Adversarial Estimators

Jonas Metzger
Stanford University

June 19, 2022

Abstract

We develop an asymptotic theory of adversarial estimators ('A-estimators'). They generalize maximum-likelihood-type estimators ('M-estimators') as their average objective is maximized by some parameters and minimized by others. This class subsumes the continuous-updating Generalized Method of Moments, Generative Adversarial Networks and more recent proposals in machine learning and econometrics. In these examples, researchers state which aspects of the problem may *in principle* be used for estimation, and an adversary learns *how to emphasize* them optimally. We derive the convergence rates of A-estimators under pointwise and partial identification, and the normality of functionals of their parameters. Unknown functions may be approximated via sieves such as deep neural networks, for which we provide simplified low-level conditions. As a corollary, we obtain the normality of neural-net M-estimators, overcoming technical issues previously identified by the literature. Our theory yields novel results about a variety of A-estimators, providing intuition and formal justification for their success in recent applications.

1   Introduction

Although it is not always obvious, nearly all population parameters that are estimated in econometrics and machine learning can be written as the solution of so-called saddle-point or adversarial objectives of the form:

$$\theta_* = \arg\min_{\theta \in \Theta} \max_{\lambda \in \Lambda} \mathbb{E} l(\theta, \lambda, Y) \tag{1.1}$$

where l is a known loss function, Y is a random variable and Θ, Λ are parameter spaces, containing the unknown parameter of interest θ_* and nuisance λ. We examine the natural estimator $\widehat{\theta}_n$ that approximately solves the empirical Nash condition:

$$\mathbb{E}_n l(\widehat{\theta}_n, \widehat{\lambda}_n, Y) \leq \inf_{\theta \in \Theta_n} \mathbb{E}_n l(\theta, \widehat{\lambda}_n, Y) + \widetilde{\eta}_n \tag{1.2}$$

$$\mathbb{E}_n l(\widehat{\theta}_n, \widehat{\lambda}_n, Y) \geq \sup_{\lambda \in \Lambda_n} \mathbb{E}_n l(\widehat{\theta}_n, \lambda, Y) - \eta_n \tag{1.3}$$

which replaces the expectation $\mathbb{E}$ of the population objective 1.1 with the average of n iid samples, $\mathbb{E}_n$. We search for the estimators over so-called *sieve spaces* $\widehat{\theta}_n, \widehat{\lambda}_n \in \Theta_n, \Lambda_n$ (Grenander [1981]), which approximate the full parameter spaces $\Theta_n, \Lambda_n \subset \Theta, \Lambda$ and grow with the sample size n. These could be neural networks for example, growing in depth and width. The sequences $\widetilde{\eta}_n$, $\eta_n = o_{\mathbb{P}}(1)$ accommodate numerical procedures which only yield approximate Nash equilibria. This class of A-estimators (A for *adversarial*) strictly generalizes so-called M-estimators (M for *maximum likelihood-type*), which are obtained by fixing Λ to be singleton.

A-Estimators have become a workhorse of econometrics and causal inference long before the advent of deep learning. Hansen et al. [1996]'s continuous-updating Generalized Methods of Moments (GMM), which looks for θ satisfying $\mathbb{E}[m(\theta, Y)] = 0$ for some known function $m(\theta, Y)$, can be written as:

$$\inf_\theta \mathbb{E}_n\left[m(\theta, Y)\right] \mathbb{E}_n\left[m(\theta, Y)m(\theta, Y)'\right]^{-1} \mathbb{E}_n\left[m(\theta, Y)\right]$$
$$= \inf_\theta \sup_\lambda \mathbb{E}_n\left[m(\theta, Y)'\lambda - (m(\theta, Y)'\lambda)^2/4\right]$$

and is therefore an A-estimator, but not an M-estimator. In statistics, an earlier example consists of the Empirical Likelihood (EL) approach pioneered in Cosslett [1981], Owen [1988, 1990], Qin and Lawless [1994]. Subsequently, EL was unified with GMM into the Generalized Empirical Likelihood (GEL) framework (Newey and Smith [2004]), also subsuming the exponential-tilting estimator (Imbens et al. [1998]), for example. All GEL estimators are A-estimators, but their adversarial formulation was rarely salient. However, some of their benefits may be owed directly to their adversarial objective: the adversary λ automatically detects which moment violations are most informative at a given parameter guess, adaptively guiding the estimation towards an efficient solution. This contrasts with earlier estimators which weighted the moments in a way that depended on choices of the researcher: the weights of Pearson's Method-of-Moments were manually set by the researcher (implicitly), resulting in inefficient root-n asymptotics. Two-step GMM (Hansen [1982]) required choosing a first-step estimator to compute the weights, yielding inefficient higher-order asymptotics (see Newey and Smith [2004]). Formally, the optimal weights are nuisance parameters, and as we will see in Section 3.2, estimating them via an adversary ensures that $\widehat{\theta}_n$ is robust to estimation errors in these nuisance parameters.

A key invention which put a spotlight on adversarial objectives in recent years were *Generative Adversarial Networks*, or GANs (Goodfellow et al. [2014]). They search for a generative model $Y \sim \mathbb{P}_\theta$ for which no adversary $\lambda(Y) \in (0,1)$ (called 'critic' or 'discriminator') could tell apart the generated data from n real samples:

$$\inf_{\mathbb{P}_\theta} \sup_{\lambda(\cdot) \in (0,1)} \mathbb{E}_{\mathbb{P}_\theta} \log \lambda(Y) + \mathbb{E}_n \log(1 - \lambda(Y))$$

The objective contains the log-likelihood of a binary classifier $\lambda(Y)$ discriminating between an equal number of real and generated samples. As we show in Section 2.1, this directly measures the Jensen-Shannon divergence between $\mathbb{P}_\theta$ and $\mathbb{P}_n$. As of today, versions of this objective are key to state-of-the-art image generation, see e.g. Jabbar et al. [2022] for a recent survey. An analogy to human-generated images makes this unsurprising: it is much easier to tell apart a photo from an image drawn by a human, than it is to draw a realistic image, or to define what makes a drawing realistic. This intuition motivates the objective: train the generator until its critic has nothing more to criticize. The ingenuity is that the researcher need not define a meaningful measure of 'realism' of a piece of data anymore. Instead, this measure is *learned* by the adversary. It is clear that the utility of this idea extends beyond image generation: in Imitation Learning, a sub-field of Robotics, it has been used to teach human behavior to artificial agents without requiring hand-crafted measures of 'humanness' (Ho and Ermon [2016]). In Econometrics, where new causal inference methods can usually only be benchmarked on simulated data sets, Athey et al. [2019] used the objective to limit the impact of researcher's subjective choices by requiring simulations to be indistinguishable from real data. Kaji et al. [2020] proposed to use the objective to estimate structural economic models which produce realistic data beyond the set of features that would otherwise be manually specified by the researcher.

More generally, other adversarial objectives have proven useful beyond fitting models to data. In Reinforcement Learning, a sub-field of Robotics where agents independently discover strategies to reach predefined goals without copying prior examples, Dai et al. [2018] proposed an A-estimator in which the adversary detects and penalizes any systematic deviation from optimal behavior. Cotter et al. [2019] proposed an estimator which extends a standard ML objective by an adversary imposing fairness

constraints across sub-populations. More recently, research in econometrics established A-estimators as a natural framework for integrating machine learning methods into causal inference, where quantities of interest are frequently identified by a continuum of restrictions. Chernozhukov et al. [2020] propose to estimate Riesz representers of causal parameters directly, via an adversary enforcing the restrictions identifying the Riesz representer. Estimating Riesz representers is key to obtaining well-behaved estimates of causal parameters in the presence of nuisance functions, and can also be useful for estimating asymptotic variances, e.g. Chen et al. [2019]. Another line of research develops novel adversarial objectives to estimate causal parameters from *conditional* moment restrictions, which naturally arise from causal assumptions (e.g. the instrumental variable setting), and are usually more informative than any finite set of unconditional moment restrictions. In this line of research, the adversary can be viewed as adaptively finding the unconditional moment restriction which is most violated at the current parameter guess, among infinitely many which are implied by the conditional moment restriction. The key works are Lewis and Syrgkanis [2018], Dikkala et al. [2020] and Bennett et al. [2019b], Bennett and Kallus [2020]. Metzger [2022] propose a semi-parametrically efficient generalization of GEL to the conditional case via adversarial networks, containing Bennett and Kallus [2020] as a special case.

In summary, a recurring theme of adversarial objectives is that instead of manually defining which specific features of the data are important for a model to capture, the researcher's role is restricted to stating a general principle which should be satisfied by all features of the correct model, and the adversary *adaptively* focuses the estimation on the model's features which violate this principle the most. Over the course of the paper, we will encounter further interesting connections between various A-estimators, such as their Neyman orthogonality, their information-theoretic foundation via f-Divergences, and their ties to Lagrangian Duality.

Despite their popularity, we are not aware of a unified statistical theory of A-estimators. For some individual estimators, consistency (Bennett et al. [2019b]) and convergence rate results (Dikkala et al. [2020], Singh et al. [2018], Liang [2021], Belomestny et al. [2021]) were obtained, but normality results are limited to parametric θ, either in Kernel settings (Bennett and Kallus [2020]) or leaving high-level assumptions about neural networks unverified (Kaji et al. [2020]). This can be attributed to two main

4

obstacles: the theory of M-estimation does not apply to A-estimators, and the arguments from which the former is built up are insufficient to e.g. obtain the required uniform convergence of the adversary. The second issue is that adversarial objectives are most popular in the context of (deep) neural networks, whose statistical analysis (particularly their asymptotic normality) is complicated, e.g. due to their non-convex sieve space. Even in M-estimation settings, it was not clear whether known, high-level conditions for normality could be verified for neural networks (cf. the Conclusion of Shen et al. [2019]). We therefore make three separate contributions:

1. We characterize the general class of A-estimators, and show that a wide range of estimators proposed in econometrics and machine learning fall into this class. We point out desirable characteristics shared between A-estimators, which help explain their recent success in practice.

2. We develop a unified statistical theory of A-estimators, yielding their consistency, convergence rates (both under point- and partial identification), and asymptotic normality of functionals of their parameters. We provide high-level conditions for arbitrary sieves, as well as low-level conditions for semiparametric settings with neural networks, to simplify verification in practice.

3. We extend the theory of neural network M-estimators (as a special case). Our convergence rates hold uniformly over families of losses, allow more general losses than Farrell et al. [2018] and attain a reduced curse-of-dimensionality which Nakada and Imaizumi [2020], Bauer et al. [2019] observed in regression settings with lower-dimensional structures. To the best of our knowledge, we provide the first normality result for functionals of deep neural networks which does not rely on Neyman-orthogonality or unverified high-level assumptions.

The remainder of the paper is structured as follows. In Section 2, we review five different A-estimators proposed in the econometrics and machine learning literatures. We present our general statistical theory of A-estimators in Section 3 and apply it in Section 4 to derive novel results about the examples of Section 2. We conclude by recapping the similar role adversaries play across all examples, providing intuition which types of problems may generally benefit from adversarial formulations. Appendix C and Online Appendix D contain the proofs omitted in Sections 3 and 4, respectively.

2 Examples

2.1 Minimum f-Divergence

A powerful class of estimation objectives asymptotically minimize an f-divergence $D_f(\mathbb{P}_\theta \| \mathbb{P})$ between the distribution of the data $Y \sim \mathbb{P} = \mathbb{P}_{\theta_*}$ and the distribution of some model $\mathbb{P}_\theta, \theta \in \Theta_n$ with support $\mathcal{Y}$. This class, introduced by Nowozin et al. [2016], subsumes GANs (Goodfellow et al. [2014]), and many follow ups such as Mao et al. [2017], Tao et al. [2018]. For a continuous, proper convex function $f : \mathbb{R} \mapsto \mathbb{R}$ satisfying $f(1) = 0$, the f-divergence is defined as $D_f(\mathbb{P}_\theta \| \mathbb{P}) = \mathbb{E}_\mathbb{P}[f(\frac{\mathrm{d}\mathbb{P}_\theta(Y)}{\mathrm{d}\mathbb{P}(Y)})]$, where $\frac{\mathrm{d}\mathbb{P}_\theta(Y)}{\mathrm{d}\mathbb{P}(Y)}$ denotes the Radon-Nikodym derivative of $\mathbb{P}_\theta$ with respect to $\mathbb{P}$ (=likelihood ratio), which we assume exists for all $\theta \in \Theta$. Notably, $D_f(\mathbb{P}_\theta \| \mathbb{P})$ admits a useful dual representation:

$$D_f(\mathbb{P}_\theta \| \mathbb{P}) := \mathbb{E}_\mathbb{P} f\left(\frac{\mathrm{d}\mathbb{P}_\theta(Y)}{\mathrm{d}\mathbb{P}(Y)}\right) = \sup_{\lambda:\mathcal{Y}\to\mathbb{R}} \mathbb{E}_{\mathbb{P}_\theta}[\lambda(Y)] - \mathbb{E}_\mathbb{P}[f_*(\lambda(Y))] \qquad (2.1)$$

where $f_*(t) := \sup_{\lambda\in\mathbb{R}} \lambda t - f(\lambda)$ denotes the *convex conjugate* of f. The equality above follows from $f = (f_*)_*$. Various choices[1] for f are presented in Table 2.1. This duality is useful because the right-hand side suggests a finite-sample analog which does not depend on unknown quantities: we obtain an A-estimator for θ_* by letting

$$l(\theta, \lambda, Y) = \mathbb{E}_{\mathbb{P}_\theta}[\lambda(Y)] - f_*(\lambda(Y)) \qquad (2.2)$$

and solving for $\widehat{\theta}_n, \widehat{\lambda}_n$ satisfying the Nash condition 1.2,1.3 in $\mathbb{E}_n[l(\theta, \lambda, Y)]$. Normalizing $f(t) \leftarrow \frac{f(t) - f'(1)(t-1)}{f''(1)}$ without loss of generality[2], assuming the second derivative f'' exists, the function λ attaining the supremum in 2.1 at some θ is $\lambda_*^\theta = f'(\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}})$. The adversary $\widehat{\lambda}_n$ therefore estimates this transformed likelihood ratio at the current guess for $\widehat{\theta}_n$, and the Nash-equilibrium corresponds to the case where it is approximately constant, i.e. the distribution $\mathbb{P}_{\widehat{\theta}_n}$ is close to that of the data. Notably, $\mathbb{E}_n[l(\theta, \lambda, Y)]$ can be evaluated using only samples from the two distributions[3]. This is crucial for

---

[1]None of the objectives are unique: $f(t) \leftarrow f(t) + c(t-1)$ for any c yields the same divergence, but changes the expressions. Note that we may also swap $\mathbb{E}_{\mathbb{P}_\theta}$ and $\mathbb{E}_n$, which yields valid objectives for the respective "reverse" f-divergences.

[2]This implies $f'(1) = 0$, $f''(1) = 1$ and $f_*(0) = 0, f_*'(0) = f_*''(0) = 1$, which merely re-scales the divergence 2.1 by a factor of $1/f''(1)$

[3]Note that we neither require explicit knowledge of $\mathbb{P}_\theta$ nor infinitely many samples from $\mathbb{P}_\theta$ at a given n: it suffices to draw $m \succ n^2$ Monte Carlo samples from $\mathbb{P}_\theta$ and solve for the corresponding

GANs, where $\mathbb{P}_\theta$ is only implicitly defined via a push-forward mapping parametrized by a neural net. As proposed by Kaji et al. [2020], this also makes it a drop-in alternative to the *Simulated Method of Moments*, which similarly estimates economic models from data they generate, but matches only a finite set of moments instead of the full distribution.

