d_\theta)}{\delta^2 \sqrt{n}} \right].$$

Proof. Since $\log x \leq 2(\sqrt{x} - 1)$ for every $x > 0$,

$$P_0 \log \tfrac{D}{D_\theta} \leq 2P_0 \left(\sqrt{\tfrac{D}{D_\theta}} - 1 \right) = \left[2P_0 \tfrac{\sqrt{D(p_0 + p_\theta)}}{\sqrt{p_0}} - \int D(p_0 + p_\theta) - \int p_0 \right] + (P_0 + P_\theta)(D - D_\theta) = -h_\theta(D, D_\theta)^2 + (P_0 + P_\theta)(D - D_\theta).$$

Similarly, $P_\theta \log \frac{1-D}{1-D_\theta} \le -h_\theta(1-D, 1-D_\theta)^2 - (P_0 + P_\theta)(D - D_\theta)$. Replacing D and $1-D$ with $(D + D_\theta)/2$ and $(1 - D + 1 - D_\theta)/2$ and summing them up yield

$$P_0 m_{D_\theta}^D + P_\theta m_{1-D_\theta}^{1-D} \le -h_\theta\Big(\tfrac{D+D_\theta}{2}, D_\theta\Big)^2 - h_\theta\Big(\tfrac{1-D+1-D_\theta}{2}, 1 - D_\theta\Big)^2.$$

Since $\sqrt{2} h_\theta(\frac{p+q}{2}, q) \le h_\theta(p, q) \le (1 + \sqrt{2}) h_\theta(\frac{p+q}{2}, q)$ (van der Vaart and Wellner, 1996, Problem 3.4.4), we obtain the first inequality. For the second inequality, observe that

$$\sqrt{n}\Big[(\tilde{\mathbb{M}}_\theta - \tilde{M}_\theta)(D) - (\tilde{\mathbb{M}}_\theta - \tilde{M}_\theta)(D_\theta)\Big] = \sqrt{n}(\mathbb{P}_0 - P_0) m_{D_\theta}^D + \sqrt{n}(\mathbb{P}_\theta - P_\theta) m_{1-D_\theta}^{1-D}.$$

Therefore, it suffices to separately bound

$$\mathbb{E}^* \sup_{D \in \mathcal{D}_{\theta,\delta}} \Big|\sqrt{n}(\mathbb{P}_0 - P_0) m_{D_\theta}^D\Big| \quad \text{and} \quad \sqrt{\tfrac{n}{m}}\, \mathbb{E}^* \sup_{D \in \mathcal{D}_{\theta,\delta}} \Big|\sqrt{m}(\mathbb{P}_\theta - P_\theta) m_{1-D_\theta}^{1-D}\Big|.$$

Since $m_{D_\theta}^D, m_{1-D_\theta}^{1-D} \ge \log(1/2)$ and $e^{|x|} - 1 - |x| \le 4(e^{x/2} - 1)^2$ for every $x \ge \log(1/2)$,

$$\|m_{D_\theta}^D\|_{P_0,B}^2 \le 8 P_0 (e^{m_{D_\theta}^D/2} - 1)^2 \le 8 h_\theta(\tfrac{D+D_\theta}{2}, D_\theta)^2 \le 4 h_\theta(D, D_\theta)^2,$$
$$\|m_{1-D_\theta}^{1-D}\|_{P_\theta,B}^2 \le 4 h_\theta(1-D, 1-D_\theta)^2.$$

By van der Vaart and Wellner (1996, Lemma 3.4.3), the first supremum is bounded by $J_{[]}(2\delta, \mathcal{M}_{\theta,\delta}^1, \|\cdot\|_{P_0,B})[1 + J_{[]}(2\delta, \mathcal{M}_{\theta,\delta}^1, \|\cdot\|_{P_0,B})/(4\delta^2\sqrt{n})]$. Let $[\ell, u]$ be an ε-bracket in $\mathcal{D}$ with respect to d_θ. Since $u - \ell \ge 0$ and $e^{|x|} - 1 - |x| \le 2(e^{x/2} - 1)^2$ for $x \ge 0$,

$$\begin{aligned}\Big\|m_{D_\theta}^u - m_{D_\theta}^\ell\Big\|_{P_0,B}^2 \le 4 \int \Big(\sqrt{\tfrac{u+D_\theta}{\ell+D_\theta}} - 1\Big)^2 p_0 &\le 4 \int \Big(\sqrt{u + D_\theta} - \sqrt{\ell + D_\theta}\Big)^2 (p_0 + p_\theta) \\ &\le 4 h_\theta(u, \ell)^2 \le 4\varepsilon^2.\end{aligned}$$

Thus, $[m_{D_\theta}^\ell, m_{D_\theta}^u]$ makes a 2ε-bracket in $\mathcal{M}_{\theta,\delta}^1$ with respect to $\|\cdot\|_{P_0,B}$, so $J_{[]}(2\delta, \mathcal{M}_{\theta,\delta}^1, \|\cdot\|_{P_0,B}) \le 2 J_{[]}(\delta, \mathcal{D}_{\theta,\delta}, d_\theta)$. Analogous argument for the second supremum yields the second inequality. ■

S.2.2 Cross-Entropy Loss

To show convergence of the objective function, we need to make an additional assumption that the tails of the discriminators in the sieve are not too thin. This assumption would be trivial if we assume a compact support for the observables X_i and $X_{i,\theta}$, which is standard in the neural network literature.

Assumption S.3 (Support compatibility)**.** Define $P(X|A)$ to be $P(X\mathbb{1}\{A\})/P(A)$ if $P(A) > 0$ and 0 otherwise. There exists M such that

$$\sup_{D\in\mathcal{D}_{\theta,\delta_n}} P_0\Big(\tfrac{D_\theta}{D} \;\Big|\; \tfrac{D_\theta}{D} \geq \tfrac{25}{16}\Big) < M, \quad \sup_{D\in\mathcal{D}_{\theta,\delta_n}} P_\theta\Big(\tfrac{1-D_\theta}{1-D} \;\Big|\; \tfrac{1-D_\theta}{1-D} \geq \tfrac{25}{16}\Big) < M.$$

Also, the brackets $\{\ell \leq D \leq u\}$ in Assumption S.1 can be taken so that $(P_0 + P_\theta)(\frac{D_\theta}{\ell}(\sqrt{u} - \sqrt{\ell})^2)$ and $(P_0 + P_\theta)(\frac{1-D_\theta}{1-u}(\sqrt{1-\ell} - \sqrt{1-u})^2)$ are $O(d_\theta(u,\ell)^2)$. □

With this, we obtain the rate for the estimated cross-entropy loss.

Theorem S.2 (Rate of convergence of objective function)**.** *Under Assumptions 2 and S.1 to S.3,* $\mathbb{M}_\theta(\hat{D}_\theta) - \mathbb{M}_\theta(D_\theta) = O_P^*(\delta_n^2)$.