Name	$f(t)$	$f_*(t)$, domain	Generative Adversarial Objective for θ
Total Variation	$|t-1|/2$	t, for $|t| \leq \frac{1}{2}$	$\sup_{|\lambda| \leq \frac{1}{2}} \mathbb{E}_{\mathbb{P}_\theta} \lambda(Y) - \mathbb{E}_n \lambda(Y)$
KL Divergence	$t \log t$	e^{t-1}	$\sup_{\lambda \in \mathbb{R}} 1 + \mathbb{E}_{\mathbb{P}_\theta} \lambda(Y) - \mathbb{E}_n e^{\lambda(Y)}$
Reverse KL	$-\log t$	$-\log(-te)$, for $t \leq 0$	$\sup_{\lambda \leq 0} 1 + \mathbb{E}_{\mathbb{P}_\theta} \lambda(Y) + \mathbb{E}_n \log(-\lambda(Y))$
χ^2 Divergence	$(t-1)^2$	$t + t^2/4$	$\sup_{\lambda \in \mathbb{R}} \mathbb{E}_{\mathbb{P}_\theta} \lambda(Y) - \mathbb{E}_n \left[\lambda(Y) + \lambda(Y)^2/4\right]$
Squared Hellinger	$(\sqrt{t}-1)^2$	$\frac{t}{1-t}$, for $t \leq 1$	$\sup_{\lambda \leq 1} \mathbb{E}_{\mathbb{P}_\theta} \lambda(Y) - \mathbb{E}_n \left[\frac{\lambda(Y)}{1-\lambda(Y)}\right]$
rescaled JS (GAN)	$t \log t - (1+t)\log(1+t)$	$-\log(1-e^t)$, for $t < 0$	$\sup_{\log \lambda < 0} \mathbb{E}_{\mathbb{P}_\theta} \log \lambda(Y) + \mathbb{E}_n \log(1-\lambda(Y))$

Table 1: Various adversarial f-divergence objectives. $f_*(t) = \infty$ outside the domain.

2.2 Generalized Empirical Likelihood

Our next example is a class of A-estimators that was proposed long before the recent success of adversarial objectives in deep learning. In econometrics, many important parameters θ_* are identified by a moment restriction of the form:

$$\mathbb{E}[m(Y,\theta)] = 0 \iff \theta = \theta_*$$

for some known, possibly vector-valued function $m(Y,\theta)$. In the Introduction, we presented the continuous-updating GMM objective (Hansen et al. [1996]) for estimating θ_*, a workhorse for causal inference in econometrics. In this section, we review the more general class of Generalized Empirical Likelihood (GEL) estimators (Newey and Smith [2004]), which solve the constrained minimization problem:

$$\inf_{\bar{\mathbb{P}}, \theta \in \Theta} D_f(\bar{\mathbb{P}} \| \mathbb{P}_n) \text{ s.t. } \mathbb{E}_{\bar{\mathbb{P}}}[m(Y,\theta)] = 0$$

That is, they seek for a parameter θ and a corresponding population distribution $\bar{\mathbb{P}}$ that is as close as possible to the sample $\mathbb{P}_n$, subject to satisfying the moment

finite sample saddle point. The resulting Monte Carlo approximation error for the expectation $\mathbb{E}_{\mathbb{P}_\theta}$ is then of order $\sqrt{m}^{-1} = n^{-1}$ and can thus be accounted for by letting $\widetilde{\eta}_n, \eta_n = O_{\mathbb{P}}(n^{-1})$ in equations 1.2,1.3, which has no impact on our asymptotic results.

constraint $\mathbb{E}_{\bar{\mathbb{P}}}[m(Y,\theta)] = 0$. At this high level, it is worth noting that GEL optimizes the same target as the objective in Section 2.1, which imposes $\bar{\mathbb{P}} = \mathbb{P}_\theta$ instead of a moment constraint. Glossing over some details, we can obtain a tractable estimator in this setting by concentrating out $\bar{\mathbb{P}}$ from the corresponding Lagrangian:

$$\inf_{\bar{\mathbb{P}},\theta\in\Theta} \sup_{\lambda\in\mathbb{R}^{\dim(m)}} D_f(\bar{\mathbb{P}}\|\mathbb{P}_n) - \lambda'\mathbb{E}_{\bar{\mathbb{P}}}[m(Y,\theta)] = \inf_{\theta\in\Theta} \sup_{\lambda\in\mathbb{R}^{\dim(m)}} \mathbb{E}_n\left[-f_*\left(\lambda'm(Y,\theta)\right)\right] \quad (2.3)$$

Which again uses the convex conjugate f_* of f (see example 2.1). For a formal proof of this equivalence, see e.g. Imbens et al. [1998]. It is easy to see that GEL is an A-estimator with $l(\theta,\lambda,Y) = -f_*(\lambda'm(Y,\theta))$ where $\Lambda_n = \mathbb{R}^{\dim(m)}$ and Θ_n is the parameter space of the economic model. A particularly popular version of this objective corresponds to the case $D_f = \chi^2$, where Table 2.1 tells us that $f(t) = (t-1)^2$ and $f_*(t) = t + t^2/4$. In this case, we can analytically solve for the optimal adversary given θ. Substituting it in, we get the continuous-updating GMM objective presented in the introduction:

$$\sup_{\lambda\in\mathbb{R}^{\dim(m)}} \mathbb{E}_n\left[-f_*\left(\lambda'm(Y,\theta)\right)\right] = \mathbb{E}_n\left[m(Y,\theta)\right]' \mathbb{E}_n\left[m(Y,\theta)m(Y,\theta)'\right]^{-1} \mathbb{E}_n\left[m(Y,\theta)\right]$$

2.3 Off-Policy Reinforcement Learning

Next, we review the Smoothed Bellman Error Embedding (SBEED) algorithm introduced by Dai et al. [2018], a popular off-policy learning method in robotics. Off-policy learning aims to learn the optimal policy for an agent from data that was generated under an entirely different policy regime. This problem is not limited to robotics: since it was identified in the monetary policy context by Lucas [1976], it became a primary concern in econometrics and its recognition played a key role in the *credibility revolution* (Angrist and Pischke [2010]) of econometrics. While problem definitions otherwise differ between these literatures, off-policy learning methods have received recent interest in econometrics (Zhan et al. [2021], Athey and Wager [2021]).

For an agent receiving reward $R(s,a)$ for taking action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$, forming an expectation over the future state $s^+ \in \mathcal{S}$, SBEED's goal is to learn the value function $V_*(s)$ and policy $a \sim P_*(\cdot|s)$ which satisfy the regularized Bellman equation:

$$V_*(s) = \max_{P(\cdot|s)} \mathbb{E}_{a\sim P(\cdot|s)}\left[R(s,a) + \beta\mathbb{E}_{s^+|s,a}[V_*(s^+)|s,a]\right] + H(P,s)$$

where the entropy $H(P, s) = -\mathbb{E}_{a \sim P(\cdot|s)}[\log P(a|s)]$ regularizes the optimal policy $P_*(\cdot|s)$ towards exploring all actions $a \in \mathcal{A}$. Given the researcher's choice of R, β, the goal is to learn P_*, V_* from finite samples $\{(s_i, a_i, s_i^+)\}_{i=1}^n$. Importantly, the actions a_i may be sampled from a *suboptimal* policy which does not equal P_*. Starting from the first-order condition of the Bellman equation, Dai et al. [2018] develop an adversarial population objective, whose finite-sample analog is the A-estimator 1.2,1.3 with loss:

$$l(\theta, \lambda, Y) = \big(R(s, a) + \beta V_\theta(s^+) - V_\theta(s) - \log P_\theta(a|s)\big)\lambda(s, a) - \frac{1}{2}\lambda(s, a)^2 \qquad (2.4)$$

where $\lambda(s, a)$, $\log P_\theta(a|s)$ and $V_\theta(s)$ are implemented as neural networks in practice.

2.4 A-Estimators for Conditional Moment Restrictions

Another powerful application for A-estimators recently pursued by the econometric literature are conditional moment estimators. These methods estimate parameters θ_* which are identified by restrictions of the form:

$$\mathbb{E}[m(X, \theta)|Z] = 0 \ \forall Z \iff \theta = \theta_* \qquad (2.5)$$

for some random variables $Y = (X, Z)$ and a known function $m(X, \theta)$. Conditions of this type occur e.g. when estimating some causal effect θ via instrumental variables, or as the first-order conditions of agents optimizing some expected utility given some information Z. As a result, nonparametric conditional moment estimators received considerable interest in econometrics, see e.g. Ai and Chen [2003, 2007], Chen and Qiu [2016]. These earlier estimators rely on first-step estimates of nuisance parameters capturing the conditional means and variances. Intuitively however, estimating the nuisance parameters via predictive objectives in a separate first step may dedicate scarce model capacity to capturing features which are not useful for the purpose of estimating θ_* downstream. This motivates recent work on adversarial objectives which unify the estimation into a single objective, more plausibly targeting the nuisance estimation towards the goal of identifying θ_*. Specifically, we will examine the estimator of Dikkala et al. [2020], with $l(\theta, \lambda, Y) = m(X, \theta)'\lambda(Z) - \frac{1}{4}\|\lambda(Z)\|_2^2$, yielding the finite sample objective

$$\inf_{\theta \in \Theta_n} \sup_{\lambda \in \Lambda_n} \mathbb{E}_n \left[m(X, \theta)'\lambda(Z) - \frac{1}{4}\lambda(Z)'\lambda(Z) \right],$$

where Λ_n is a class of neural networks. The methods proposed by Bennett et al. [2019a], Bennett and Kallus [2020] are closely related, but differ in the penalty they impose on λ. Dikkala et al. [2020] consider the case of instrumental variable regression, where $X = (y, x)$ and $m(X, \theta) = y - \theta(x)$, but we will examine the general case. We note that Example 2.3 (SBEED, Dai et al. [2018]) can be viewed as a special case of re-scaled version of this objective, with $X = (s, a, s^+)$ and $Z = (s, a)$, although both literatures seem to be unaware of their connection. One can analytically solve for the optimal adversary $\lambda_*^\theta(Z) = 2\mathbb{E}[m(X, \theta)|Z]$ to rewrite the population objective as:

$$\mathbb{E}[l(\theta, Y)] := \mathbb{E}[l(\theta, \lambda_*^\theta, Y)] = \mathbb{E}[\|\mathbb{E}[m(X, \theta)|Z]\|_2^2] = \mathbb{E}[\|\mathbb{E}[m(X, \theta) - m(X, \theta_*)|Z]\|_2^2]$$

which can be understood as a measure of distance between θ and θ_*, which clearly attains its minimum at $\theta = \theta_*$, when $\mathbb{E}[m(X, \theta)|Z] \equiv 0$. In Section 4.4, we will apply our theory to derive the asymptotic distribution of this estimator and show that is in fact *inefficient*. We further discuss how the adversarial formulation of GMM can directly inform a simple modification similar to Bennett and Kallus [2020] which yields an efficient A-estimator.

2.5 Estimating Riesz Representers

Chernozhukov et al. [2020] propose a distinct A-estimator to estimate *Riesz representers* for structural parameters ϕ_* which can be written as linear functionals $\phi_* = \phi(g_*) = \mathbb{E}[m(Y, g_*)]$. Here, $g_* = \mathbb{E}[y|x]$ is an unknown function for which an estimate $\widehat{g}_n$ is available from some first-stage regression of y on x, where $Y = (y, x, w)$. Quantities like ϕ_* are common in the average treatment effect or asset pricing literature, for example. Unfortunately, especially if $\widehat{g}_n$ is estimated via machine learning, the 'naive' estimator

$$\widehat{\phi}_n = \mathbb{E}_n[m(Y, \widehat{g}_n)]$$

is often not well behaved: $\sqrt{n}(\widehat{\phi}_n - \phi_*)$ may not converge in distribution to a Gaussian limit and thus one cannot provide confidence intervals around the estimate. Under the conditions of the Riesz representation theorem however, there may exist a function $\theta_* \in \Theta$ called the *Riesz representer* of the functional $\phi(g)$, which satisfies:

$$\phi(g) = \mathbb{E}[\theta_*(x)g(x)] \quad \forall g \in \Theta$$

If a well-behaved estimate $\widehat{\theta}_n$ of θ_* is available, it can be combined with $\widehat{g}_n$ to define the so-called *orthogonalized* estimator:

$$\tilde{\phi}_n = \mathbb{E}_n[m(Y, \widehat{g}_n) - \widehat{\theta}_n(x)(y - \widehat{g}_n(x))]$$

which attains asymptotic normality under rather weak conditions on $\widehat{g}_n$ (see Lemma 17 of Chernozhukov et al. [2020]). Chernozhukov et al. [2020] propose a generalized procedure to estimate $\widehat{\theta}_n$ via an A-estimator, which we will simplify as follows:

$$\inf_{\theta \in \Theta_n} \sup_{\lambda \in \Lambda_n} \mathbb{E}_n[m(Y, \lambda) - \theta(x)\lambda(x) - \lambda(x)^2/2]$$

where Θ_n, Λ_n are neural networks. To clarify why this objective works, is it useful to analytically solve the adversarial component of the corresponding population objective:

$$\sup_{\lambda} \mathbb{E}[m(Y, \lambda) - \theta(x)\lambda(x) - \lambda(x)^2/2] = \frac{1}{2}\mathbb{E}[(\theta_*(x) - \theta(x))^2]$$

As we will show in Section 4.5, our theory directly yields the convergence rates for $\widehat{\theta}_n$ that Chernozhukov et al. [2020]'s Lemma 17 requires for the asymptotic normality of $\tilde{\phi}_n$. It does so at a reduced curse of dimensionality in x for rather general function classes - i.e. under weaker conditions on smoothness and dimension of the data - complementing the original work.

3 General Theory

Roadmap. This Section will present our general theory of A-estimators. Subsection 3.1 briefly discusses an alternative definition of A-estimators that may be more natural to some readers. In Subsection 3.2, we establish that A-estimators satisfy the desirable condition of Neyman-orthogonality with respect to the adversary and discuss its implications. Next, we characterize the convergence rates of A-estimators: Section 3.3 provides a high-level result for arbitrary sieves such as splines or wavelets, not just neural nets. Under more easily verifiable low-level conditions, Subsection 3.4 provides convergence rates for semiparametric settings involving neural networks, showing they exhibit a reduced curse-of-dimensionality. Finally, we characterize the asymptotic normality of smooth functionals of A-estimators. We again begin with a

general, high-level result for arbitrary sieves, followed with the low-level conditions for the normality of neural networks. Notably, we show that a combination of under-smoothing and regularizing towards a convex target space suffices to overcome a key issue for normality proofs of neural networks: their non-convex sieve space.

Notation. Throughout, we consider random variable Y with support $\mathcal{Y}$, distribution $\mathbb{P}$ and corresponding expectation operator $\mathbb{E}$. We also denote the variance operator by $\mathbb{V}[f(Y)] = \mathbb{E}(f(Y) - \mathbb{E}[f(Y)])^2$ for any function $f : \mathcal{Y} \mapsto \mathbb{R}$. We denote the sample average, i.e. the expectation under the empirical distribution $\mathbb{P}_n$, by $\mathbb{E}_n$. Throughout, $\mathbb{E}$ will treat estimated parameters as deterministic sequences indexed by n, as is common in the literature. We also consider subvectors of Y, denoted by $x \in \mathcal{X}, \bar{x} \in \bar{\mathcal{X}}$, with their respective supports $\mathcal{X}, \bar{\mathcal{X}}$ being subspaces of $\mathcal{Y}$. We require various norms: throughout, $\|x\|_q$ will denote the ℓ^q norm of a finite dimensional vector x, with $\|x\| = \|x\|_2$ being the Euclidean norm. For a possibly vector-valued function $f(x)$, we denote its L^q function norm over some subset $\widetilde{\mathcal{X}} \subset \mathcal{X}$ by $\|f\|_{L^q(\widetilde{\mathcal{X}})} = \mathbb{E}[\|f(x)\|_q^q | x \in \widetilde{\mathcal{X}}]^{1/q}$. We denote the supremum norm of a vector x with components x_i by $\|x\|_\infty = \max_i |x_i|$. The supremum norm of f over $\widetilde{\mathcal{X}}$ will be denoted by $\|f\|_{\widetilde{\mathcal{X}}} = \sup_{x \in \widetilde{\mathcal{X}}} \|f(x)\|_\infty$. For $\widetilde{\mathcal{X}} = \mathcal{X}$, we may omit the dependence on $\mathcal{X}$ by writing $\|f\|_\infty := \|f\|_{\mathcal{X}}$. We will often write $a \prec b$ to denote $a = O(b)$, implying that a sufficiently large global constant $\infty > C > 0$ exists such that $a \leq Cb$, where C does not depend on any varying aspects of the problem, such as any parameters, sample sizes, et cetera. We write $a \asymp b$ if $a \prec b \prec a$. We will also write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. Throughout, we will write $l^\theta(\lambda, Y) = l(\theta, \lambda, Y)$ and $l(\theta, Y) = l(\theta, \lambda_*^\theta, Y)$ for short, where $\lambda_*^\theta = \arg\max_{\lambda \in \Lambda} \mathbb{E}l(\theta, \lambda, Y)$. We denote by π_n a (not necessarily linear) projection onto the respective sieves, i.e. $\pi_n\theta \in \arg\inf_{\theta' \in \Theta_n} \|\theta' - \theta\|_\infty$ for any $\theta \in \Theta$ and $\pi_n\lambda \in \arg\inf_{\lambda' \in \Lambda_n} \|\lambda' - \lambda\|_\infty$ for any $\lambda \in \Lambda$.