Proof. Since $\mathbb{M}_\theta(\hat{D}_\theta) - \mathbb{M}_\theta(D_\theta) \geq -O_P(\delta_n^2)$ by Assumption S.2, we need only to prove the reverse inequality. With $\log(x) \leq 2(\sqrt{x} - 1)$ for $x > 0$, for every D,

$$\begin{aligned}&\mathbb{M}_\theta(D) - \mathbb{M}_\theta(D_\theta)\\ &\quad\leq 2P_0\Big(\sqrt{\tfrac{D}{D_\theta}} - 1\Big) + 2P_\theta\Big(\sqrt{\tfrac{1-D}{1-D_\theta}} - 1\Big) + (\mathbb{P}_0 - P_0)\log\tfrac{D}{D_\theta} + (\mathbb{P}_\theta - P_\theta)\log\tfrac{1-D}{1-D_\theta}.\end{aligned}$$

As in Lemma S.1, the first two terms are equal to $-d_\theta(D, D_\theta)^2$. Since Theorem S.1 implies $d_\theta(\hat{D}_\theta, D_\theta)^2 = O_P^*(\delta_n^2)$, it remains to show that the last two terms are of the same order. We bound the suprema,

$$\mathbb{E}^* \sup_{D\in\mathcal{D}_{\theta,\delta_n}} \Big|\sqrt{n}(\mathbb{P}_0 - P_0)\log\tfrac{D}{D_\theta}\Big| \quad \text{and} \quad \mathbb{E}^* \sup_{D\in\mathcal{D}_{\theta,\delta_n}} \Big|\sqrt{m}(\mathbb{P}_\theta - P_\theta)\log\tfrac{1-D}{1-D_\theta}\Big|.$$

Under Assumption S.3, it follows from (the remark after) Lemma S.4 that for $D \in \mathcal{D}_{\theta,\delta_n}$,

$$\Big\|\tfrac{1}{2}\log\tfrac{D}{D_\theta}\Big\|_{P_0,B}^2 \leq 2(1+M)h_\theta(D, D_\theta)^2, \;\; \Big\|\tfrac{1}{2}\log\tfrac{1-D}{1-D_\theta}\Big\|_{P_\theta,B}^2 \leq 2(1+M)h_\theta(1-D, 1-D_\theta)^2.$$

Assumption S.3 also implies that an ε-bracket in $\mathcal{M}_{\theta,\delta}^1$ induces

$$\begin{aligned}\Big\|\log\tfrac{u}{D_\theta} - \log\tfrac{\ell}{D_\theta}\Big\|_{P_0,B}^2 &\leq 4P_0\Big(\sqrt{\tfrac{u}{\ell}} - 1\Big)^2 = 4(P_0 + P_\theta)\tfrac{D_\theta}{\ell}(\sqrt{u} - \sqrt{\ell})^2 \leq Cd_\theta(u,\ell)^2,\\ \Big\|\log\tfrac{1-\ell}{1-D_\theta} - \log\tfrac{1-u}{1-D_\theta}\Big\|_{P_\theta,B}^2 &\leq 4(P_0 + P_\theta)\tfrac{1-D_\theta}{1-u}(\sqrt{1-\ell} - \sqrt{1-u})^2 \leq Cd_\theta(u,\ell)^2,\end{aligned}$$

for some $C > 0$. By similar arguments as in the proof of Lemma S.1, the two suprema are of orders $\sqrt{n}\delta_n^2$ and $\sqrt{m}\delta_n^2$.[1] With Assumption 2 follows the theorem. ■

[1] We can write $\|\frac{1}{2}\log\frac{D}{D_\theta}\|_{P_0,B}^2 \leq [2(1+M) \vee C]h_\theta(D, D_\theta)^2$ and $\|\log\frac{u}{D_\theta} - \log\frac{\ell}{D_\theta}\|_{P_0,B}^2 \leq [2(1+$

S.2.3 Neural Network Discriminator

The results in Appendix S.2.1 apply to any nonparametric sieve discriminator. Given a particular sieve, the specific convergence rate is determined by the δ_n that satisfies Assumption S.1. In the nonparametric estimation literature, it is often observed that δ_n gets slower as the dimension d of the input X_i increases. In the context of nonparametric regression, Bauer and Kohler (2019) show that a particular type of neural network estimator does not have a rate that slows with d but only with d^*, the "underlying dimension" of the target function.[2] We believe that the structure they impose on the target function arises very naturally in economic models, and want to incorporate the "remedy for the curse of dimensionality" aspect into our theory.

In light of this, we develop the "classification counterpart" of the results in Bauer and Kohler (2019). Instead of the target regression function, we exploit the low-dimensional composite structure on the log likelihood ratio $\log(p_0/p_\theta)$. We note that our theory does not *require* that there is such a low-dimensional structure; if there is none, we have $d^* = d$ and our result reduces to a regular nonparametric rate with the curse of dimensionality.

Intuitively, the low-dimensional composite structure is described as follows. Note that the log likelihood ratio $\log(p_0/p_\theta)$ takes a d-dimensional input X as its argument, where d can be large. We need that this ratio admits a representation as a nested composition of smooth functions, each of which takes a possibly smaller number d^* of arguments. In the first layer of composition, we assume a linear index structure to reduce d arguments into d^* intermediate outputs.

To develop a precise definition, we start with the notion of smoothness we use.

Definition ((p, C)-smoothness; Bauer and Kohler, 2019, Definition 1)**.** Let $p = q + s$ for some $q \in \mathbb{N}_0$ and $0 < s \leq 1$. A function $m : \mathbb{R}^d \to \mathbb{R}$ is called (p, C)-*smooth* if for every $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = q$, the partial derivative $\frac{\partial^q m}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$ exists and satisfies

$$\left| \frac{\partial^q m}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}(x) - \frac{\partial^q m}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}(z) \right| \leq C \|x - z\|^s$$

for every $x, z \in \mathbb{R}^d$ where $\|\cdot\|$ denotes the Euclidean norm.

With this, the nested composition structure is defined as follows.

$M) \vee C] d_\theta(u, \ell)^2$ to apply the same argument as Theorem S.1.

[2] Bauer and Kohler (2019) call d^* the *order*.