3.1 Nash vs Minimax

We presented our preferred definition for A-estimators in the introduction, as satisfying a Nash condition of the empirical loss. All results of this paper will apply to this definition. However, the reader may have noticed that the "simultaneous" Nash condition of the estimator is symmetric in $\widehat{\theta}_n$ and $\widehat{\lambda}_n$, unlike the 'sequential' mini-max

population objective, which nests a family of inner maximizations:

$$\lambda_*^\theta = \arg\max_{\lambda \in \Lambda} \mathbb{E}l(\theta, \lambda, Y) \qquad (3.1)$$

where the loss l and as a result the solutions λ_*^θ are indexed by the parameter $\theta \in \Theta$. The reader may therefore wonder if we could define an A-estimator for θ_* in a similar 'sequential' mini-max fashion. That is, we could consider a family of M-estimators $\widehat{\lambda}_n^\theta$ approximately maximizing the empirical loss at any value of $\theta \in \Theta$:

$$\widehat{\lambda}_n^\theta \in \Lambda_n : \quad \mathbb{E}_n l\left(\theta, \widehat{\lambda}_n^\theta, Y\right) \geq \sup_{\lambda \in \Lambda_n} \mathbb{E}_n l(\theta, \lambda, Y) - \eta_n \quad \forall \theta \in \Theta_n \qquad (3.2)$$

And then look for $\widehat{\theta}_n \in \Theta_n$ satisfying:

$$\mathbb{E}_n l(\widehat{\theta}_n, \widehat{\lambda}_n^{\widehat{\theta}_n}, Y) \leq \inf_{\theta \in \Theta_n} \mathbb{E}_n l(\theta, \widehat{\lambda}_n^\theta, Y) + \bar{\eta}_n \qquad (3.3)$$

where $\bar{\eta}_n = o_{\mathbb{P}}(1)$ again accommodates approximate minimization. Fortunately, it turns out that any $\widehat{\theta}_n$ satisfying the more compact Nash condition from the introduction always satisfies the mini-max condition presented above, as summarized by the following Lemma:

Lemma 3.1. *Any $\widehat{\theta}_n$, satisfying 1.2 and 1.3 for some $\widehat{\lambda}_n$, also satisfies 3.3 with some $\widehat{\lambda}_n^\theta$ for which 3.2, $\widehat{\lambda}_n^{\widehat{\theta}_n} = \widehat{\lambda}_n$ and $\bar{\eta}_n = \widetilde{\eta}_n + \eta_n$ holds.*

Proof. Pick any $\widehat{\theta}_n$ satisfying 1.2 and 1.3 for some $\widehat{\lambda}_n$. Now pick some arbitrary family $\widehat{\lambda}_n^\theta$ satisfying 3.2 for all $\theta \neq \widehat{\theta}_n$, and define $\widehat{\lambda}_n^{\widehat{\theta}_n} := \widehat{\lambda}_n$. Note that 1.3 directly implies that this $\widehat{\lambda}_n^\theta$ also satisfies 3.2 at $\theta = \widehat{\theta}_n$. It remains to show that the resulting $\widehat{\theta}_n$ and $\widehat{\lambda}_n^\theta$ satisfy 3.3:

$$\mathbb{E}_n l(\widehat{\theta}_n, \widehat{\lambda}_n^{\widehat{\theta}_n}, Y) \leq \inf_{\theta \in \Theta_n} \mathbb{E}_n l(\theta, \widehat{\lambda}_n, Y) + \widetilde{\eta}_n \leq \inf_{\theta \in \Theta_n} \mathbb{E}_n l(\theta, \widehat{\lambda}_n^\theta, Y) + \widetilde{\eta}_n + \eta_n$$

where the first inequality used $\widehat{\lambda}_n^{\widehat{\theta}_n}$ and the Nash condition 1.2, and the second used the fact that $\widehat{\lambda}_n^\theta$ was constructed to satisfy 3.2. $\qquad \square$

This reassures us that it suffices to find one set of values $\widehat{\theta}_n, \widehat{\lambda}_n$ which satisfy the Nash condition from the introduction, rather than a continuum of solutions $\widehat{\lambda}_n^\theta$ indexed by θ. The final $\widehat{\theta}_n$ will satisfy the mini-max condition regardless, for some (unknown) $\widehat{\lambda}_n^\theta$. For our theory, it was crucial to derive the uniform convergence of $\widehat{\lambda}_n^\theta$, hence we

will state the rate results for the more general mini-max definition. For the normality result, it was more convenient to work with the stronger Nash definition.

3.2 Adversaries are Neyman-Orthogonal

For many A-estimators, one could construct non-adversarial estimators which capture the same population objective. Whenever the adversarial nuisance parameter λ is a function, this usually requires a non-parametric first-step estimation of an alternative nuisance parameter. However, such an alternative estimator may not have a desirable property that is guaranteed for A-estimators: *Neyman-orthogonality* of θ_* with respect to the nuisance parameter.

This property has a long history in statistics, dating back at least to Neyman [1959]. It was popularized in econometrics by Chernozhukov et al. [2017] as a key setting in which standard machine learning methods can be applied without invalidating causal inference, which sparked follow-up work such as Chernozhukov et al. [2021] seeking to reformulate non-orthogonal problems as orthogonal ones. The notion applies to parameters which are identified by a moment restriction of the form:

$$\mathbb{E}[\varphi(\theta, \nu_*, Y)] = 0 \iff \theta = \theta_* \tag{3.4}$$

where φ is known and ν_* is an unknown nuisance parameter which has to be estimated in a first step. A popular estimator $\widehat{\theta}_n$ in this setting would be Hansen [1982]'s GMM, for example. The moment condition above is called (Neyman-)orthogonal whenever:

$$\nabla_{\nu_* \to \nu} \mathbb{E}[\varphi(\theta_*, \nu_*, Y)] = 0 \quad \forall \nu \tag{3.5}$$

Intuitively, this states that the condition identifying θ_* is "locally robust" against perturbations in ν_*. This guarantees that the uncertainty introduced by an appropriate first-step estimation of ν_* has no first-order effect on the GMM estimator $\widehat{\theta}_n$. Specifically, the asymptotic distribution of $\widehat{\theta}_n$ is the same as in the case in which ν_* is known. In contrast, when moment restrictions do not satisfy this orthogonality condition, uncertainty about ν_* generally amplifies the asymptotic variance of $\widehat{\theta}_n$, see e.g. Chen and Liao [2015], and normality may break down altogether.

Notably, if (and only if) θ_* is parametric, we can examine the first order condition

for θ_* that is implied by the A-estimation objective 1.1 in this moment restriction framework[4] : let $\varphi(\theta, \nu_*, Y) = \nabla_\theta l(\theta, \nu_*(\theta), Y)$, where $\nu_* : \Theta \mapsto \Lambda$ denotes the functional evaluating to $\nu_*(\theta) = \lambda_*^\theta$. Orthogonality then follows from the continuum of first-order conditions identifying ν_*:

$$\nabla_{\nu_* \to \nu} \mathbb{E}[l(\cdot, \nu_*(\cdot), Y)] \equiv \mathbf{0} \implies \nabla_{\nu_* \to \nu} \mathbb{E}[\varphi(\theta_*, \nu_*, Y)] = \nabla_{\nu_* \to \nu} \nabla_{\theta_*} \mathbb{E}[l(\theta_*, \nu_*(\theta_*), Y)] = \nabla_{\theta_*} \mathbf{0}$$

since the derivative operators are exchangeable. This implies that as $\widehat{\theta}_n$ approaches θ_*, an A-estimator $\widehat{\theta}_n$ is robust to estimation errors in the adversary $\widehat{\lambda}_n^\theta$ relative to λ_*^θ, meaning they do not reduce the accuracy of $\widehat{\theta}_n$, to a first-order.

Consider the example of Section 2.1, which estimates θ_* minimizing the f-Divergence between the model $\mathbb{P}_\theta$ and the data $\mathbb{P}$. As a non-adversarial alternative, we could re-parametrize the problem and estimate $\nu_* := d\mathbb{P}$ via a first-step Kernel density estimator $\widehat{\nu}_n(Y) = \widehat{d\mathbb{P}}_n(Y)$, and subsequently approximate the f-Divergence as the average over $f\left(\frac{d\mathbb{P}_\theta(Y)}{\widehat{\nu}_n(Y)}\right)$. However, the first-order condition for θ_* would not satisfy orthogonality, hence a GMM estimator based on this condition may not attain the variance of the analogous GMM estimator using ν_* instead. In contrast, the A-estimator of Section 2.1 *does* attain this variance - due to its orthogonal adversary - which we formally establish in Section 4.1. Moreover, this remains true when generalizing to a setting in which θ_* contains unknown functions, where no analogous GMM estimator exists that could capture the continuum of first-order conditions in θ_*.

3.3 Convergence Rate of A-Estimators

We begin with a general theorem characterizing the convergence rates of sieve A-estimators, for arbitrary loss functions and parameter spaces. It can be viewed as a generalization of Shen and Wong [1994]'s M-estimator result. Its proof is provided in Appendix C.1, with the main challenge being that Shen and Wong [1994]'s chaining arguments need to be carefully modified to hold uniformly over Θ. Our theorem adopts a more compact formulation than Shen and Wong [1994] which does not require any norm over Θ, Λ to state our assumptions, although convergence rates are

---

[4]Note however that even when θ_* is parametric, we usually cannot estimate it via GMM as $\nabla_\theta l(\theta, \widehat{\lambda}_n^\theta, Y)$ will not exist if $\widehat{\lambda}_n^\theta$ is a typical sieve, such as a neural network. For the same reason, the theory developed in this paper must not rely on any finite-sample first order conditions. Instead, it will use only the approximate Nash condition 1.3, 1.2.

obtained for any (pseudo-)norm $d(\theta, \theta_*)$ which is dominated by the objective.

Theorem 3.1 (Convergence Rate of A-Estimators). *Assume that:*

- *C1: The criterion variance is bounded by a power $\gamma > 0$ of its expectation:*

$$\mathbb{V}[l(\theta, Y) - l(\theta_*, Y)] \prec \mathbb{E}[l(\theta, Y) - l(\theta_*, Y)]^\gamma \tag{3.6}$$

$$\mathbb{V}[l^\theta(\lambda_*^\theta, Y) - l^\theta(\lambda, Y)] \prec \mathbb{E}[l^\theta(\lambda_*^\theta, Y) - l^\theta(\lambda, Y)]^\gamma \tag{3.7}$$

for all $\theta \in \Theta, \lambda \in \Lambda$ for which the right hand sides are less than some constant.

- *C2: For all small $\varepsilon > 0$, the covering number (Def. 1) is bounded via*

$$\log \mathcal{N}(\varepsilon, \{l(\theta, \lambda, \cdot) : \theta \in \Theta_n, \lambda \in \Lambda_n\}, \|\cdot\|_\infty) \prec n^s(\varepsilon^{-r} - 1)/r \tag{3.8}$$

for $0 \le s < 1$ and $r \ge 0$, where $r = 0$ represents $\lim_{r \to 0} n^s(\varepsilon^{-r} - 1)/r = n^s \log(1/\varepsilon)$.

Then the following conclusions hold.

i) *The criterion converges at rate:*

$$\mathbb{E}[l(\theta_*, Y) - l(\widehat{\theta}_n, Y)] = O_\mathbb{P}(n^{-\tau(\gamma, s, r, n)} + \epsilon_n + \eta_n + \bar{\epsilon}_n + \bar{\eta}_n) \tag{3.9}$$

$$\sup_{\theta \in \Theta_n} \mathbb{E}[l^\theta(\lambda_*^\theta, Y) - l^\theta(\widehat{\lambda}_n^\theta, Y)] = O_\mathbb{P}(n^{-\tau(\gamma, s, r, n)} + \epsilon_n + \eta_n) \tag{3.10}$$

where $\bar{\epsilon}_n = \mathbb{E}[l(\pi_n \theta_*, Y) - l(\theta_*, Y)]$ and $\epsilon_n = \sup_{\theta \in \Theta_n} \mathbb{E}[l^\theta(\lambda_*^\theta, Y) - l^\theta(\pi_n \lambda_*^\theta, Y)]$ are the sieve approximation errors. 3.10 also holds without 3.6. $\tau(\gamma, s, r, n)$ represents:

$$\tau(\gamma, s, r, n) = \begin{cases} 1 - s - \frac{\log \log n}{\log n}, & \text{if } r = 0, \gamma \ge 1 \\ \frac{1-s}{2-\gamma}, & \text{if } r = 0, \gamma < 1 \\ \frac{1-s}{2 - \min(1, \gamma)(2-r)/2}, & \text{if } 0 < r < 2 \\ \frac{1-s}{2} - \frac{\log \log n}{\log n}, & \text{if } r = 2 \\ \frac{1-s}{r}, & \text{if } r > 2 \end{cases}$$

ii) *Hence, $d(\widehat{\theta}_n, \theta_*) = o_\mathbb{P}(1)$ for any (pseudo-)norm $d(\cdot, \cdot)$ under which $\mathbb{E}[l(\theta, Y)]$ compact and continuous. If also $d(\theta, \theta_*)^{1/q} \prec \mathbb{E}[l(\theta, Y) - l(\theta_*, Y)]$ for $q > 0$, we get:*

$$d(\widehat{\theta}_n, \theta_*) = O_\mathbb{P}(n^{-\tau(\gamma, s, r, n)q} + \epsilon_n^q + \eta_n^q + \bar{\epsilon}_n^q + \bar{\eta}_n^q)$$

Remark 3.1 (Discussion of Assumptions). The theorem extends Shen and Wong [1994]'s convergence rate result for sieve M-estimators to A-estimators. There is a direct mapping between our assumptions and theirs: our C1 combines their assumptions C1 and C2, and our C2 corresponds to their C3. Our proof in Appendix C.1 is structured

in the same way as that of Shen and Wong [1994], although we need to modify their Lemmas to obtain the uniform convergence of the adversary in 3.10, which is crucial to the main result 3.9. The key modifications to our assumptions, which allow us to do so are: C1) that the constant factor implicit in the "$\prec$" relation of 3.7 must not depend on θ, as implied by the definition of "$\prec$" at the beginning of this section and C2) that the complexity of the *joint* sieve space $\Theta_n \times \Lambda_n$ satisfies the entropy bound. Otherwise, the assumptions are conceptually the same and we refer the reader to Shen and Wong [1994] for a more detailed discussion.

Remark 3.2. Using similar arguments as ours, one may establish the uniform convergence of A-estimators over a third parameter space, generalizing the setting to arbitrary finite sequences of min's and max's over different parameter spaces: e.g. $\min_\theta \max_\lambda \min_\gamma \mathbb{E}[l(\theta, \lambda, \gamma, Y)]$. This would yield convergence rates towards more general *Stackelberg equilibria* in so-called *empirical games*, for which we are currently only aware of a consistency result by Tuyls et al. [2018].

Remark 3.3. Beyond convergence rates for $\widehat{\theta}_n$ and $\widehat{\lambda}_n^\theta$, it is often useful to control the empirical process of arbitrary functions $f(\theta, \lambda, Y)$ of the parameters, e.g. to establish conditions for asymptotic normality required by Theorem 3.3. For this purpose, we provide Lemma B.5 in Appendix B.

3.4 Semiparametric Rates with Neural Networks

Next, we will apply the general result of the previous section to derive the convergence rates for neural network A-estimators. For generality, we will consider the semiparametric setting in which θ, λ may contain both Euclidean vectors and functions. These lower-level conditions are easy to verify in practice, but are general enough to apply to all estimators considered in Section 2. We will include the proof as it is short and an instructive application of Theorem 3.1. The theorem allows for two types of function classes, both of which can be viewed as generalizations of traditional Hölder functions with D-dimensional domain, with their own notion of an *intrinsic dimension $d^* \leq D$*, which may be smaller than that of the data D. As we will review in Remark 3.5, we observe that neural networks achieve a reduced curse of dimensionality in these settings.