Definition (Generalized hierarchical interaction model; Bauer and Kohler, 2019, Definition 2)**.** Let $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$, and $m : \mathbb{R}^d \to \mathbb{R}$. We say that the function m satisfies a *generalized hierarchical interaction model of order* d^* *and level* 0, if there exist $a_1 \in \mathbb{R}^d, \dots, a_{d^*} \in \mathbb{R}^d$, and $f : \mathbb{R}^{d^*} \to \mathbb{R}$ such that

$$m(x) = f(a_1^\top x, \dots, a_{d^*}^\top x)$$

for every $x \in \mathbb{R}^d$. We say that m satisfies a *generalized hierarchical interaction model of order* d^* *and level* $l+1$ *with* K *components* if there exist $g_k : \mathbb{R}^{d^*} \to \mathbb{R}$ and $f_{1,k}, \dots, f_{d^*,k} : \mathbb{R}^d \to \mathbb{R}$ $(k = 1, \dots, K)$ such that $f_{1,k}, \dots, f_{d^*,k}$ $(k = 1, \dots, K)$ satisfy a generalized hierarchical model of order d^* and level l and

$$m(x) = \sum_{k=1}^{K} g_k(f_{1,k}(x), \dots, f_{d^*,k}(x))$$

for every $x \in \mathbb{R}^d$. We say that the generalized hierarchical interaction model is (p, C)*-smooth* if all functions occurring in its definition are (p, C)-smooth.

For example, a conditional binary choice model yields a log likelihood ratio that satisfies a generalized hierarchical interaction model of order $d^* \leq 3$ and level 0, irrespectively of the dimension of the covariates.

Example S.1 (Binary choice model)**.** Let $y_i = \mathbb{1}\{x_i^\top \alpha + \varepsilon_i > 0\}$, $\varepsilon_i \sim P_\varepsilon$, be the true DGP and $y_i = \mathbb{1}\{x_i^\top \beta + \tilde{\varepsilon}_i > 0\}$, $\tilde{\varepsilon}_i \sim \tilde{P}_\varepsilon$, be the structural model. Then,

$$\log \frac{p_0(y, x)}{p_\theta(y, x)} = y \log \frac{1 - P_\varepsilon(-x^\top \alpha)}{1 - \tilde{P}_\varepsilon(-x^\top \beta)} + (1 - y) \log \frac{P_\varepsilon(-x^\top \alpha)}{\tilde{P}_\varepsilon(-x^\top \beta)}.$$

Therefore, we can write this as $f(a_1^\top z, a_2^\top z, a_3^\top z)$ where $z = (y, x^\top)^\top$, $a_1 = (1, 0, \dots, 0)^\top$, $a_2 = (0, -\alpha^\top)^\top$, $a_3 = (0, -\beta^\top)^\top$, and $f(y, x_1, x_2) = y[\log(1 - P_\varepsilon(x_1)) - \log(1 - \tilde{P}_\varepsilon(x_2))] + (1 - y)[\log P_\varepsilon(x_1) - \log \tilde{P}_\varepsilon(x_2)]$. □

Neural networks approximate functions by a nested composition of activation functions. For theoretical development, we define the following structure on the neural network estimator.

Definition (Hierarchical neural network; Bauer and Kohler, 2019, Section 2)**.** Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a q-admissible activation function. For $M^* \in \mathbb{N}$, $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$,

and $\alpha > 0$, let $\mathcal{F}_{M^*,d^*,d,\alpha}$ be the class of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that

$$f(x) = \sum_{i=1}^{M^*} \mu_i \sigma\left(\sum_{j=1}^{4d^*} \lambda_{i,j} \sigma\left(\sum_{v=1}^{d} \theta_{i,j,v} x_v + \theta_{i,j,0}\right) + \lambda_{i,0}\right) + \mu_0$$

for some $\mu_i, \lambda_{i,j}, \theta_{i,j,v} \in \mathbb{R}$, where $|\mu_i| \leq \alpha$, $|\lambda_{i,j}| \leq \alpha$, and $|\theta_{i,j,v}| \leq \alpha$. For $l = 0$, define the set of neural networks with two hidden layers by $\mathcal{H}^{(0)}_{M^*,d^*,d,\alpha} \coloneqq \mathcal{F}_{M^*,d^*,d,\alpha}$; for $l > 0$, define the set of neural networks with $2l + 2$ hidden layers by

$$\mathcal{H}^{(l)}_{M^*,d^*,d,\alpha} \coloneqq$$
$$\left\{h : \mathbb{R}^d \to \mathbb{R} : h(x) = \sum_{k=1}^{K} g_k(f_{1,k}(x), \ldots, f_{d^*,k}(x)), g_k \in \mathcal{F}_{M^*,d^*,d^*,\alpha},\ f_{j,k} \in \mathcal{H}^{(l-1)}\right\}.$$

Now, we assume that the log likelihood ratio admits a hierarchical representation and that the neural network has a corresponding hierarchical structure.

Assumption S.4 (Neural network discriminator)**.** Let P_0 and P_θ have subexponential tails and finite first moments.[3] Let $\log(p_0/p_\theta)$ satisfy a (p, C)-smooth generalized hierarchical interaction model of order d^* and finite level l with K components for $p = q + s$, $q \in \mathbb{N}_0$, and $s \in (0, 1]$. Let $\mathcal{H}^{(l)}_{M^*,d^*,d,\alpha}$ be the class of neural networks with the Lipschitz activation function with Lipschitz constant 1 for

$$M_* = \left\lceil \binom{d^* + q}{d^*} (q+1) \left(\left\lceil \frac{(\log \delta_n)^{2(2q+3)}}{\delta_n} \right\rceil^{\frac{1}{p}} + 1 \right)^{d^*} \right\rceil,$$
$$\alpha = \left\lceil \frac{(\log \delta_n)^{2(2q+3)}}{\delta_n} \right\rceil^{\frac{d^* + p(2q+3)+1}{p}} \frac{\log n}{\delta_n^2},$$

and $\delta_n = [(\log n)^{\frac{p+2d^*(2q+3)}{p}} / n]^{\frac{p}{2p+d^*}}$. Denote by $\mathcal{D}_n \coloneqq \{\Lambda(f) : f \in \mathcal{H}^{(l)}_{M^*,d^*,d,\alpha}\}$ the sieve of neural network discriminators for the standard logistic cdf Λ. □

Assumption S.4 gives a sufficient condition for Assumption S.1, so we use this to derive the rate of convergence of the neural network discriminator. If, in addition, $d^* < 2p$, we have $\delta_n = o_P(n^{-1/4})$; this is easier to satisfy if the underlying dimension of the log likelihood ratio is low, regardless of the dimension of the input.

Proposition S.3 (Rate of convergence of neural network discriminator)**.** *Under Assumptions 2, S.2, and S.4,* $d_\theta(\hat{D}_\theta, D_\theta) = O^*_P(\delta_n)$.

[3] We say that P on $\mathbb{R}^d$ has *subexponential tails* if $\log P(\|X\|_\infty > a) \lesssim -a$ for large a.

Proof. We use Lemma S.2 to bound the bracketing number in Assumption S.1. For now, let us assume that $\mathcal{D}_n$ in Assumption S.4 satisfies the network structure of Lemma S.2; later, we calibrate the constants in reflection of the network structure in Assumption S.4. Since D is nonnegative, we can extend d_θ to accommodate arbitrary functions f_1 and f_2 by $d_\theta(f_1, f_2) := d_\theta(0 \vee f_1, 0 \vee f_2)$. In the notation of Lemma S.2,

$$\begin{aligned}\|\varepsilon^2 F\|_{d_\theta}^2 &= \sup_{D\in\mathcal{D}} d_\theta(D - \varepsilon^2 F/2, D + \varepsilon^2 F/2)^2 \le h_\theta(0, \varepsilon^2 F)^2 + h_\theta(0, \varepsilon^2 F)^2 \\ &= 2\varepsilon^2(P_0 + P_\theta)F = 2\varepsilon^2[2\sigma_0 + (P_0 + P_\theta)\|X\|_\infty] =: B\varepsilon^2.\end{aligned}$$

Since P_0 and P_θ have bounded first moments, $B < \infty$. Replacing ε with $\varepsilon/\sqrt{B}$ yields $\|\frac{\varepsilon^2}{B}F\|_{d_\theta} \le \varepsilon$. Therefore, with Lemma S.2,

$$\log N_{[]}(\varepsilon, \mathcal{D}_n, d_\theta) \le \log N_{[]}(\|\tfrac{\varepsilon^2}{B}F\|_{d_\theta}, \mathcal{D}_n, d_\theta) \le S \log\Big\lceil \tfrac{2B(L+1)(\tilde{U}C)^{L+1}d}{\varepsilon^2}\Big\rceil.$$

Observe that for $0 < \delta \le e^a$,

$$\int_0^\delta \sqrt{1 + a - \log\varepsilon}\, d\varepsilon = \tfrac{\sqrt{\pi}e^a}{2}\operatorname{erfc}\Big(\sqrt{1 + a - \log\delta}\Big) + \delta\sqrt{1 + a - \log\delta} \lesssim \delta\sqrt{1 + a - \log\delta}.$$

Therefore,

$$\begin{aligned}J_{[]}(\delta, \mathcal{D}_n, h_\theta) &\lesssim \int_0^\delta \sqrt{1 + S[\log(2B(L+1)(\tilde{U}C)^{L+1}d) - 2\log\varepsilon]_+}\, d\varepsilon \\ &\lesssim \delta\sqrt{1 + S[\log(2B(L+1)(\tilde{U}C)^{L+1}d) - 2\log\delta]_+} \lesssim \delta\sqrt{1 \vee [SL\log(\tilde{U}C) - S\log\delta]}.\end{aligned}$$

Therefore, if we set

$$\delta_n = O\Big(\sqrt{\tfrac{SL\log(\tilde{U}C) + S\log n}{n}}\Big), \tag{1}$$

$\mathcal{D}_n$ satisfies Assumption S.1 with $\alpha = 1.5$. Now, we must choose S, L, $\tilde{U}$, and C so that this rate is attainable and fast. For the rate to be attainable, we must also have Assumption S.2, for which we need that $\mathcal{D}_{\theta,\delta}$ is nonempty. That is, the sieve $\mathcal{D}_n$ must contain an element in the δ_n-neighborhood of D_θ, i.e., $\inf_{D\in\mathcal{D}_n} d_\theta(D, D_\theta) \lesssim \delta_n$.

Since $\mathcal{D}_n = \Lambda(\mathcal{H}^{(l)})$, we use Bauer and Kohler (2019, Theorem 3) to find the network configuration that attains this inequality. For this, we need to choose "N, η_n, a_n, M_n" in their notation; in doing so, we find "S, L, $\tilde{U}$, C" in our notation. First, we set $N = q$ and $\eta_n = \delta_n^2$. By subexponentiality, we have $\log P_0(\|X\|_\infty > a) + \log P_\theta(\|X\|_\infty > a) \lesssim -a$ for large a. Therefore, we want $a_n \gg -2\log\delta_n$ so that the remainder term in Bauer and Kohler (2019, Theorem 3) is small enough, that

is, $(P_0 + P_\theta)(\|X\|_\infty > a_n) \lesssim \delta_n^2$.[4] We can do this by setting, e.g., $a_n = (-\log \delta_n)^2$. Finally, we want to choose M_n so that $a_n^{N+q+3} M_n^{-p} \sim \delta_n$ since then Bauer and Kohler (2019, Theorem 3) can bound the supremum term that appears below; set $M_n = (\log \delta_n)^{2(N+q+3)/p}/\delta_n^{1/p}$. Let $A \subset [-a_n, a_n]^d$ be the set for which $(P_0 + P_\theta)(A) \le c\eta_n$ in Bauer and Kohler (2019, Theorem 3). Then,

$$
\begin{aligned}
h_\theta(D, D_\theta)^2 &\le \left(\int_{\|x\|_\infty > a_n} + \int_A + \int_{\{\|x\|_\infty \le a_n\}\setminus A}\right)(\sqrt{D} - \sqrt{D_\theta})^2 (p_0 + p_\theta) \\
&\le (P_0 + P_\theta)(\|X\|_\infty > a_n) + (P_0 + P_\theta)(A) + \int_{\{\|x\|_\infty \le a_n\}\setminus A} (\sqrt{D} - \sqrt{D_\theta})^2 (p_0 + p_\theta).
\end{aligned}
$$

The first two terms are bounded by $\delta_n^2 + c\delta_n^2$. For $D = \Lambda(f)$,

$$
\begin{aligned}
\int_{\{\|x\|_\infty \le a_n\}\setminus A} (\sqrt{D} - \sqrt{D_\theta})^2 (p_0 + p_\theta) &= \int_{\{\|x\|_\infty \le a_n\}\setminus A} \left(\sqrt{\Lambda(f)} - \sqrt{\Lambda(\Lambda^{-1} \circ D_\theta)}\right)^2 (p_0 + p_\theta) \\
&\le \tfrac{2}{27} \|f - \Lambda^{-1} \circ D_\theta\|^2_{\infty, \{\|x\|_\infty \le a_n\}\setminus A} = \tfrac{2}{27} \|f - \log \tfrac{p_0}{p_\theta}\|^2_{\infty, \{\|x\|_\infty \le a_n\}\setminus A},
\end{aligned}
$$

since $\sqrt{\Lambda(\cdot)}$ is Lipschitz with constant $1/(3\sqrt{3})$. We may likewise bound $h_\theta(1-D, 1-D_\theta)^2$. By Bauer and Kohler (2019, Theorem 3), $\inf_{f \in \mathcal{H}^{(l)}} \|f - \log \frac{p_0}{p_\theta}\|_{\infty, \{\|x\|_\infty \le a_n\}\setminus A} \lesssim \delta_n$. Thus, we obtain $\inf_{D \in \mathcal{D}_n} d_\theta(D, D_\theta) \lesssim \delta_n$.