Theorem 3.2 (Semiparametric Rates with Neural Networks)**.**
Consider the semiparametric setting in which $\Theta = \bar{\mathcal{B}} \times \bar{\mathcal{A}}$ and $\Lambda = \mathcal{B} \times \mathcal{A}$, where $\bar{\mathcal{B}}, \mathcal{B}$ are subsets of some Euclidean spaces and $\bar{\mathcal{A}}, \mathcal{A}$ are some function spaces. Let Λ, Θ be compact under $\|\cdot\|_\infty$. For all $\lambda, \lambda' \in \Lambda,\ \theta, \theta' \in \Theta$, assume the following conditions

hold:

- *A0: Assume that $\theta_* \in \Theta_*$ satisfies either*

  a) $\Theta_* \subset \bar{\mathcal{B}} \times \mathcal{H}(\bar{p}, \bar{\mathcal{X}})$ *on some* $\bar{\mathcal{X}} \subset [0,1]^{\bar{D}}$ *with* $\dim_M \bar{\mathcal{X}} = \bar{d}^* \leq \bar{D}$ *(see Def. 3 and 4)*

  b) $\Theta_* \subset \bar{\mathcal{B}} \times \mathcal{G}(\bar{p}, \bar{d}^*, [0,1]^{\bar{D}})$ *(see Def. 6)*

  and that $\{\lambda_*^\theta : \theta \in \Theta\} \subset \Lambda_*$ satisfies either

  a) $\Lambda_* \subset \mathcal{B} \times \mathcal{H}(p, \mathcal{X})$ *on some* $\mathcal{X} \subset [0,1]^D$ *with* $\dim_M \mathcal{X} = d^* \leq D$

  b) $\Lambda_* \subset \mathcal{B} \times \mathcal{G}(p, d^*, [0,1]^D)$

- *A1: $l(\theta, \lambda, Y) - l(\theta', \lambda', Y) \prec \|\theta - \theta'\|_{\bar{\mathcal{X}}} + \|\lambda - \lambda'\|_{\mathcal{X}}$*

- *A2: $\mathbb{V}[l(\theta, Y) - l(\theta_*, Y)] \prec \mathbb{E}[l(\theta, Y) - l(\theta_*, Y)] \prec \|\theta - \theta_*\|_{\widetilde{\mathcal{X}}}^2 + \mathbb{P}(\bar{x} \notin \widetilde{\mathcal{X}}) \; \forall \widetilde{\mathcal{X}} \subset \bar{\mathcal{X}}$*

- *A3: $\mathbb{V}[l^\theta(\lambda_*^\theta, Y) - l^\theta(\lambda, Y)] \prec \mathbb{E}[l^\theta(\lambda_*^\theta, Y) - l^\theta(\lambda, Y)] \prec \|\lambda - \lambda_*^\theta\|_{\widetilde{\mathcal{X}}}^2 + \mathbb{P}(x \notin \widetilde{\mathcal{X}}) \; \forall \widetilde{\mathcal{X}} \subset \mathcal{X}$*

Pick any two values $\bar{r} > \underline{r} \geq \left(\frac{d^*}{p} \vee \frac{\bar{d}^*}{\bar{p}} \right)$. Consider the A-estimator 3.2 with $\eta_n, \bar{\eta}_n = o_\mathbb{P}(n^{-2/(2+\bar{r})})$ where $\Lambda_n = \mathcal{B} \times \mathcal{F}_\sigma(L, W_n, w_n, \kappa_n)$ and $\Theta_n = \bar{\mathcal{B}} \times \mathcal{F}_\sigma(\bar{L}, \bar{W}_n, \bar{w}_n, \bar{\kappa}_n)$ implement neural networks (cf. Definition 2) satisfying $W_n, \bar{W}_n, w_n, \bar{w}_n \asymp n^{\underline{r}/(\underline{r}+2)}$ and $\kappa_n, \bar{\kappa}_n \asymp n^c$ for any large enough choice of $L, \bar{L}, c > 0$. For A0a) choose $\sigma(x) = \mathrm{ReLU}(x)$ and for A0b) choose $\sigma(x) = \tanh(x)$. Then:

$$\mathbb{E}[l(\widehat{\theta}_n, Y) - l(\theta_*, Y)] = o_\mathbb{P}(n^{-2/(2+\bar{r})})$$

$$\sup_{\theta \in \Theta_n} \mathbb{E}[l^\theta(\lambda_*^\theta, Y) - l^\theta(\widehat{\lambda}_n^\theta, Y)] = o_\mathbb{P}(n^{-2/(2+\bar{r})})$$

Hence, $d(\widehat{\theta}_n, \theta_*) = o_\mathbb{P}(1)$ for any (pseudo-)norm $d(\cdot, \cdot)$ under which $\mathbb{E}[l(\theta, Y)]$ is compact and continuous. Further, if $d(\theta, \theta_*)^{1/q} \prec \mathbb{E}[l(\theta, Y) - l(\theta_*, Y)]$ for $q > 0$, we get:

$$d(\widehat{\theta}_n, \theta_*) = o_\mathbb{P}(n^{-2q/(2+\bar{r})})$$

Proof. We will verify the conditions of Theorem 3.1. A2 and A3 imply C1 (3.6 and 3.7) with $\gamma = 1$. Lipschitzness A1 together with Lemma B.1 imply C2 (B.3) with $s = t/(t+2)$ for any $t : \bar{r} > t > \underline{r}$ and $r = 0$. Therefore Theorem 3.1 applies with $n^{-\tau(\gamma, s, r, n)} = n^{2/(2+t)} \log n \prec n^{2/(2+\bar{r})}$, which dominates η_n and $\bar{\eta}_n$ by assumption. We are therefore left with bounding ϵ_n and $\bar{\epsilon}_n$. By A3, we can bound $\epsilon_n \prec \sup_{\theta \in \Theta_n} \|\pi_n \lambda_*^\theta - \lambda_*^\theta\|_{\widetilde{\mathcal{X}}}^2 + \mathbb{P}(x \notin \widetilde{\mathcal{X}})$ for any $\widetilde{\mathcal{X}} \subset \mathcal{X}$. In the case of A0a), we set $\widetilde{\mathcal{X}} = \mathcal{X}$ and use Lemma B.2 to obtain $\sup_{\theta \in \Theta_n} \|\pi_n \lambda_*^\theta - \lambda_*^\theta\|_{\mathcal{X}}^2 \prec (W_n \wedge w_n)^{-2p/d^*} \prec n^{-2p\underline{r}/d^*/(2+\underline{r})} \prec n^{2/(2+\underline{r})}$ which yields $\epsilon_n = o(n^{2/(2+\bar{r})})$. For A0b), Lemma B.3 yields

the same bound as Lemma B.2, but only over a subset $\widetilde{\mathcal{X}} \subset \mathcal{X}$ with $P(x \notin \widetilde{\mathcal{X}}) \prec n^{-k}$ for some arbitrarily large constant $k > 0$, which only affects the constant c in the bound on κ_n. Hence we conclude that $\epsilon_n \prec n^{2/(2+\underline{r})} + n^{-k} \prec n^{2/(2+\underline{r})}$. Analogous arguments yield the same bound for $\bar{\epsilon}_n$. $\qquad\square$

Remark 3.4 (Discussion of Assumptions). A0 defines the function classes addressed by the Theorem. Both are generalizations of traditional Hölder classes which arise for $d^* = D$, see Remark 3.5. Condition A1 requires the loss to be Lipschitz in both parameters, which simplifies (but is not necessary for) the verification of C2. Condition A2 (and analogously A3) consists of two parts. First, it states that for a given parameter, the variance of the criterion difference must be bounded by its expectation, a simplified version of Assumption C1 of Theorem 3.1 which happens to be satisfied in all of our examples, but versions of this Theorem with $\gamma \neq 1$ can be derived via the same steps as the proof above. The second part of the condition bounds the expected loss by a squared sup-norm over any subset $\widetilde{\mathcal{X}}$ of the function domain $\mathcal{X}$. For the case of A0a), it would have sufficed to state the condition with $\widetilde{\mathcal{X}} = \mathcal{X}$ only, but for A0b) we require arbitrary subsets $\widetilde{\mathcal{X}}$ to apply the approximation result of Lemma B.3. A2 is implied, for example, by $\mathbb{E}[l(\theta, Y) - l(\theta_*)] \prec \|h(\theta) - h(\theta*)\|^2_{\mathcal{L}^q(\mathcal{X})}$ for some q and Lipschitz map $h : \Theta \mapsto \Theta$. The assumption is significantly weaker than Shen et al. [2019] or Farrell et al. [2018] who impose $\mathbb{E}[l(\theta, Y) - l(\theta_*)] \asymp \|\theta - \theta*\|^2_{\mathcal{L}^2(\mathcal{X})}$, which would not hold for Examples 2.2 or 2.4. It could be generalized further to allow for arbitrary powers of the sup-norms (and proved in the same way via Theorem 3.1), but the squares arise rather universally via Taylor expansions.

Remark 3.5. Theorem 3.2 clarifies that neural networks do not necessarily exhibit the curse of dimensionality, as the lower bound on $\bar{r}$ does not depend on the dimension D of the data. Instead, what matters is the *intrinsic dimension* d^* of the target function. In the setting A0a), introduced by Nakada and Imaizumi [2020], d^* refers to the Minkowski dimension of the manifold $\mathcal{X}$ which supports the data. It has been observed that $d^* \ll D$ for many high-dimensional types of data: intuitively, d^* is low whenever there is strong statistical dependency between the individual dimensions of the data. Examples include the characteristics of physical products, images and natural language. In the setting A0b), introduced by Bauer et al. [2019], d^* refers to the order of a generalized hierarchical interaction model. It is common for structural models in e.g. economics or optimal control to suggest that an unknown function is hierarchically composed of some finite number of individual functions which only

depend on $d^* \ll D$ inputs at a time. The result underscores that neural networks can *adaptively* - that is, without the researcher modifying the estimation procedure - exploit *structures* in the target function which allow them to model the relationships more efficiently than what standard convergence results suggest.

3.5   Asymptotic Normality of A-Estimators

In applications, it we a often interested in estimating a quantity of the form $F(\theta_*)$, where $F : \Theta \mapsto \mathbb{R}$ is some known functional. To derive confidence intervals around the plug-in estimate $F(\widehat{\theta}_n)$, we need its asymptotic distribution. To this end, we present Theorem 3.3, which can roughly be viewed as a generalization of Shen [1997] to A-Estimators. For this section, we make use of the pathwise derivative presented in Definition 7. We require a particular inner product over the space Θ:

$$\langle \theta, \theta' \rangle := \nabla_{\theta_* \to \theta} \nabla_{\theta_* \to \theta'} \mathbb{E}[l(\theta_*, Y)]$$

As discussed in Definition 7, the notation $\nabla_{\theta_* \to \theta}$ implicitly assumes that the corresponding limit exists and is linear in θ. For short, we write $\lambda_*'^\theta[v] := \nabla_{\theta \to v} \lambda_*^\theta$, $l'(\theta, Y)[v] := \nabla_{\theta \to v} l(\theta, Y)$ and $l'(\theta, \lambda, Y)[v, w] := \nabla_{\theta \to v} l(\theta, \lambda, Y) + \nabla_{\lambda \to w} l(\theta, \lambda, Y)$.

Theorem 3.3 (General Normality of A-Estimators).
Consider the estimators $\widehat{\theta}_n, \widehat{\lambda}_n$ satisfying the Nash conditions 1.2 and 1.3. Fix a sequence $e_n = o(n^{-1/2})$. Assume F is smooth enough and $\widehat{\theta}_n, \widehat{\lambda}_n$ converge fast enough such that a Riesz representer $v_* \in \Theta_*$ exists, satisfying:

$$\sup_{\theta \in \widehat{\Theta}_n} |F(\theta) - F(\theta_*) - \langle \theta - \theta_*, v_* \rangle| = O_{\mathbb{P}}(e_n) \tag{3.11}$$

Where $\widehat{\Theta}_n$ and $\widehat{\Lambda}_n(\theta)$ are the shrinking neighborhoods defined in Lemma B.5. For $v \in \{v_*, -v_*\}$, define the local perturbations $\bar{\theta}_n(\theta) = \theta - e_n v$ and $\bar{\lambda}_n^\theta(\lambda) = \lambda + e_n \lambda_*'^\theta[v]$ and assume:

CONDITION N1: Stochastic Equicontinuity

$$\sup_{\theta \in \widehat{\Theta}_n, \lambda \in \widehat{\Lambda}_n(\theta)} (\mathbb{E}_n - \mathbb{E}) l'(\theta, \lambda, Y)[v, \lambda_*'^\theta[v]] - l'(\theta_*, Y)[v] = O_{\mathbb{P}}(e_n)$$

CONDITION N2: Population Criterion Difference

$$\sup_{\theta \in \widehat{\Theta}_n, \lambda \in \widehat{\Lambda}_n(\theta)} \mathbb{E} l'(\theta, \lambda, Y)[v, \lambda_*'^\theta[v]] - l'(\theta_*, Y)[v] - \langle \theta - \theta_*, v \rangle = O_{\mathbb{P}}(e_n)$$

CONDITION N3: Approximation Error

$$\sup_{\theta \in \widehat{\Theta}_n, \lambda \in \widehat{\Lambda}_n(\theta)} \mathbb{E}_n l'(\theta, \lambda, Y)[\bar{\theta}_n(\theta) - \pi_n \bar{\theta}_n(\theta), \bar{\lambda}_n^\theta(\lambda) - \pi_n \bar{\lambda}_n^\theta(\lambda)] = O_{\mathbb{P}}(e_n^2)$$

If 1.2 and 1.3 are satisfied with $\widetilde{\eta}_n, \eta_n = O_{\mathbb{P}}(e_n^2)$, *then:*

$$\sqrt{n}\left(F(\widehat{\theta}_n) - F(\theta_*)\right) \xrightarrow{d} \mathcal{N}(0, V), \quad where \ V = \mathbb{V}\left(l'(\theta_*, Y)[v_*]\right)$$

Remark 3.6 (Discussion of Assumptions). In contrast to our convergence rate result, our proof requires the A-estimator to satisfy the (stronger) Nash condition from the introduction. Our conditions N1-3 are analogues of Shen [1997]'s and play the same roles in our proof. N1 combines their assumptions A and D, N2 corresponds to their B, and N3 to their C. Shen [1997]'s high-level discussion of their assumptions therefore applies to ours as well, and we again refer the reader there for additional context. The main difference is that their conditions are formulated to control the remainder of a second order Taylor expansion, whereas we look at the convergence of the first derivative, which results in $O_{\mathbb{P}}(e_n) = o_{\mathbb{P}}(n^{-1/2})$ requirements for N1 and N2, rather than the $O_{\mathbb{P}}(e_n^2) = o_{\mathbb{P}}(n^{-1})$ found in Shen [1997]'s conditions A and B.

Remark 3.7. Condition N3 is a version of a known condition on approximation error in M-estimation settings (see Condition C4 in Shen et al. [2019] and Condition C in Shen and Wong [1994]). Its verification usually exploits convexity of Θ_n, such that $\pi_n \bar{\theta}_n(\theta) = \theta + e_n \pi_n v_*$. This holds for series or kernel based estimators, but not neural networks. Shen et al. [2019] therefore leave it as an explicit assumption, concluding that it is unclear how to verify it for neural networks. In Theorem 3.4, we resolve this issue, showing that N3 can be verified for non-convex sieves such as neural networks by adhering to two simple implementation choices: 1) *undersmoothing*, i.e. choosing a sieve which grows faster than rate-optimal, achieving an approximation error of $o(n^{-1})$ and 2) regularizing the sieves towards the convex target classes containing θ_*, λ_*.

3.6 Semiparametric Normality with Neural Networks

Next, we present Theorem 3.4, which strengthens the assumptions of our previous neural network convergence rate result (Theorem 3.2) in a way that allows us to derive the asymptotic normality of functionals $F(\widehat{\theta}_n)$ via Theorem 3.3. A crucial innovation is that we are able to work around the non-convexity issues of deep neural networks discussed in Remark 3.7, to obtain a normality result from *low-level* conditions, which

only consist of general properties that the loss function must satisfy (A4-A7), and certain *implementation choices* for the neural networks that must be followed. To the best of our knowledge, the theorem therefore also provides the first low-level conditions for the normality of smooth functionals of deep neural network *M-estimators* (as the special case where Λ is singleton).