These configurations can be translated into our constants as $S = O(dd^* M_* K^l) \sim M_*$, $\tilde{U} = M_* \vee (4d^*) \vee K \sim M_*$, $C = \alpha$, and $L = 2 + 3l = O(1)$, where Bauer and Kohler (2019, Theorem 3) define

$$
\begin{aligned}
M_* &= \binom{d^* + N}{d^*} (N+1)(M_n + 1)^{d^*} \sim M_n^{d^*} = \frac{(\log \delta_n)^{2d^*(N+q+3)/p}}{\delta_n^{d^*/p}}, \\
\alpha &= \frac{M_n^{d^* + p(2N+3)+1}}{\eta_n} \log n = \frac{(\log \delta_n)^{2(N+q+3)[d^* + p(2N+3)+1]/p}}{\delta_n^{2+[d^* + p(2N+3)+1]/p}} \log n.
\end{aligned}
$$

With these, (1) becomes $\delta_n^2 \sim M_* \frac{\log(M_*\alpha) + \log n}{n} \sim [(\log n)^{\frac{p + 2d^*(N+q+3)}{p}}/n]^{\frac{p}{2p+d^*}}$. The result follows by substituting $N = q$ and invoking Theorem S.1. ■

The following lemma bounds the bracketing number of a (possibly sparse) neural network with bounded weights and Lipschitz activation functions. The notation of the neural network is defined as follows. Denote the hidden-layer activation function by $\sigma : \mathbb{R} \to \mathbb{R}$ and the output activation function by $\Lambda : \mathbb{R} \to \mathbb{R}$. Let L be the number of hidden and output layers. Let $w_{\ell ij}$ be the weight for the ith node in the $(\ell + 1)$th

[4] If we set $a_n \sim -2 \log \delta_n$, we can only say $(P_0 + P_\theta)(\|X\|_\infty > a_n) \lesssim \delta_n^c$ for some c.

layer on the jth node in the ℓth layer; for example, the input to the second node in the first layer is $w_{021}x_1 + \cdots + w_{02U}x_U$, where $X = (x_1, \ldots, x_U)$ is the input to the network. Let $w_{\ell i} = (w_{\ell i1}, \ldots, w_{\ell iU})^\top$ be the column vector of weights for the ith node in the $(\ell+1)$th layer. Let $w_\ell = (w_{\ell 1}, \ldots, w_{\ell U})$ be the matrix with columns $w_{\ell i}$; note that for $\ell = L$, w_L is just a column vector as there is only one output. Let w be the vector of all parameters. Then, the discriminator is given by[5]

$$D(X; w) = \Lambda(w_L^\top \sigma(w_{L-1}^\top \sigma(\cdots w_1^\top \sigma(w_0^\top X)))),$$

where $\sigma(\cdot)$ for a vector argument is elementwise application.

Lemma S.2 (Bracketing number of neural network with bounded weights)**.** *Let $\mathcal{F}$ be a class of neural networks defined as above. Denote the total number of nonzero weights by S and the maximum number of nonzero weights in each node (except for the first layer taking inputs) by $\tilde{U}$.*[6] *Assume that σ and Λ are Lipschitz with constant 1 and $\|w\|_\infty \le C$ for some C. Assume innocuously that $\tilde{U}C \ge 2$ and let $\sigma_0 \coloneqq |\sigma(0)|$. Define an envelope $F : \mathbb{R}^d \to \mathbb{R}$ by $F(x) \coloneqq \sigma_0 + \|x\|_\infty$. Then, for every premetric $d_{\mathcal{F}}$ and $\|f\|_{d_{\mathcal{F}}} \coloneqq \sup_{g\in\mathcal{F}} d_{\mathcal{F}}(g - f/2, g + f/2)$,*

$$N_{[]}(\|\varepsilon F\|_{d_{\mathcal{F}}}, \mathcal{F}, d_{\mathcal{F}}) \le \left\lceil \frac{2(L+1)(\tilde{U}C)^{L+1}d}{\varepsilon} \right\rceil^S.$$

For a fully connected network, $\tilde{U} = U$ and $S = (LU+1)U + (d-U)U$. For a hierarchical network in Bauer and Kohler (2019), $S = O(\tilde{U}^{(L+4)/3}d)$.

Proof. The neural network is expressed as $f(x; w) = \Lambda(w_L^\top \sigma(w_{L-1}^\top \sigma(\cdots w_1^\top \sigma(w_0^\top x))))$. We can bound the outputs of the ℓth layer by

$$\begin{aligned}
\|\sigma(w_{\ell-1}^\top \sigma(\cdots))\|_\infty &\le \sigma_0 + \|w_{\ell-1}^\top \sigma(\cdots)\|_\infty \le \sigma_0 + \tilde{U}C\|\sigma(\cdots)\|_\infty \\
&\le [1 + \tilde{U}C + \cdots + (\tilde{U}C)^{\ell-1}]\sigma_0 + \tilde{U}^{\ell-1}C^\ell d\|x\|_\infty \\
&\le \tilde{U}^{\ell-1}C^\ell(\tilde{U}\sigma_0 + d\|x\|_\infty) \le (\tilde{U}C)^\ell d(\sigma_0 + \|x\|_\infty),
\end{aligned}$$

where the fourth inequality holds for $\tilde{U}C \ge 2$. For two sets of weights, w and $\tilde{w}$,

$$|f(x; w) - f(x; \tilde{w})| \le \tilde{U}\|w_L - \tilde{w}_L\|_\infty(\|\sigma(w_{L-1}^\top \sigma(\cdots))\|_\infty \vee \|\sigma(\tilde{w}_{L-1}^\top \sigma(\cdots))\|_\infty)$$

[5] If we include a constant input and a constant node (also known as the "bias" term), it is assumed to be already incorporated in X and w.

[6] The number of nonzero elements in each row of each matrix w_ℓ, $\ell \ge 1$, is bounded by $\tilde{U}$.

$$
\begin{aligned}
&\quad + \tilde{U}C\|\sigma(w_{L-1}^\top \sigma(\cdots)) - \sigma(\tilde{w}_{L-1}^\top \sigma(\cdots))\|_\infty \\
&\le \tilde{U}^{L+1} C^L d \|w_L - \tilde{w}_L\|_\infty (\sigma_0 + \|x\|_\infty) + \cdots \\
&+ \tilde{U}^{L+1} C^L d \|w_1 - \tilde{w}_1\|_\infty (\sigma_0 + \|x\|_\infty) + \tilde{U}^L C^L d \|w_0 - \tilde{w}_0\|_\infty \|x\|_\infty \\
&\le (L+1)\tilde{U}^{L+1} C^L d \|w - \tilde{w}\|_\infty (\sigma_0 + \|x\|_\infty).
\end{aligned}
$$

Let $A \coloneqq (L+1)\tilde{U}^{L+1}C^L d$. Partitioning the weight space $[-C, C]^S$ into cubes of length $2\varepsilon/A$ creates $\lceil CA/\varepsilon \rceil^S$ cubes. Hence, the covering number is bounded as $N(\varepsilon, [-C, C]^S, \|\cdot\|_\infty) \le \lceil CA/\varepsilon \rceil^S$. The bound on the bracketing number then follows from van der Vaart and Wellner (1996, Theorem 2.7.11), observing that the proof thereof works for a premetric with modification of $2\varepsilon\|F\|$ to $\|2\varepsilon F\|_{d_{\mathcal{F}}}$.