Theorem 3.4 (Semiparametric Normality with Neural Networks).
Let all assumptions of Theorem 3.2 be satisfied with $\frac{d^*}{p} \vee \frac{\bar{d}^*}{\bar{p}} < 1/4$, and choose $2 \geq \bar{r} > \underline{r} > 2/3$. Let Θ_*, Λ_* be convex and $\theta_*, \lambda_*^\theta$ lie in their interior. Replace the neural network sieves Θ_n, Λ_n with the following regularized versions:

$$\Theta_n \leftarrow \{\theta \in \Theta_n : \inf_{\theta' \in \Theta_*} \|\theta - \theta'\|_{\bar{\mathcal{X}}} \prec n^{-1-\epsilon}\}, \quad \Lambda_n \leftarrow \{\lambda \in \Lambda_n : \inf_{\lambda' \in \Lambda_*} \|\lambda - \lambda'\|_{\mathcal{X}} \prec n^{-1-\epsilon}\}$$

for any $\epsilon > 0$ which is small enough to guarantee that $\Theta_n, , \Lambda_n$ are nonempty. Further, for all $\theta, v \in \Theta, \ \lambda, w \in \Lambda$, assume:

- *A4: Lipschitz Derivative: $l'(\theta, \lambda, Y)[v, w] - l'(\theta', \lambda', Y)[v, w] \prec \|\theta - \theta'\|_{\bar{\mathcal{X}}} + \|\lambda - \lambda'\|_{\mathcal{X}}$*

- *A5: The perturbations are smooth: $v_* \in \Theta_*, \ \lambda_*'^\theta[v_*] \in \Lambda_*$*

- *A6: The Taylor remainders vanish with the loss:*

    i) $|\mathbb{E}l'(\theta, \lambda, Y)[v_*, \lambda_*^\theta[v_*]] - l'(\theta, Y)[v_*]| \prec \mathbb{E}[l(\theta, \lambda_*^\theta, Y) - l(\theta, \lambda, Y)]$
    ii) $|\mathbb{E}l'(\theta, Y)[v_*] - l'(\theta_*, Y)[v_*] - \langle \theta - \theta_*, v_* \rangle| \prec \mathbb{E}[l(\theta, Y) - l(\theta_*, Y)]$

- *A7: For non-Donsker classes, the variance of the derivatives is bounded by the loss:*

    i) $\mathbb{V}[l'(\theta, \lambda, Y)[v, \lambda_*'^\theta[v]] - l'(\theta, \lambda_*^\theta, Y)[v, \lambda_*'^\theta[v]]] \prec \mathbb{E}[l(\theta, \lambda_*^\theta, Y) - l(\theta, \lambda, Y)]$ or Λ_* is $\mathbb{P}$-Donsker
    ii) $\mathbb{V}[l'(\theta, Y)[v] - l'(\theta_*, Y)[v]] \prec \mathbb{E}[l(\theta, Y) - l(\theta_*, Y)]$ or Θ_* is $\mathbb{P}$-Donsker

If $\widehat{\theta}_n, \widehat{\lambda}_n$ satisfy the Nash condition 1.2,1.3 with $\eta_n, \tilde{\eta}_n = o_\mathbb{P}(n^{-1})$, then:

$$\sqrt{n}\left(F(\widehat{\theta}_n) - F(\theta_*)\right) \overset{d}{\longrightarrow} \mathcal{N}(0, V), \quad where \ V = \mathbb{V}\left(l'(\theta_*, Y)[v_*]\right)$$

Remark 3.8 (Discussion of Assumptions). The Theorem requires that the neural network sieves Θ_n, Λ_n are implemented to undersmooth (i.e. grow faster than the rate-optimal sieve would) via the condition on $\underline{r}$, while being regularized towards the convex target spaces Θ_*, Λ_*. Note that this does not affect the sieve's approximation power towards these spaces, and there always exists an $\epsilon > 0$ for which Θ_n, Λ_n are non-empty due to their $o(n^{-1})$ approximation rates. While in principle just an implementation choice, the current sup-norm regularization is arguably not practical and

future work may be able to clarify whether e.g. an appropriate L2 penalty on the weights suffices. Conditions A4-A7 are general conditions on the loss function which can be satisfied in all our examples. A4 is a simple Lipschitz condition analogous to A1. The smoothness of the Riesz representer (A5) is most easily verified by computing and examining a given $v_*, \lambda'^\theta_*[v_*]$ directly, although the Riesz representation theorem can provide general conditions under which v_* lives in the same space as θ_*. A6 is a standard condition controlling the Taylor remainder. For a discussion, see e.g. Assumptions 4.5 in Ai and Chen [2003] and Ai and Chen [2007], or Assumption 3.5ii) in Chen and Pouzo [2015]. Whether it holds depends on how non-linear the objective is: e.g. for the quadratic objective of Dikkala et al. [2020], the left-hand side is zero. A7 serves to control the empirical process (N1). It can be easily satisfied either by bounding the variances of the derivatives, or by relying on the Donsker property of the target space (cf. Remark 3.9).

Remark 3.9. Note that the Donsker property and thus A7 always holds if $p > D/2$, where standard results using bracketing number bounds imply that the Hölder spaces Θ_*, Λ_* satisfy the Donsker property. We conjecture that this analogously holds for our lower-dimensional classes A0a) and A0b) whenever $d^*/p < 2$, which would make the verification of A7 unnecessary in general, since we require $\frac{d^*}{p} \vee \frac{\bar{d}^*}{\bar{p}} < 1/4$. Verifying this conjecture is beyond the scope of this paper however, hence we provide A7 as an explicit assumption for maximum flexibility.

4   Application to Examples

4.1   Minimum f-Divergence

Applying our general Theorem 3.2 to the estimator of Section 2.1 yields Proposition 4.1, which provides the convergence rate of semiparametric $\widehat{\theta}_n$ if Λ and all unknown functions in Θ are approximated by classes of neural networks.

Proposition 4.1. *Let* $\theta_* \in \Theta_* \subset \Theta, \lambda^\theta_* = f'(\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}}) \in \Lambda_* \subset \Lambda,$ *where* Θ, Λ *are compact under* $\| \cdot \|_\infty$ *and path-connected, and the target function classes* Θ_*, Λ_* *satisfy A0 in Theorem 3.2. Fix some* $C < \infty$. *For any* $\theta \in \Theta$, *let* $0 < f'' \left(\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}}(Y) \right) < C$ *wp1 and for any* $\lambda \in \Lambda$, *let* $0 < f''_* (\lambda(Y)) < C$ *wp1. Let* $\left\| \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}} - \frac{\mathrm{d}\mathbb{P}_{\theta'}}{\mathrm{d}\mathbb{P}} \right\|_\infty \prec \|\theta - \theta'\|_\infty$. *Let* Θ_n, Λ_n *be constructed as in Theorem 3.2, with all neural networks growing in width*

at some rate $n^{\underline{r}/(\underline{r}+2)}$ *satisfying* $\underline{r} \geq \frac{d^*}{p} \vee \frac{\bar{d}^*}{\bar{p}}$. *Then for any* $\bar{r} > \underline{r}$:

$$D_f(\mathbb{P}_{\widehat{\theta}_n}\|\mathbb{P}) = o_{\mathbb{P}}(n^{-2/(2+\bar{r})})$$

Remark 4.1. In general, the convergence rate of $\widehat{\theta}_n$ is faster the slower the growth rate $n^{\underline{r}/(\underline{r}+2)}$ of the neural network. However, the growth must be fast enough to control the approximation error of the sieves Θ_n, Λ_n relative to the target function classes Θ_*, Λ_*. This lower bound depends on the ratio of the *smoothness* of the target classes p and $\bar{p}$ and their *intrinsic dimensions* d^* and $\bar{d}^*$, which may be smaller than that of the data Y, in which case f-GANs attain a reduced curse-of-dimensionality relative to traditional nonparametric density estimators.

Remark 4.2. This convergence rate result stands in contrast to Arora et al. [2017], who argued that Generative Adversarial Networks do not generalize with respect to the metric given by the population objective, only under a weaker "neural net distance" which they introduce. The convergence rate result above clarifies that the broad class of f-GANs in fact *does* converge quickly under population divergence.

While a fast convergence rate of the model distribution $\mathbb{P}_{\widehat{\theta}_n}$ is a key goal in semi- and nonparametric estimation, whenever some function $F(\widehat{\theta}_n)$ of the estimate informs downstream decision-making, we are often interested in obtaining confidence intervals around $F(\widehat{\theta}_n)$. To this end, we derive the asymptotic normality of the adversarial f-Divergence objective - an entirely novel result at this level of generality, to the best of our knowledge. First, we compute the inner product defined in Section 3.5, which can be expressed concisely:

$$\langle \theta, \theta' \rangle = \nabla_{\theta_* \to \theta} \nabla_{\theta_* \to \theta'} \mathbb{E}\left[f\left(\frac{d\mathbb{P}_{\theta_*}(Y)}{d\mathbb{P}(Y)} \right) \right] = \mathbb{E}\left[\nabla_{\theta_* \to \theta} \log d\mathbb{P}_{\theta_*}(Y) \cdot \nabla_{\theta_* \to \theta'} \log d\mathbb{P}_{\theta_*}(Y) \right]$$

Where $\nabla_{\theta_* \to \theta} \log d\mathbb{P}_{\theta_*}(Y) = \nabla_{\theta_* \to \theta} \frac{d\mathbb{P}_{\theta_*}(Y)}{d\mathbb{P}(Y)}$ is a pathwise derivative of the Radon-Nikodym derivative. Conditions under which the normality result of Section 3.5 applies are presented in Proposition 4.2.

Proposition 4.2. *Consider a functional* $F(\theta)$ *for which a Riesz representer* v_* *exists satisfying 3.11 with* $\langle \cdot, \cdot \rangle$ *defined above. Let all assumptions of Theorem 4.1 be satisfied for* $d^*/p \vee \bar{d}^*/\bar{p} < 1/4$ *and assume that* $\times_*$ *is Donsker. Let* Θ_*, Λ_* *be convex, let* $\theta_*, \lambda_*^\theta$ *lie in their interior, and let them contain* $v_*, \lambda_*'^{,\theta}[v_*]$. *Assume the Lipschitz condition* $\|\nabla_{\theta \to v} \frac{d\mathbb{P}_\theta}{d\mathbb{P}} - \nabla_{\theta' \to v} \frac{d\mathbb{P}_{\theta'}}{d\mathbb{P}}\|_\infty \prec \|\theta - \theta'\|_\infty$ *and let* f'' *be Lipschitz. Pick* $2 \geq \bar{r} > \underline{r} > 2/3$ *and regularize* Θ_n, Λ_n *as in Theorem 3.4. Finally, for any* $\widetilde{\theta}, \widetilde{\theta}'$ *on a path between* θ_*

and θ, *assume that:*

$$\nabla_{\widetilde{\theta} \to \widetilde{\theta}' - \theta_*} \nabla_{\widetilde{\theta} \to \theta - \theta_*} \nabla_{\widetilde{\theta} \to v_*} D_f(\mathbb{P}_{\widetilde{\theta}} \| \mathbb{P}_{\theta_*}) \prec D_f(\mathbb{P}_\theta \| \mathbb{P}_{\theta_*})$$

Then:

$$\sqrt{n}(F(\theta_n) - F(\theta_*)) \xrightarrow{d} \mathcal{N}(0, \langle v_*, v_* \rangle) \tag{4.1}$$

Remark 4.3. In applications, the key difficulty lies in verifying that the third derivative above is bounded by the loss. This condition serves to control the higher order term of the Taylor expansion. Such assumptions are common in the semiparametric literature, e.g. Ai and Chen [2003]'s Assumptions 4.5 and 4.6 play the same role. It is easiest to verify in the parametric setting, where

$$\nabla_{\widetilde{\theta} \to \widetilde{\theta}' - \theta_*} \nabla_{\widetilde{\theta} \to \theta - \theta_*} \nabla_{\widetilde{\theta} \to v_*} D_f(\mathbb{P}_\theta \| \mathbb{P}_{\theta_*}) \asymp \| \theta - \theta_* \|_2^2 \asymp D_f(\mathbb{P}_\theta \| \mathbb{P}_{\theta_*})$$

Note that $\langle \cdot, \cdot \rangle$ and hence the asymptotics of $\widehat{\theta}_n$ are independent of f, so the f-divergences are asymptotically equivalent. An example for a smooth functional $F(\theta)$ that is of particular interest in the semiparametric setting $\theta = (\beta, \alpha)$ is $F(\theta) = \beta' \zeta$, which "picks out" a linear combination of the parametric components. This allows us to derive the asymptotic normality of the vector $\sqrt{n}(\widehat{\beta} - \beta_*)$ in the following Corollary, which makes use of the orthogonal scores assumption that is standard in the semiparametric literature.

Corollary 4.2.1. *In addition to the assumptions of Proposition 4.2, assume the orthogonal scores condition holds:*

$$\mathbb{E}\left[\nabla_{\beta_* \to \beta} \log d\mathbb{P}_{\beta_*, \alpha_*}(Y) \nabla_{\alpha_* \to \alpha} \log d\mathbb{P}_{\beta_*, \alpha_*}(Y) \right] = 0 \quad \forall \beta, \alpha$$

Then the parametric component $\widehat{\beta}_n$ attains the Cramér-Rao bound:

$$\sqrt{n}(\widehat{\beta}_n - \beta_*) \xrightarrow{d} \mathcal{N}\left(0, I^{-1}\right), \text{ where } I = \mathbb{E}\left[\nabla_{\beta_*} \log d\mathbb{P}_{\beta_*, \alpha_*}(Y) \cdot \nabla_{\beta_*'} \log d\mathbb{P}_{\beta_*, \alpha_*}(Y) \right]$$

Proof. We simply choose $v_* = (I^{-1}\zeta, \mathbf{0})$, such that $\langle \theta - \theta_*, v_* \rangle = (\beta - \beta_*)' \zeta = F(\theta) - F(\theta_*)$. Since $\langle v_*, v_* \rangle = \zeta' I^{-1} \zeta$, Proposition 4.2 yields $\sqrt{n}(\widehat{\beta}_n - \beta_*)' \zeta \xrightarrow{d} \mathcal{N}(0, \zeta' I^{-1} \zeta)$. The result then follows via the Cramér-Wold device. $\square$

The f-GAN objective therefore attains the efficient asymptotics of maximum likelihood, but does not require explicit knowledge of the model density $\mathbb{P}_\theta$.