For a fully connected network, the number of all weights is dU (weights for the first layer) plus $(L-1)U^2$ (weights for the remaining hidden layers) plus U (weights in the output layer), summing to $(LU+1)U+(d-U)U$.[7] For a network $\mathcal{H}^{(0)}$ in Bauer and Kohler (2019) (in their notation), the number of all weights is $A^{(0)} \coloneqq d(4d^*M_*) + 4d^*M_* + M_* = 4(1+d)d^*M_* + M_*$. For $\mathcal{H}^{(1)}$, $A^{(1)} \coloneqq A^{(0)}K + K(4d^*M_*) + 4d^*M_* + M_* = A^{(0)}K + 4(1+K)d^*M_* + M_*$. For $\mathcal{H}^{(l)}$, $A^{(l)} \coloneqq A^{(l-1)}K + 4(1+K)d^*M_* + M_* = A^{(0)}K^l + \sum_{j=0}^{l-1} K^j[4(1+K)d^*M_* + M_*] = 4d^*M_*[(1+d)K^l + \frac{1-K^l}{1-K}(1+K)] + M_*\frac{1-K^{l+1}}{1-K} = O(dd^*M_*K^l)$. Then use $L = 2+3l$ and $\tilde{U} = M_* \vee (4d^*) \vee K$. ■

Remark. Lemma S.2 assumes a Lipschitz property for the activation and output functions, which accommodates ReLU, softplus, and sigmoid, but not perceptron.

S.3 SUPPORTING LEMMAS FOR THE MAIN TEXT

The following lemma shows local convergence of the loss needed for Theorem 3.

Lemma S.3 (Asymptotic distribution of objective function)**.** *Under Assumptions 2 and 5, for every compact $K \subset \Theta$, uniformly in $h \in K$,*

$$
\begin{aligned}
n[\mathbb{M}_{\theta_0+h/\sqrt{n}}(D_{\theta_0+h/\sqrt{n}}) - \mathbb{M}_{\theta_0}(D_{\theta_0})] &= -\sqrt{n}\mathbb{P}_0 h^\top \dot{\ell}_{\theta_0} + \sqrt{n}(\mathbb{P}_0 + \mathbb{P}_{\theta_0+h/\sqrt{n}}) D_{\theta_0+h/\sqrt{n}} h^\top \dot{\ell}_{\theta_0} \\
&\quad + n[(\mathbb{P}_{\theta_0+h/\sqrt{n}} - P_{\theta_0+h/\sqrt{n}}) - (\mathbb{P}_{\theta_0} - P_{\theta_0})] \log(1 - D_{\theta_0}) + \frac{h^\top \tilde{I}_{\theta_0} h}{4} + o_P(1).
\end{aligned}
$$

[7] If the network has a bias term, the actual variable weights are slightly fewer, but it does not change the order.

Proof. Let $\theta \coloneqq \theta_0 + h/\sqrt{n}$, $W \coloneqq \sqrt{D_\theta/D_{\theta_0}} - 1$, $\tilde{W} \coloneqq \sqrt{p_{\theta_0}/p_\theta} - 1$. Observe that

$$n[\mathbb{M}_\theta(D_\theta) - \mathbb{M}_{\theta_0}(D_{\theta_0})] = n(\mathbb{P}_0 + \mathbb{P}_\theta)\log\tfrac{D_\theta}{D_{\theta_0}} - n\mathbb{P}_\theta \log\tfrac{p_{\theta_0}}{p_\theta} + n(\mathbb{P}_\theta - \mathbb{P}_{\theta_0})\log(1 - D_{\theta_0}).$$

We examine each term separately. By Assumption 5,

$$\begin{aligned} n(P_\theta - P_{\theta_0})\log(1 - D_{\theta_0}) &= n\int (\sqrt{p_\theta} + \sqrt{p_{\theta_0}})(\sqrt{p_\theta} - \sqrt{p_{\theta_0}})\log(1 - D_{\theta_0}) \\ &= \int \Big(\sqrt{n}h^\top \dot{\ell}_{\theta_0} + \tfrac{h^\top \ddot{\ell}_{\theta_0} h}{2} + \tfrac{h^\top \dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}^\top h}{2}\Big) p_{\theta_0}\log(1 - D_{\theta_0}) + o(1). \end{aligned}$$

The first term is zero since $M_\theta(D_\theta) - M_{\theta_0}(D_{\theta_0}) \geq 0$ and $M_\theta(D_\theta) - M_{\theta_0}(D_{\theta_0}) = 2\int D_{\theta_0}(\sqrt{p_\theta} - \sqrt{p_{\theta_0}})^2 + o(h(\theta,\theta_0)^2) + (P_\theta - P_{\theta_0})\log(1 - D_{\theta_0})$.[8] Therefore, $n(P_\theta - P_{\theta_0})\log(1 - D_{\theta_0}) = \frac{1}{2}P_{\theta_0}(h^\top \ddot{\ell}_{\theta_0} h + h^\top \dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}^\top h)\log(1 - D_{\theta_0}) + o(1)$.

Using $\log x = 2(\sqrt{x} - 1) - (\sqrt{x} - 1)^2 + (\sqrt{x} - 1)^2 R(\sqrt{x} - 1)$ for $R(x) = O(x)$,

$$n(\mathbb{P}_0 + \mathbb{P}_\theta)\log\tfrac{D_\theta}{D_{\theta_0}} = 2n(\mathbb{P}_0 + \mathbb{P}_\theta)W - n(\mathbb{P}_0 + \mathbb{P}_\theta)W^2 + n(\mathbb{P}_0 + \mathbb{P}_\theta)W^2 R(W_n).$$