4.2 Generalized Empirical Likelihood

For the class of Generalized Empirical Likelihood estimators introduced in Section 2.2, the $\sqrt{n}$-normality and asymptotic efficiency of $\widehat{\theta}_n$ is long established in the parametric case (Imbens et al. [1998], Imbens [2002], Newey and Smith [2004]). However, our theoretical framework still allows us to extend the known results to the semiparametric case where θ may contain unknown functions, which are approximated by a class of neural networks Θ_n which may grow with n. In this case, we can characterize the convergence rate of $\widehat{\theta}_n$ to the identified set $\Theta_* = \{\theta \in \Theta : \mathbb{E}[m(Y,\theta)] = 0\}$, which is unlikely to be singleton given that an infinite-dimensional parameter is hardly pinned down by a finite number of unconditional moment restrictions. We obtain the following result:

Proposition 4.3. *Let $D_f = \chi^2$ and consider the A-estimator $\widehat{\theta}_n, \widehat{\lambda}_n$ satisfying 1.2,1.3 with $l(\theta, \lambda, Y) = -f_*(\lambda' m(Y, \theta))$. Let Θ_*, Θ_n be as in Theorem 3.2 and $\Lambda_* = \Lambda_n = \mathbb{R}^{\dim(m)}$, with $\bar{r} > \frac{d^*}{p}$. Assume that $m(Y, \theta) - m(Y, \theta') \prec \|\theta - \theta'\|_\infty$ and $|m(Y, \theta)| < \infty$. Then:*

$$\mathbb{E}\left[m(Y, \widehat{\theta}_n)\right] = o_{\mathbb{P}}(n^{-1/(2+\bar{r})})$$

Proof. We verify the conditions of Theorem 3.2. Assumption A0 holds by assumption, and A1 follows from the Lipschitzness of $m(Y, \cdot)$ and that of $f_*(t) = t + t^2/4$. To verify Assumption 2, note that $l(\theta_*, Y) = 0$ and boundedness of $m(Y, \theta)$ imply:

$$\mathbb{V}[l(\theta, Y) - l(\theta_*, Y)] \prec \mathbb{E}[(m(Y, \theta)' \lambda_*^\theta)^2] \asymp \mathbb{E}[l(\theta, Y) - l(\theta_*, Y)]$$

For the second part of condition A2, simply verify that $\mathbb{E}[l(\theta, Y) - l(\theta_*, Y)] \prec \|\lambda_*^\theta - \lambda_*^{\theta_*}\|_2^2 \prec \|\theta, \theta_*\|_{\widetilde{\mathcal{X}}}^2 + \mathbb{P}(\bar{x} \notin \widetilde{\mathcal{X}})$, which follows by applying the Lipschitzness of m in θ and the tower-property of $\mathbb{E}$ to $\lambda_*^\theta = -2\mathbb{E}[m(Y, \theta)m(Y, \theta)']^{-1}\mathbb{E}[m(Y, \theta)]$, akin to the proof of 4.1. Assumption A3 can be verified for the Euclidean λ via a Taylor expansion, yielding: $\mathbb{V}[l(\theta, \lambda, Y) - l(\theta, \lambda_*^\theta, Y)] \asymp \|\lambda - \lambda_*^\theta\|_2^2 \asymp \mathbb{E}[l(\theta, \lambda, Y) - l(\theta, \lambda_*^\theta, Y)]$. $\quad\square$

4.3 Off-Policy Reinforcement Learning

Next, we will use our theory to the extend the known results about SBEED, the off-policy RL algorithm of Dai et al. [2018] introduced in Section 2.3. Theorem 3.2 makes it easy to obtain the convergence rates of the corresponding A-estimator:

Proposition 4.4. *Consider the A-estimator $\widehat{\theta}_n, \widehat{\lambda}_n$ satisfying 1.2,1.3 with $l(\theta, \lambda, Y)$ as in 2.4. Assume the observations are iid for simplicity, and that $P_* = P_{\theta_*}$ and $V_* = V_{\theta_*}$, where $\theta_* \in \Theta_*, \lambda_*^\theta \in \Lambda_*$ satisfy A0 in Theorem 3.2 with $\mathcal{X} = \bar{\mathcal{X}} = \mathcal{S} \times \mathcal{A}$. Let $\Theta_* \subset \Theta, \Lambda_* \subset \Lambda$, with Θ, Λ compact under $\|\cdot\|_\infty$ and path-connected. Let $R(\cdot, \cdot), V_\theta(\cdot), P_\theta(\cdot|\cdot)$ be continuous. Let the parametrizations P_θ, V_θ satisfy the Lipschitz conditions $\|\log P_\theta - \log P_{\theta'}\|_\infty \prec \|\theta - \theta'\|_\infty$ and $\|V_\theta - V_{\theta'}\|_\infty \prec \|\theta - \theta'\|_\infty$. Let the neural network classes Θ_n, Λ_n be constructed as in Theorem 3.2, for any $\underline{r} \geq \frac{d^*}{p} \vee \frac{\bar{d}^*}{\bar{p}}$. Then for any $\bar{r} > \underline{r}$:*

$$\mathbb{E}_{s,a}\left[\left(R(s,a) + \beta \mathbb{E}[V_{\widehat{\theta}_n}(s^+)|s,a] - V_{\widehat{\theta}_n}(s) - \log P_{\widehat{\theta}_n}(a|s)\right)^2\right] = o_\mathbb{P}(n^{-2/(2+\bar{r})})$$

Remark 4.4. In contrast to the original work, our result also applies in the case where $\mathcal{A}$ and $\mathcal{S}$ are continuous, and we characterize the optimal rate of growth for the neural network function approximators, which optimally trade off bias and variance. While following almost trivially from the general Theorem 3.2, our result yields significantly faster convergence rates than the $o_\mathbb{P}(\sqrt{n})$ rates obtained by Dai et al. [2018], and our rates further exhibit the reduced curse of dimensionality of neural networks.

Remark 4.5. We noticed that SBEED can be viewed as a special case of some of the econometric conditional moment estimators treated in Example 2.4, such as Dikkala et al. [2020]. We therefore refer the reader to Section 4.4 for an application of our asymptotic normality result. Interestingly, neither literature seems to be aware of this connection. Dai et al. [2018] cite convex conjugation and the interchangeability principle as the inspiration for their objective, whereas the adversarial conditional moment estimators in econometrics were inspired by Hansen [1982]'s Generalized Method of Moments.

4.4 A-Estimators for Conditional Moment Restrictions

We will now apply our theory to examine the asymptotic behavior of the conditional moment estimator of Dikkala et al. [2020], introduced in Section 2.4. We can apply Theorem 3.2 to obtain the rate at which $\widehat{\theta}_n$ converges:

Proposition 4.5. *Let $\Theta_n, \Theta_*, \Lambda_n, \Lambda_*$ be as in Theorem 3.2. Let $m(X, \theta)$ be $\|\cdot\|_\infty$-Lipschitz in θ. Let the support of Y be bounded. Then, for any $\bar{r} > \frac{d^*}{p} \vee \frac{\bar{d}^*}{\bar{p}}$, we get:*

$$\mathbb{E}\left[\left\|\mathbb{E}[m(X, \widehat{\theta}_n) - m(X, \theta_*)|Z]\right\|_2^2\right] = o_\mathbb{P}(n^{2/(2+\bar{r})})$$

For the instrumental variable regression setting studied by Dikkala et al. [2020], where

$m(X, \theta) = y - \theta(x)$, *this implies:*

$$\mathbb{E}\left[\left\|\mathbb{E}\left[\widehat{\theta}_n(x) - \theta_*(x)\big|Z\right]\right\|_2^2\right] = o_{\mathbb{P}}(n^{2/(2+\bar{r})})$$

Proof. Condition A0 is satisfied by assumption, and A1 follows from Lipschitzness of $m(X, \cdot)$ and boundedness. Assumptions A2 and A3 can be verified by using boundedness to establish

$$\mathbb{V}[l(\theta, Y) - l(\theta_*, Y)] \prec \|\lambda_*^\theta\|_{\mathcal{L}(\mathcal{Z})^2}^2 \asymp \mathbb{E}[l(\theta, Y) - l(\theta_*, Y)]$$
$$\mathbb{V}[l(\theta, \lambda, Y) - l(\theta, \lambda_*^\theta, Y)] \prec \|\lambda - \lambda_*^\theta\|_{\mathcal{L}(\mathcal{Z})^2}^2 \asymp \mathbb{E}[l(\theta, \lambda, Y) - l(\theta, \lambda_*^\theta, Y)]$$

$\square$

Remark 4.6. Note that just like in the previous Example 2.2, this result does not require the parameter θ_* to be identified by the restriction 2.5. If that is the case however, the above rates can be translated into similar rates in any norm $\|\widehat{\theta}_n - \theta_*\|$ which is dominated by the objective, usually by construction. See Ai and Chen [2003] for an example of such a norm in the semi-parameteric setting.

Remark 4.7. In contrast to Dikkala et al. [2020], our convergence rate result allows for general m and possibly vector-valued, semiparametric Θ in which unknown functions are approximated by neural networks. Our rates are also exhibit the reduced curse of dimensionality of neural networks.

Next, we use Theorem 3.4 to derive the asymptotic variance of the estimator, showing that the estimator is in inefficient in general. For this purpose, it suffices to only consider the simpler parametric setting.

Proposition 4.6. *Consider the parametric case where $\Theta_n = \Theta_* = \Theta$ is Euclidean. In addition to the assumptions of Proposition 4.5, assume that the identification condition 2.5 holds. Let $d(X, \theta) := \nabla_\theta m(X, \theta)$ be bounded and satisfy the Lipschitz condition $|d(X, \theta) - d(X, \theta')| \prec \|\theta - \theta'\|_\infty$. Assume that $\mathbb{E}[l(\theta, Y)]$ is three times differentiable in θ. For all $\theta \in \Theta$, let $\lambda_*^\theta := 2\mathbb{E}[m(X, \theta)|Z] \in \Lambda_*$ for a Λ_* satisfying A0 with $\frac{d^*}{p} < \frac{1}{4}$, let $\theta_*, \lambda_*^\theta$ lie in the respective interiors of Θ, Λ_*, and let $\lambda_*'^\theta[v_*](\cdot) := 2v_*'\mathbb{E}[d(X, \theta)|Z = \cdot] \in \Lambda_*$ for any $v_* \in \Theta$. Let Λ_n be regularized as in Theorem 3.4. Then:*

$$\sqrt{n}(\widehat{\theta}_n - \theta_*) \xrightarrow{d} \mathcal{N}(0, V)$$

where $V = \mathbb{E}\left[\mathbb{E}\left[\nabla_{\theta_*}m(X, \theta_*)'|Z\right]\mathbb{E}\left[m(X, \theta_*)m(X, \theta_*)'|Z\right]\mathbb{E}\left[\nabla_{\theta_*}m(X, \theta_*)|Z\right]\right]^{-1}$.

Chamberlain [1987] derived the efficiency bound for the parametric conditional mo-

ment setting, corresponding to the smallest (in a p.s.d. sense) $\sqrt{n}$-asymptotic variance for any unbiased estimator. It is given by the covariance matrix:

$$V_* = \mathbb{E}\left[\mathbb{E}\left[\nabla_{\theta_*}m(X,\theta_*)'|Z\right]\mathbb{E}\left[m(X,\theta_*)m(X,\theta_*)'|Z\right]^{-1}\mathbb{E}\left[\nabla_{\theta_*}m(X,\theta_*)|Z\right]\right]^{-1}$$

Note that $V \neq V_*$ in general, implying that $\widehat{\theta}_n$ is an inefficient estimator. By extension, this also applies to the Reinforcement Learning algorithm of Example 2.3. Comparing the GMM objective of Example 2.2 - which is known to be efficient in the unconditional moment setting - to the population objective of the present example, this may be unsurprising: in contrast to GMM, the population objective of Dikkala et al. [2020] corresponds to a regular ℓ^2 norm, without the inverse covariance weighting which is crucial for asymptotic efficiency in the unconditional case. Generalizing GEL to the conditional moment setting by replacing the constant adversary with a neural network Λ_n, Metzger [2022] therefore proposes the A-estimator given by:

$$\inf_{\theta \in \Theta_n} \sup_{\lambda \in \Lambda_n} \mathbb{E}_n\left[-f_*\left(m(X,\theta)'\lambda(Z)\right)\right]$$

which nests a simplified variant of Bennett and Kallus [2020] for $D_f = \chi^2$, and for $D_f = D_{KL}$ can be viewed as alternative to the Kernel approach of Kitamura et al. [2004]. Metzger [2022] provides a similar information theoretic foundation as the GEL estimator and - building on the theory developed in the present paper - derives the convergence rates and asymptotic efficiency of this estimator, where Θ_n may contain unknown functions which are modeled as neural networks.

4.5   Estimating Riesz Representers

Finally, we show that Theorem 3.2 can be used to quickly derive the convergence rates of Chernozhukov et al. [2020]'s adversarial estimator for Riesz representers, which we introduced in Section 2.5.

Proposition 4.7. *Let* $\Theta_n, \Theta_*, \Lambda_n, \Lambda_*$ *be as in Theorem 3.2. Let* $m(Y,\lambda) = m(Y,\lambda(x))$ *be Lipschitz in* $\lambda(x)$. *Let the support of* Y *be bounded. Then, for any* $\bar{r} > \frac{d^*}{p} \vee \frac{\bar{d}^*}{\bar{p}}$:

$$\|\widehat{\theta}_n - \theta_*\|_{\mathcal{L}^2(x)} = o_{\mathbb{P}}(n^{1/(2+\bar{r})})$$

This result clarifies that the Riesz representer of Chernozhukov et al. [2020] can similarly benefit from the adaptivity properties of neural networks, which yield faster

rates for our target classes if $d_* < D$. In combination with their Lemma 17, this implies that compared to other non-parametric sieves, neural networks guarantee the asymptotic normality of the orthogonalized estimator $\tilde{\phi}_n$ under weaker conditions on smoothness and D. Since the normality of $\tilde{\phi}_n - \phi_*$ is of primary interest and already follows from Chernozhukov et al. [2020]'s Lemma 17 given our convergence rates, we refrain from deriving it for arbitrary functionals $\widehat{\phi}_n(g) = \mathbb{E}[\widehat{\theta}_n(x)g(x)]$, although it would be possible to use Theorem 3.4 to derive $\sqrt{n}(\widehat{\phi}_n(g) - \phi(g)) \xrightarrow{d} \mathcal{N}(0, V_g)$ for some V_g for example.

5   Conclusion

We characterize the general class of adversarial estimators *('A-estimators')*, subsuming many estimators independently proposed in the fields of econometrics and machine learning. Our unified framework suggests interesting commonalities between A-estimators: their adversary is always Neyman-orthogonal with respect to the main model, guaranteeing that its estimation errors have no first-order asymptotic impact on the estimated model. Most objectives have versions which asymptotically minimize an f-divergence criterion and are asymptotically efficient. Typically, A-estimators adaptively learn how to optimally emphasize the restrictions implied researcher's estimation assumptions, performing particularly well when this set is large. This makes them a promising framework for incorporating machine learning methods into causal inference, where even simple target parameters often satisfy a continuum of restrictions. We characterize the convergence rates of A-estimators, as well as the asymptotic normality of smooth functionals of their parameters. We also provide low-level analogues of these results for semi-parametric models, in which unknown functions are approximated by deep neural networks. Our convergence and normality results also extend the theory of neural network M-estimators, as a special case: building on recent results in approximation theory, our neural network converge rates exhibit a reduced curse of dimensionality for more general losses than previously examined, which hold uniformly over a second parameter space. Our normality result overcomes a problem previously posed by the non-convexity of neural network sieves, showing that a particular regularization, combined with under-smoothing, can be used to satisfy a strong, high-level approximation error condition which the literature left hitherto unverified.

References

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71 (6):1795–1843, 2003. doi: https://doi.org/10.1111/1468-0262.00470. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00470.

Chunrong Ai and Xiaohong Chen. Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141(1):5–43, 2007. URL https://EconPapers.repec.org/RePEc:eee:econom:v:141:y:2007:i:1:p:5-43.

Joshua D. Angrist and Jörn-Steffen Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2):3–30, June 2010. doi: 10.1257/jep.24.2.3. URL https://www.aeaweb.org/articles?id=10.1257/jep.24.2.3.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 224–232. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arora17a.html.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021. doi: https://doi.org/10.3982/ECTA15732. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15732.

Susan Athey, Guido W Imbens, Jonas Metzger, and Evan M Munro. Using wasserstein generative adversarial networks for the design of monte carlo simulations. Technical report, National Bureau of Economic Research, 2019.

Benedikt Bauer, Michael Kohler, et al. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Annals of Statistics*, 47(4):2261–2285, 2019.

Denis Belomestny, Eric Moulines, Alexey Naumov, Nikita Puchkin, and Sergey Sam-

sonov. Rates of convergence for density estimation with gans. *arXiv preprint arXiv:2102.00199*, 2021.

Andrew Bennett and Nathan Kallus. The variational method of moments. *CoRR*, abs/2012.09422, 2020. URL `https://arxiv.org/abs/2012.09422`.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. In *NeurIPS*, 2019a.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL `https://proceedings.neurips.cc/paper/2019/file/15d185eaa7c954e77f5343d941e25fbd-Pap`

Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987.

Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Statistical guarantees of generative adversarial networks for distribution estimation, 2020.

Xiaohong Chen and Zhipeng Liao. Sieve semiparametric two-step gmm under weak dependence. *Journal of Econometrics*, 189(1):163–186, 2015. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2015.07.001. URL `https://www.sciencedirect.com/science/article/pii/S0304407615002031`.

Xiaohong Chen and Demian Pouzo. Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, 2015.

Xiaohong Chen and Yin Jia Qiu. Methods for nonparametric and semiparametric regressions with endogeneity: a gentle guide. Cowles Foundation Discussion Papers 2032, Cowles Foundation for Research in Economics, Yale University, 2016. URL `https://EconPapers.repec.org/RePEc:cwl:cwldpp:2032`.

Xiaohong Chen, Oliver Linton, and Ingrid Van Keilegom. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5): 1591–1608, 2003.

Xiaohong Chen, Demian Pouzo, and James L. Powell. Penalized sieve gel for weighted average derivatives of nonparametric quantile iv regressions. *Journal of Econometrics*, 2019.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *The American Economic Review*, 107:261–265, 2017.

Victor Chernozhukov, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. Adversarial estimation of riesz representers, 2020.

Victor Chernozhukov, Whitney Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. Automatic debiased machine learning via neural nets for generalized linear regression. 2021.

Stephen Cosslett. Maximum likelihood estimator for choice-based samples. *Econometrica*, 49:1289–1316, 02 1981. doi: 10.2307/1912755.

Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. Two-player games for efficient non-convex constrained optimization. In Aurélien Garivier and Satyen Kale, editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 300–332. PMLR, 22–24 Mar 2019. URL https://proceedings.mlr.press/v98/cotter19a.html.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. *CoRR*, abs/1712.10285, 2018. URL http://arxiv.org/abs/1712.10285.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *arXiv preprint arXiv:2006.07201*, 2020.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *arXiv preprint arXiv:1809.09953*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Ulf Grenander. Abstract inference. Technical report, 1981.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912775.

Lars Peter Hansen, John Heaton, and Amir Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3): 262–280, 1996. ISSN 07350015. URL http://www.jstor.org/stable/1392442.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *NIPS*, abs/1606.03476, 2016. URL http://arxiv.org/abs/1606.03476.

Guido W. Imbens. Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics*, 20(4):493–506, 2002. ISSN 07350015. URL http://www.jstor.org/stable/1392419.

Guido W. Imbens, Richard H. Spady, and Phillip Johnson. Information theoretic approaches to inference in moment condition models. *Econometrica*, 66(2):333–357, 1998. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2998561.

Abdul Jabbar, Xi Li, and Bourahla Omar. A survey on generative adversarial networks: Variants, applications, and training. *ACM Computing Surveys (CSUR)*, 54: 1 – 49, 2022.

Tetsuya Kaji, Elena Manresa, and Guillaume Pouliot. An adversarial approach to structural estimation, 2020.

Yuichi Kitamura, Gautam Tripathi, and Hyungtaik Ahn. Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*, 72(6):1667–1714, 2004.

Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments, 2018.

Tengyuan Liang. How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(228):1–41, 2021.

Robert E. Lucas. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46, 1976. ISSN

0167-2231. doi: https://doi.org/10.1016/S0167-2231(76)80003-6. URL https://www.sciencedirect.com/science/article/pii/S0167223176800036.

Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks, 2017.

Jonas Metzger. Adversarial conditional moment estimation, 2022.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020. URL http://jmlr.org/papers/v21/20-002.html.

Whitney K. Newey and Richard J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/3598854.

Jerzy Neyman. Optimal asymptotic tests of composite statistical hypotheses. *Probability and Statistics: The Harald Cramer Volume*, 1959.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization, 2016.

Art Owen. Empirical Likelihood Ratio Confidence Regions. *The Annals of Statistics*, 18(1):90 – 120, 1990. doi: 10.1214/aos/1176347494. URL https://doi.org/10.1214/aos/1176347494.

Art B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 06 1988. ISSN 0006-3444. doi: 10.1093/biomet/75.2. 237. URL https://doi.org/10.1093/biomet/75.2.237.

Jin Qin and Jerry Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325, 1994. ISSN 00905364. URL http://www.jstor.org/stable/2242455.

Xiaotong Shen. On methods of sieves and penalization. *The Annals of Statistics*, 25(6):2555–2591, 1997. ISSN 00905364. URL http://www.jstor.org/stable/2959045.

Xiaotong Shen and Wing Hung Wong. Convergence rate of sieve estimates. *Ann. Statist.*, 22(2):580–615, 06 1994. doi: 10.1214/aos/1176325486. URL `https://doi.org/10.1214/aos/1176325486`.

Xiaoxi Shen, Chang Jiang, Lyudmila Sakhanenko, and Qing Lu. Asymptotic properties of neural network sieve estimators. *arXiv preprint arXiv:1906.00875*, 2019.

Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabás Póczos. Nonparametric density estimation under adversarial losses. *arXiv preprint arXiv:1805.08836*, 2018.

Chenyang Tao, Liqun Chen, Ricardo Henao, Jianfeng Feng, and Lawrence Carin Duke. Chi-square generative adversarial network. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4887–4896. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/tao18b.html`.

K Tuyls, J Perolat, M Lanctot, JZ Leibo, and T Graepel. A generalised method for empirical game theoretic analysis. In *AAMAS'18: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 77–85. ACM, 2018.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

Ruohan Zhan, Vitor Hadad, David A. Hirshberg, and Susan Athey. Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 2125–2135, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467456. URL `https://doi.org/10.1145/3447548.3467456`.

A   Definitions

Definition 1 (Covering Number)**.**
For some norm $\| \cdot \|$ over some metric space Λ, the covering number $\mathcal{N}(\delta, \Lambda, \| \cdot \|)$ is

defined as the cardinality of the smallest set $C \subset \Lambda$ such that $\sup_{\lambda \in \Lambda} \inf_{c \in C} \|\lambda - c\| \leq \delta$. The quantity $\log \mathcal{N}(\delta, \Lambda, \|\cdot\|)$ is also called metric entropy.

Definition 2 (Deep Neural Networks).

We define the class of deep σ networks $f \in \mathcal{F}_\sigma(L, W, w, \kappa, B)$ as parametrized functions of the form:

$$f(x) = A^{(L)} \cdot \sigma \left(A^{(L-1)} \cdots \sigma \left(A^{(1)} x + b^{(1)} \right) \cdots + b^{(L-1)} \right) + b^{(L)}$$

where the $A^{(l)}$'s are weight matrices and $b^{(l)}$'s are intercept vectors with real-valued elements, and $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is applied element-wise. For example, the choice $\sigma(x) = \mathrm{ReLU}(x) = \max\{0, x\}$ (rectified linear unit) gives rise to the class of deep ReLU networks, and $\sigma(x) = \tanh(x)$ gives rise to the class of tanh networks. We say the network is L layers deep and call the upper bound $\sup_l \dim(b^{(l)}) \leq w$ its width. Further, we assume that

$$\max_{i,j,l} \left| A_{ij}^{(l)} \right| \leq \kappa, \max_{i,l} |b_i^{(l)}| \leq \kappa, \sum_{l=1}^L \left\| A^{(l)} \right\|_0 + \left\| b^{(l)} \right\|_0 \leq W, \text{ for } i = 1, \ldots, L$$

i.e. all elements in the $A^{(l)}$'s and $b^{(l)}$'s are bounded in absolute value by κ, and there are at most W non-zero parameters in total. Finally, we assume $\|f\|_\infty \leq B < \infty$ for all f. If the particular value B is an arbitrary large enough constant, we may suppress the notation and write $\mathcal{F}_\sigma(L, W, w, \kappa, B) = \mathcal{F}_\sigma(L, W, w, \kappa)$.

Definition 3 (Minkowski Dimension).

The (upper) Minkowski dimension of a set $\mathcal{X} \subset [0, 1]^D$ is defined as

$$\dim_M \mathcal{X} := \inf \left\{ d^* \geq 0 \mid \limsup_{\varepsilon \downarrow 0} \mathcal{N}(\varepsilon, \mathcal{X}, \|\cdot\|_\infty) \varepsilon^{d^*} = 0 \right\}$$

where $\mathcal{N}(\varepsilon, \mathcal{X}, \|\cdot\|_\infty)$ is given by Definition 1. As shown in Nakada and Imaizumi [2020], this definition generalizes many other notions of intrinsic dimension, such as the manifold dimension.

Definition 4 (Hölder Space).

For a function $f : \mathbb{R}^D \to \mathbb{R}, \partial_d f(x)$ is a partial derivative with respect to a d-th component, and $\partial^\alpha f := \partial_1^{\alpha_1} \cdots \partial_D^{\alpha_D} f$ using multi-index $\alpha = (\alpha_1, \ldots, \alpha_D)$. For $z \in \mathbb{R}$ $\lfloor z \rfloor$ denotes the largest integer that is less than z. Let $p > 0$ be a degree of smoothness. For $f : [0, 1]^D \to \mathbb{R}$, the Höder norm is defined as

$$\|f\|_{\mathcal{H}(p, [0,1]^D)} := \max_{\alpha : \|\alpha\|_1 < \lfloor p \rfloor} \sup_{x \in [0,1]^D} |\partial^\alpha f(x)| + \max_{\alpha : \|\alpha\|_1 = \lfloor p \rfloor} \max_{x, x' \in [0,1]^D, x \neq x'} \frac{|\partial^\alpha f(x) - \partial^\alpha f(x')|}{\|x - x'\|_\infty^{p - \lfloor p \rfloor}}$$

Then, the Hölder space on $[0, 1]^D$ is defined as

$$\mathcal{H}\left(p, [0, 1]^D \right) = \left\{ f \in C^{\lfloor p \rfloor} \left([0, 1]^D \right) \mid \|f\|_{\mathcal{H}(p, [0,1]^D)} < \infty \right\}$$

Also, $\mathcal{H}\left(p, [0,1]^D, M\right) = \left\{ f \in \mathcal{H}\left(p, [0,1]^D\right) \mid \|f\|_{\mathcal{H}(p,[0,1]^D)} \leq M \right\}$ denotes the M-radius closed ball in $\mathcal{H}\left(p, [0,1]^D\right)$.

Definition 5 ((p, C)-smoothness).
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 $\frac{\partial^q m}{\partial x_1^{\alpha} \ldots \partial x_d^{\alpha_d}}$ exists and satisfies

$$\left| \frac{\partial^q m}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}(x) - \frac{\partial^q m}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}(z) \right| \leq C \cdot \|x - z\|_2^s$$

for all $x, z \in \mathbb{R}^d$.

Definition 6 (Generalized Hierarchical Interaction Models).
Let $C \in \mathbb{R}_{\geq 0}$, $D \in \mathbb{N}$, $d^* \in \{1, \ldots, D\}$, $m : \mathbb{R}^D \to \mathbb{R}$ and $p = q + s$ for some $q \in \mathbb{N}_0$ and $0 < s \leq 1$.

a) We say that m satisfies a generalized hierarchical interaction model of order d^* and level 0 with bound C, if there exist $a_1, \ldots, a_{d^*} \in \mathbb{R}^D$ and some $f : \mathbb{R}^{d^*} \to \mathbb{R}$ such that
$$m(x) = f\left(a_1^T x, \ldots, a_{d^*}^T x\right) \quad \text{for all } x \in \mathbb{R}^D$$

and where f is Lipschitz continuous with constant C and all of its partial derivatives of order less than or equal to q are bounded in absolute value by by C.

b) We say that m satisfies a generalized hierarchical interaction model of order d^* and level $l + 1$ with bound C if there exist $K \in \mathbb{N}$, $g_k : \mathbb{R}^{d^*} \to \mathbb{R}(k = 1, \ldots, K)$ and $f_{1,k}, \ldots, f_{d^*,k} : \mathbb{R}^D \to \mathbb{R}(k = 1, \ldots, K)$ such that $f_{1,k}, \ldots, f_{d^*,k}(k = 1, \ldots, K)$ satisfy a generalized hierarchical interaction model of order d^* and level l and

$$m(x) = \sum_{k=1}^K g_k\left(f_{1,k}(x), \ldots, f_{d^*,k}(x)\right) \quad \text{for all } x \in \mathbb{R}^D$$

where g_k are Lipschitz continuous with constant C and all of their partial derivatives of order less than or equal to q are bounded by some constant C.

c) We say that the generalized hierarchical interaction model defined above is (p, C)-smooth, if all functions occurring in its definition are (p, C)-smooth, cf. Definition 5.

d) We define $\mathcal{G}(p, d^*, C, [0,1]^D)$ as the class of all functions $m : [0,1]^D \to \mathbb{R}$ satisfying a (p, C)-smooth generalized hierarchical interaction model of order d^* and level l with bound C, where $l \leq C$. Since the particular value of C is not important as long as $C < \infty$, we also write $\mathcal{G}(p, d^*, [0,1]^D)$.

Definition 7 (Pathwise Derivatives).
For some $\theta \in \Theta$, $\lambda \in \Lambda$ and some functional $l : \Theta \times \Lambda \mapsto \mathbb{R}^d$, we define the first pathwise derivative in the direction $\theta' \in \Theta$ as

$$\nabla_{\theta \to \theta'} l(\theta, \lambda) := \lim_{\tau \to 0} \frac{\partial}{\partial \tau} l(\theta + \tau \theta', \lambda)$$

for some real number $\tau \in \mathbb{R}$. Throughout this paper, the usage of $\nabla_{\theta \to \theta'}$ implicitly assumes that the derivative and limit on the RHS exists and is linear in θ'.

B   Supporting Lemmas

Lemma B.1 (Covering Number of Neural Networks).
Consider the class of deep neural networks $f \in \mathcal{F}_\sigma(L, W, w, \kappa)$ (Definition 2), with activation σ satisfying $\sigma : |\sigma(x)| \leq x, |\sigma(x) - \sigma(x')| \leq |x - x'| \; \forall x, x' \in \mathbb{R}$ (e.g. ReLU, tanh) and consider the norm $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$ for some $\mathcal{X} \subset [0, 1]^D$ where $D \leq w$. Its δ-covering number (Definition 1) can be bounded by:

$$\mathcal{N}\left(\delta, \mathcal{F}_\sigma(L, W, w, \kappa), \|\cdot\|_\infty\right) \leq \left(\frac{2L^2(w+2)(\kappa w)^{L+1}}{\delta}\right)^W$$

Proof. This is Lemma 7 in Chen et al. [2020]. While they only state the Lemma for the case of ReLU networks $\sigma(x) = \max(0, x)$, their proof works for any activation σ satisfying $|\sigma(x)| \leq x$ and $|\sigma(x) - \sigma(x')| \leq |x - x'|$ for all $x, x' \in \mathbb{R}$. We substituted the bound $B = 1$ and renamed some variables. $\qquad\square$

Lemma B.2 (Approximation by Deep ReLU Networks on Low Dimensional Data).
Consider the Hölder space $\mathcal{H} \equiv \mathcal{H}\left(p, [0, 1]^D\right)$ (Definition 4) and some support $\mathcal{X} \subset [0, 1]^D$ with Minkowski dimension (Definition 3) bounded by $\dim_M \mathcal{X} \leq d^* \leq D$. For any small enough $\epsilon > 0$, the class of deep ReLU networks $\mathcal{F} \equiv \mathcal{F}_{\text{ReLU}}(L, W(\epsilon), w(\epsilon), \kappa(\epsilon))$ (Definition 2) satisfies:

$$\sup_{f_* \in \mathcal{H}} \inf_{f \in \mathcal{F}} \sup_{x \in \mathcal{X}} |f(x) - f_*(x)| < \epsilon$$

as long as $W(\epsilon) \geq c_1 \epsilon^{-d^*/p}$, $w(\epsilon) \geq c_2 \epsilon^{-d^*/p}$, $\kappa(\epsilon) \geq c_3 \epsilon^{-c_4}$ for any large enough choice of L, c_1, c_2, c_3, $c_4 > 0$.

Proof. The case $d^* < D$ is covered by Theorem 5 in Nakada and Imaizumi [2020]. While they do not state a bound on the width $w(\epsilon)$, it is easy to see that any network described by Definition 2 with at most $W(\epsilon)$ non-zero parameters can be represented by a network with width bounded by $w(\epsilon) \leq W(\epsilon)$. In the case of $d^* = D$, the Lemma simply states the approximation error for conventional Hölder spaces as established in Yarotsky [2017]. $\qquad\square$

Lemma B.3 (Approximation of Generalized Interaction Models by Deep ReLU Networks).