Let $\breve{I}_{\theta_0} \coloneqq 2P_{\theta_0} D_{\theta_0}\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}^\top$. Observe that

$$(P_0 + P_\theta)\Big(\sqrt{n}W + \tfrac{h^\top \dot{\ell}_{\theta_0}}{2}(1 - D_\theta)\Big)^2 = n\int \Big[\sqrt{p_0 + p_{\theta_0}} - \sqrt{p_0 + p_\theta} + \tfrac{h^\top \dot{\ell}_{\theta_0}}{2\sqrt{n}}\sqrt{(1 - D_\theta)p_\theta}\Big]^2,$$

which is $o(\|h\|^2/n)$ by Lemma S.6 and Assumption 5. Thus, the RHS converges to zero uniformly over every compact $K \subset \Theta$. We draw two observations: (i) the mean and variance of $(\sqrt{n}W + (1 - D_\theta)h^\top \dot{\ell}_{\theta_0}/2)(X_i)$, $X_i \sim (P_0 + P_{\theta_n})/2$, converge to zero and so does the variance of $\sqrt{n}(\mathbb{P}_0 + \mathbb{P}_\theta)(\sqrt{n}W + (1 - D_\theta)h^\top \dot{\ell}_{\theta_0}/2)$ under Assumption 2;[9] (ii) $(P_0 + P_\theta)|nW^2 - (1 - D_\theta)^2(h^\top \dot{\ell}_{\theta_0}/2)^2| \to 0$, so $n(\mathbb{P}_0 + \mathbb{P}_\theta)W^2 = (\mathbb{P}_0 + \mathbb{P}_\theta)(1 - D_\theta)^2(h^\top \dot{\ell}_{\theta_0}/2)^2 + o_P(1) \to h^\top I_{\theta_0} h/4 - h^\top \breve{I}_{\theta_0} h/8$. Next,

$$\begin{aligned} n(P_0 + P_\theta)W &= -\tfrac{n}{2}h(p_0 + p_{\theta_0}, p_0 + p_\theta)^2 \longrightarrow -\tfrac{h^\top I_{\theta_0} h}{8} + \tfrac{h^\top \breve{I}_{\theta_0} h}{16}, \\ \sqrt{n}(P_0 + P_\theta)(1 - D_\theta)\tfrac{h^\top \dot{\ell}_{\theta_0}}{2} &= \sqrt{n}P_\theta \tfrac{h^\top \dot{\ell}_{\theta_0}}{2} = \sqrt{n}(P_\theta - P_{\theta_0})\tfrac{h^\top \dot{\ell}_{\theta_0}}{2} \to \tfrac{h^\top I_{\theta_0} h}{2}. \end{aligned}$$

This implies that the mean of $\sqrt{n}(\mathbb{P}_0 + \mathbb{P}_\theta)(\sqrt{n}W + (1 - D_\theta)h^\top \dot{\ell}_{\theta_0}/2)$ converges to $3h^\top I_{\theta_0} h/8 + h^\top \breve{I}_{\theta_0} h/16$. Combining with (i), we find

$$n(\mathbb{P}_0 + \mathbb{P}_\theta)W = -\sqrt{n}(\mathbb{P}_0 + \mathbb{P}_\theta)(1 - D_\theta)\tfrac{h^\top \dot{\ell}_{\theta_0}}{2} + \tfrac{3h^\top I_{\theta_0} h}{8} + \tfrac{h^\top \breve{I}_{\theta_0} h}{16} + o_P(1).$$

[8] The term $P_{\theta_0} h^\top \dot{\ell}_{\theta_0}\log(1 - D_{\theta_0})$ is the only term that is linear in $h = h(\theta,\theta_0)$, so if it is not zero, then $M_\theta(D_\theta) - M_{\theta_0}(D_{\theta_0}) \geq 0$ is violated.

[9] This does not imply that the mean of $\sqrt{n}(\mathbb{P}_0 + \mathbb{P}_\theta)(\sqrt{n}W + (1 - D_\theta)h^\top \dot{\ell}_{\theta_0}/2)$ converges to zero.

The remainder term $n(\mathbb{P}_0+\mathbb{P}_\theta)W^2R(W_n)$ vanishes by the same logic as van der Vaart (1998, Theorem 7.2).

Next, observe that $n\mathbb{P}_\theta \log \frac{p_{\theta_0}}{p_\theta} = 2n\mathbb{P}_\theta\tilde{W} - n\mathbb{P}_\theta\tilde{W}^2 + n\mathbb{P}_\theta\tilde{W}^2R(\tilde{W})$ and

$$P_\theta\Big(\sqrt{n}\tilde{W} + \tfrac{h^\top\dot{\ell}_{\theta_0}}{2}\Big)^2 = n\int\Big[\sqrt{p_{\theta_0}} - \sqrt{p_\theta} + \tfrac{h^\top\dot{\ell}_\theta}{2\sqrt{n}}\sqrt{p_\theta}\Big]^2 = o\Big(\tfrac{\|h\|^2}{n}\Big).$$

Again, (i) the mean and variance of $(\sqrt{n}\tilde{W} + h^\top\dot{\ell}_{\theta_0}/2)(X_i)$, $X_i \sim P_\theta$, converge to zero and so does the variance of $\sqrt{n}\mathbb{P}_\theta(\sqrt{n}\tilde{W} + h^\top\dot{\ell}_{\theta_0}/2)$ under Assumption 2; (ii) $P_\theta|n\tilde{W}^2 - (h^\top\dot{\ell}_{\theta_0}/2)^2| \to 0$, so $n\mathbb{P}_\theta\tilde{W}^2 \to P_\theta(h^\top\dot{\ell}_{\theta_0}/2)^2 \to h^\top I_{\theta_0}h/4$. Next, $nP_\theta\tilde{W} = -nh(\theta,\theta_0)^2/2 \to -h^\top I_{\theta_0}h/8$ and $\sqrt{n}P_\theta h^\top\dot{\ell}_{\theta_0}/2 \to h^\top I_{\theta_0}h/2$. This implies that the mean of $\sqrt{n}\mathbb{P}_\theta(\sqrt{n}\tilde{W} + h^\top\dot{\ell}_{\theta_0}/2)$ converges to $3h^\top I_{\theta_0}h/8$. Thus, we find

$$n\mathbb{P}_\theta\tilde{W} = -\sqrt{n}\mathbb{P}_\theta\tfrac{h^\top\dot{\ell}_{\theta_0}}{2} + \tfrac{3h^\top I_{\theta_0}h}{8} + o_P(1).$$

We may once again ignore the remainder term $n\mathbb{P}_\theta\tilde{W}^2R(\tilde{W})$. Altogether, with $\tilde{I}_{\theta_0}$ defined in Assumption 5,

$$\begin{aligned} n[\mathbb{M}_\theta(D_\theta) - \mathbb{M}_{\theta_0}(D_{\theta_0})] = &-\sqrt{n}\mathbb{P}_0h^\top\dot{\ell}_{\theta_0} + \sqrt{n}(\mathbb{P}_0+\mathbb{P}_\theta)D_\theta h^\top\dot{\ell}_{\theta_0} + \tfrac{h^\top\tilde{I}_{\theta_0}h}{4} \\ &+ n[(\mathbb{P}_\theta - \mathbb{P}_{\theta_0}) - (P_\theta - P_{\theta_0})]\log(1-D_{\theta_0}) + o_P(1). \end{aligned}$$

■

The Bernstein "norm" of a function f is defined as $\|f\|_{P,B} := \sqrt{2P(e^{|f|} - 1 - |f|)}$; this induces a premetric without the triangle inequality (van der Vaart and Wellner, 1996, p. 324).[10] The next lemma bounds the Bernstein "norm" of a log likelihood ratio by the Hellinger distance without assuming a bounded likelihood ratio.

Lemma S.4 (Bernstein "norm" of log likelihood ratio; Kaji and Ročková, 2022, Lemma 2.1 (iv))**.** *For any pair of probability measures P and P_0 such that $P_0(p_0/p) < \infty$,*

$$\Big\|\tfrac{1}{2}\log\tfrac{p}{p_0}\Big\|_{P_0,B}^2 \le 2h(p,p_0)^2\Big[1 + P_0\Big(\tfrac{p_0}{p}\;\Big|\;\tfrac{p_0}{p} \ge \tfrac{25}{16}\Big)\Big],$$

where $P_0(p_0/p \mid p_0/p \ge a) = 0$ if $P_0(p_0/p \ge a) = 0$.

[10] A *premetric* on a class of functions $\mathcal{F}$ is a function $d : \mathcal{F}\times\mathcal{F} \to \mathbb{R}$ that satisfies $d(f,f) = 0$ and $d(f,g) = d(g,f) \ge 0$ for every $f, g \in \mathcal{F}$.

Remark. Similarly, we have

$$\left\|\tfrac{1}{2}\log\tfrac{D}{D_\theta}\right\|_{P_0,B}^2 \le 2h_\theta(D,D_\theta)^2\Big[1+P_0\Big(\tfrac{D_\theta}{D}\;\Big|\;\tfrac{D_\theta}{D}\ge\tfrac{25}{16}\Big)\Big],$$
$$\left\|\tfrac{1}{2}\log\tfrac{1-D}{1-D_\theta}\right\|_{P_\theta,B}^2 \le 2h_\theta(1-D,1-D_\theta)^2\Big[1+P_\theta\Big(\tfrac{1-D_\theta}{1-D}\;\Big|\;\tfrac{1-D_\theta}{1-D}\ge\tfrac{25}{16}\Big)\Big].$$

Lemma S.5 (Bernstein "norm" of log discriminator ratio)**.** *For every* $\theta_1,\theta_2\in\Theta$*,*

$$\left\|\log\tfrac{D_{\theta_1}}{D_{\theta_2}}\right\|_{P_0,B}^2 \le 8h(\theta_1,\theta_2)^2, \qquad \left\|\log\tfrac{(1-D_{\theta_1})\circ T_{\theta_1}}{(1-D_{\theta_2})\circ T_{\theta_2}}\right\|_{\tilde P_0,B}^2 \le 8\tilde h(\theta_1,\theta_2)^2.$$

Proof. Since $e^{|x|}-1-|x|\le 2(e^{x/2}-1)^2$ for $x\ge 0$,

$$\begin{aligned}\left\|\log\tfrac{D_{\theta_1}}{D_{\theta_2}}\right\|_{P_0,B}^2 &\le 4P_0\Big(\sqrt{\tfrac{D_{\theta_1}}{D_{\theta_2}}}-1\Big)^2\mathbb{1}\{D_{\theta_1}\ge D_{\theta_2}\}+4P_0\Big(\sqrt{\tfrac{D_{\theta_2}}{D_{\theta_1}}}-1\Big)^2\mathbb{1}\{D_{\theta_1}<D_{\theta_2}\}\\ &\le 4P_0\Big(\sqrt{\tfrac{p_0+p_{\theta_2}}{p_0+p_{\theta_1}}}-1\Big)^2+4P_0\Big(\sqrt{\tfrac{p_0+p_{\theta_1}}{p_0+p_{\theta_2}}}-1\Big)^2\\ &\le 8\int(\sqrt{p_0+p_{\theta_1}}-\sqrt{p_0+p_{\theta_2}})^2\le 8\int(\sqrt{p_{\theta_1}}-\sqrt{p_{\theta_2}})^2\le 8h(\theta_1,\theta_2)^2.\end{aligned}$$

Similarly,

$$\left\|\log\tfrac{(1-D_{\theta_1})\circ T_{\theta_1}}{(1-D_{\theta_2})\circ T_{\theta_2}}\right\|_{\tilde P_0,B}^2 \le 4\tilde P_0\Big(\sqrt{\tfrac{(1-D_{\theta_1})\circ T_{\theta_1}}{(1-D_{\theta_2})\circ T_{\theta_2}}}-1\Big)^2+4\tilde P_0\Big(\sqrt{\tfrac{(1-D_{\theta_2})\circ T_{\theta_2}}{(1-D_{\theta_1})\circ T_{\theta_1}}}-1\Big)^2\le 8\tilde h(\theta_1,\theta_2)^2$$

since

$$\begin{aligned}\tilde P_0\Big(\sqrt{\tfrac{(1-D_{\theta_1})\circ T_{\theta_1}}{(1-D_{\theta_2})\circ T_{\theta_2}}}-1\Big)^2 &\le \tilde P_0\Big(\tfrac{1}{\sqrt{(1-D_{\theta_2})\circ T_{\theta_2}}}-\tfrac{1}{\sqrt{(1-D_{\theta_1})\circ T_{\theta_1}}}\Big)^2\\ &\le \tilde P_0\Big(\sqrt{\tfrac{p_0}{p_{\theta_2}}\circ T_{\theta_2}}-\sqrt{\tfrac{p_0}{p_{\theta_1}}\circ T_{\theta_1}}\Big)^2=\tilde h(\theta_1,\theta_2)^2.\end{aligned}$$ ■

Lemma S.6 (Hellinger distance of sums of densities)**.** *For arbitrary densities* p*,* p_0*,* p_1*,*

$$h(p+p_0,p+p_1)^2=\int\tfrac{p_0}{p+p_0}(\sqrt{p_0}-\sqrt{p_1})^2+o(h(p_0,p_1)^2).$$

Proof. Since $\sqrt{p+x^2}$ is uniformly differentiable in x with derivative $x/\sqrt{p+x^2}$, the result follows by expanding $\sqrt{p_1}$ around $\sqrt{p_0}$. ■

REFERENCES

Bauer, B. and M. Kohler (2019): "On Deep Learning as a Remedy for the Curse of Dimensionality in Nonparametric Regression," *Annals of Statistics*, 47, 2261–2285.

Kaji, T. and V. Ročková (2022): "Supplementary Material for 'Metropolis–Hastings via Classification'," *Journal of the American Statistical Association*, forthcoming.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge: Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer.

——— (2011): "A Local Maximal Inequality under Uniform Entropy," *Electronic Journal of Statistics*, 5, 192–203.