Consider the function class $\mathcal{G} \equiv \mathcal{G}(p, d^*, [0,1]^D)$ (Definition 6d) and consider some arbitrary random variable $x \in [0,1]^D$ with probability measure $\mathbb{P}_x$. For any small enough $\epsilon, \eta > 0$, the class of deep tanh networks $\mathcal{F} \equiv \mathcal{F}_{\tanh}(L, W(\epsilon), w(\epsilon), \kappa(\epsilon))$ (Definition 2) satisfies:

$$\sup_{f_* \in \mathcal{G}} \inf_{f \in \mathcal{F}} \sup_{x \in \mathcal{X}} |f(x) - f_*(x)| < \epsilon$$

for some subset $\mathcal{X} \subset [0,1]^D$ with $\mathbb{P}_x(x \notin \mathcal{X}) \leq \eta$ as long as $W(\epsilon) = c_1 \epsilon^{-d^*/p}$, $w(\epsilon) = c_2 \epsilon^{-d^*/p}$, $\kappa(\epsilon) = c_3 \epsilon^{-c_4}/\eta$ for any large enough choice of L, c_1, c_2, c_3, $c_4 > 0$.

Proof. This directly follows from Theorem 3 in Bauer et al. [2019], however our notation is greatly simplified by the fact that we are not interested in most of their constants, and that we offloaded most of the assumptions into Definition 6. What matters is that the network they construct has a depth that bounded by a constant (their equation (6)), and a number of non-zero parameters that is proportional to what they call $(M_n + 1)^{d^*}$ in their Theorem 3 (by their equations (7) and (5) and the definition of M^* in their Theorem 3). Since we assumed bounded support (leaving their a_n as a constant), their bound yields an approximation error of $\epsilon =: cM_n^{-p}$ for some $c > 0$, such that the number of non-zero parameters can be bounded as $W(\epsilon_n) = O\left((M_n + 1)^{d^*}\right) = O(\epsilon^{-d^*/p})$. They bound $\kappa(\epsilon)$ (α in their notation) in terms of M_n and η yielding $\kappa(\epsilon) = O(\epsilon^{-c_4}/\eta)$ for some large enough constant $c_4 > 0$. Finally, their theorem holds only for activation functions σ which satisfy a property they call *N-admissible*. While this is technically not satisfied by $\sigma(x) = \tanh(x)$, it is easy to verify that this property is satisfied by the activation function $\widetilde{\sigma}(x) = 1/2 + \tanh(x)/2$. Since for any $\tilde{f} \in \mathcal{F}_{\widetilde{\sigma}}(L, W, w, \kappa)$ there exists some $f \in \mathcal{F}_{\tanh}(L, W, w, 2\kappa + 1/2)$ such that $\tilde{f} = f$, the same approximation bound holds with $\sigma(x) = \tanh(x)$. $\square$

Lemma B.4 (Empirical Process of Donsker Classes).

If $Y \in \mathcal{Y}$ is iid and $\{l(\theta, Y) : \theta \in \Theta\}$ is $\mathbb{P}$-Donsker for some $l : \Theta \times \mathcal{Y} \mapsto \mathbb{R}$ satisfying

$$\lim_{\theta \in \Theta: \|\theta - \theta_*\| \to 0} \mathbb{E}\left[(l(\theta, Y) - l(\theta_*, Y))^2\right] = 0,$$

then

$$\sup_{\theta \in \Theta: \|\theta - \theta_*\| \leq \delta_n} (\mathbb{E} - \mathbb{E}_n)[l(\theta, Y) - l(\theta_*, Y)] = o_{\mathbb{P}}(n^{-1/2})$$

for any $\delta_n = o_{\mathbb{P}}(1)$.

Proof. This directly follows from Lemma 1 in Chen et al. [2003]. $\square$

Lemma B.5 (Empirical Process Rates for A-Estimators).

Under the assumptions of Theorem 3.1, for any function $f(\theta, \lambda, Y)$ satisfying the following conditions:

- *For any sequence $e_n \geq 0$ and all $\theta \in \Theta, \lambda \in \Lambda$:*

$$\mathbb{V}[f(\theta, \lambda_*^\theta, Y) - f(\theta_*, \lambda_*^{\theta_*}, Y)] \prec \mathbb{E}[l(\theta, Y) - l(\theta_*, Y) + e_n]^\gamma \qquad \text{(B.1)}$$

$$\mathbb{V}[f(\theta, \lambda, Y) - f(\theta, \lambda_*^\theta, Y)] \prec \mathbb{E}[l^\theta(\lambda_*^\theta, Y) - l^\theta(\lambda, Y) + e_n]^\gamma \qquad \text{(B.2)}$$

at least if the right hand sides are smaller than some $C > 0$.

- *For all small $\varepsilon > 0$, we have:*

$$\log \mathcal{N}(\varepsilon, \{f(\theta, \lambda, \cdot) : \theta \in \Theta_n, \lambda \in \Lambda_n\}, \|\cdot\|_\infty) \prec n^s(\varepsilon^{-r} - 1)/r \qquad \text{(B.3)}$$

we obtain the following empirical processes bounds:

$$\sup_{\theta \in \widehat{\Theta}_n} (\mathbb{E} - \mathbb{E}_n)[f(\theta, \pi_n \lambda_*^\theta, Y) - f(\theta_*, \lambda_*^{\theta_*}, Y)] = O_\mathbb{P}(n^{-\tau(\gamma, s, r, n)} + \epsilon_n + \eta_n + \bar{\epsilon}_n + \bar{\eta}_n + e_n)$$

$$\sup_{\substack{\theta \in \widehat{\Theta}_n \\ \lambda \in \widehat{\Lambda}_n(\theta)}} (\mathbb{E} - \mathbb{E}_n)[f(\theta, \lambda, Y) - f(\theta, \pi_n \lambda_*^\theta, Y)] = O_\mathbb{P}(n^{-\tau(\gamma, s, r, n)} + \epsilon_n + \eta_n + e_n)$$

where $\widehat{\Lambda}_n(\theta) := \{\lambda \in \Lambda_n : \mathbb{E}[l^\theta(\lambda_*^\theta, Y) - l^\theta(\lambda, Y)] \prec \mathbb{E}[l^\theta(\lambda_*^\theta, Y) - l^\theta(\widehat{\lambda}_n^\theta, Y)]\}$ and $\widehat{\Theta}_n := \{\theta \in \Theta : \mathbb{E}[l(\theta, Y) - l(\theta_*, Y)] \prec \mathbb{E}[l(\widehat{\theta}_n, Y) - l(\theta_*, Y)]\}$ are shrinking neighborhoods around λ_*^θ and θ_* containing $\widehat{\lambda}_n^\theta$ and $\widehat{\theta}_n$.

C Proofs

C.1 Theorem 3.1 and Lemma B.5

Theorem 3.1 and Lemma B.5 are simplified versions of the slightly more general Theorems C.1 and C.2, which modify Shen and Wong [1994]'s M-estimator convergence rate arguments to hold uniformly over another parameter space and accommodate estimators which are finite-sample optimal up to some stochastic remainder. Theorem C.1 is presented in C.1.1 and derives the uniform convergence rates for $\widehat{\lambda}_n^\theta$. Theorem C.2 is presented in C.1.2 and derives the rates for $\widehat{\theta}_n$. In C.1.3, we then discuss how Theorem 3.1 and Lemma B.5 follow from these results.

C.1.1 Uniform convergence rate of $\widehat{\lambda}_n^\theta$

Theorem C.1 (Uniform Convergence Rates of Sieve M-Estimators). *Let $\rho^\theta(\cdot, \cdot)$ be a pseudo-distance on Λ, possibly indexed by $\theta \in \Theta$. For the estimator $\widehat{\lambda}_n^\theta$ of 3.2, assume:*

CONDITION C1a. For some constants $A_1 > 0$ and $\alpha > 0$, and all small $\varepsilon > 0$:

$$\inf_{\{\rho^\theta(\lambda, \lambda_*^\theta) \geq \varepsilon, \lambda \in \Lambda, \theta \in \Theta\}} \mathbb{E}\left[l^\theta\left(\lambda_*^\theta, Y\right) - l^\theta(\lambda, Y)\right] \geq 2A_1 \varepsilon^{2\alpha}$$

CONDITION C1b. For some constants $A_2 > 0$ and $\beta > 0$, and all small $\varepsilon > 0$:

$$\sup_{\left\{ \rho^\theta(\lambda, \lambda_*^\theta) \leq \varepsilon, \lambda \in \Lambda, \theta \in \Theta \right\}} \mathbb{V}\left[l^\theta(\lambda, Y) - l^\theta\left(\lambda_*^\theta, Y\right) \right] \leq A_2 \varepsilon^{2\beta}$$

CONDITION C2. Let $\mathcal{F}_n = \left\{ l^\theta(\lambda, \cdot) - l^\theta\left(\pi_n \lambda_*^\theta, \cdot\right) : \lambda \in \Lambda_n, \theta \in \Theta_n \right\}$. For some $r_0 < \frac{1}{2}$, $A_3 > 0$ and all small $\varepsilon > 0$, its entropy (Def. 1) is bounded as:

$$\log \mathcal{N}\left(\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty \right) \leq A_3 n^{2r_0} \varepsilon^{-r}$$

where either $r > 0$ or $r = 0^+$, which is understood to represent $\varepsilon^{-0^+} = \log(1/\varepsilon)$.

Let $\epsilon_n := \sup_{\theta \in \Theta_n} \rho^\theta\left(\pi_n \lambda_*^\theta, \lambda_*^\theta\right) \vee \left| \mathbb{E}\left[l^\theta(\lambda_*^\theta, Y) - l(\pi_n \lambda_*^\theta, Y) \right] \right|^{1/2\alpha}$, *then*

$$\sup_{\theta \in \Theta_n} \rho^\theta\left(\widehat{\lambda}_n^\theta, \lambda_*^\theta \right) = O_\mathbb{P}\left(n^{-\tau} + \epsilon_n + \eta_n^{1/2\alpha} \right),$$

where $\tau = \tau(\alpha, \beta, r, r_0, n)$ is given by:

$$\tau = \begin{cases} \frac{1-2r_0}{2\alpha} - \frac{\log\log n}{2\alpha \log n}, & \text{if } r = 0^+, \beta \geq \alpha \\ \frac{1-2r_0}{4\alpha - 2\beta}, & \text{if } r = 0^+, \beta < \alpha \\ \frac{1-2r_0}{4\alpha - \min(\alpha,\beta)(2-r)}, & \text{if } 0 < r < 2 \\ \frac{1-2r_0}{4\alpha} - \frac{\log\log n}{2\alpha \log n}, & \text{if } r = 2 \\ \frac{1-2r_0}{2\alpha r}, & \text{if } r > 2 \end{cases}$$

And for any $f(\theta, \lambda, Y)$ satisfying C1a and C2 when $l^\theta(\lambda, Y)$ is replaced by $f(\theta, \lambda, Y)$, we can bound the empirical process as follows:

$$\sup_{\theta \in \Theta_n, \lambda \in \widehat{\Lambda}_n(\theta)} (\mathbb{E} - \mathbb{E}_n)[f(\theta, \lambda, Y) - f(\theta, \pi_n \lambda_*^\theta, Y)] = O_\mathbb{P}(n^{-\tau} + \epsilon_n + \eta_n^{1/2\alpha}) \qquad \text{(C.1)}$$

where $\widehat{\Lambda}_n(\theta)$ is defined as in Lemma B.5.

Proof. The theorem generalizes Theorem 1 of Shen and Wong [1994] such that it holds uniformly over a family of losses indexed by the parameter $\theta \in \Theta_n$, and to allow for the finite-sample optimum to hold approximately up to a possibly stochastic sequence η_n. Fortunately, the proof can remain almost identical. Shen and Wong [1994] prove the Theorem by induction, through a chaining argument. They use their Lemma 2 to derive an initial, slow rate which corresponds to the induction start, yielding the assumptions of their Lemma 3 at step $k = 2$. Next, their Lemma 3 is repeatedly applied as the induction steps until the rates of Theorem C.1 are obtained. We do not reproduce these algebraic steps, as they are the same as in Shen and Wong [1994]. Like Shen and Wong [1994], we also do not provide the proof for the induction start as it is similar, but simpler than the proof of the induction step, which we present in Lemma C.1. $\square$

Lemma C.1 (Induction Step for Theorem C.1).

Suppose Conditions C1a, C1b and C2 hold. If at Step $k-1$ we have a rate $\varepsilon_n^{(k-1)} = n^{-\alpha_{k-1}} > \max\left(n^{-(1-2r_0)/[\alpha(r+2)]}, \epsilon_n\right)$ so that

$$\mathbb{P}\left(\sup_{\theta\in\Theta_n} \rho^\theta\left(\widehat{\lambda}_n^\theta, \lambda_*^\theta\right) \geq D\varepsilon_n^{(k-1)}\right) \leq 5\left[\exp\left(-(1-\varepsilon)\max\left(D^{4\alpha}, D^{2\alpha}\right)M_1 n^{2r_0}\right) + (k-1)\exp\left(-Ln^{\delta_*}\right)\right]$$

where $\delta_* = \min\left(\frac{r+4r_0}{r+2}, \frac{\beta r(1-2r_0)}{4\alpha} + r_0\right)$ and $L = (1-\varepsilon)\min\left(M_2 D^{2\alpha}, M_3 D^{4\alpha-\beta(2-r)/2}\right)$
Then at Step k, we can find an improved rate

$$\varepsilon_n^{(k)} = \max\left(n^{-\alpha_k}, n^{-(1-2r_0)/[\alpha(r+2)]}, \epsilon_n, \eta_n^{1/2\alpha}\right)$$

where $\alpha_k = (1-2r_0)/(4\alpha) + \alpha_{k-1}\beta(2-r)/(4\alpha)$, so that

$$\mathbb{P}\left(\sup_{\theta\in\Theta_n} \rho^\theta\left(\widehat{\lambda}_n^\theta, \lambda_*^\theta\right) \geq D\varepsilon_n^{(k)}\right) \leq 5\left[\exp\left(-(1-\varepsilon)\max\left(D^{4\alpha}, D^{2\alpha}\right)M_1 n^{2r_0}\right) + k\exp\left(-Ln^{\delta_*}\right)\right]$$

And for any function $f(\theta, \lambda, Y)$ satisfying C1b and C2 when $l^\theta(\lambda, Y)$ is replaced by $f(\theta, \lambda, Y)$, and $\widehat{\Lambda}_n(\theta)$ is defined as in Lemma B.5, the same bound applies to:

$$\mathbb{P}\left(\sup_{\theta\in\Theta_n, \lambda\in\widehat{\Lambda}_n(\theta)} (\mathbb{E} - \mathbb{E}_n)[f(\theta, \lambda, Y) - f(\theta, \pi_n\lambda_*^\theta, Y)] \geq A_1\left(D\varepsilon_n^{(k)}\right)^{2\alpha}\right)$$

Proof. We assume $D > 1$ (wlog) and we only prove the case of $4\alpha \geq \beta(2-r)/2$. Let $B_n^{(i)} = \left\{D\varepsilon_n^{(i)} \leq \sup_{\theta\in\Theta_n} \rho^\theta\left(\widehat{\lambda}_n^\theta, \lambda_*^\theta\right) < D\varepsilon_n^{(i-1)}\right\}$ for $i = 2, \ldots, k$. Then

$$\mathbb{P}\left(\sup_{\theta\in\Theta_n} \rho^\theta\left(\widehat{\lambda}_n^\theta, \lambda_*^\theta\right) \geq D\varepsilon_n^{(k)}\right) \leq \mathbb{P}\left(\sup_{\theta\in\Theta_n} \rho^\theta\left(\widehat{\lambda}_n^\theta, \lambda_*^\theta\right) \geq D\varepsilon_n^{(k-1)}\right) + \mathbb{P}\left(B_n^{(k)}\right)$$

To prove the Lemma , we only need to tackle $\mathbb{P}\left(B_n^{(k)}\right)$. By Condition C1a,

$$\inf_{\left\{\rho^\theta\left(\lambda, \lambda_*^\theta\right) \geq D\varepsilon_n^{(k)}, \lambda\in\Lambda_n, \theta\in\Theta_n\right\}} \mathbb{E}\left[l^\theta\left(\pi_n\lambda_*^\theta, Y\right) - l^\theta(\lambda, Y)\right] - \eta_n$$
$$\geq 2A_1\left(D\varepsilon_n^{(k)}\right)^{2\alpha} - \sup_{\theta\in\Theta_n} \mathbb{E}\left[l^\theta\left(\lambda_*^\theta, Y\right) - l^\theta\left(\pi_n\lambda_*^\theta, Y\right)\right] - \eta_n \geq A_1\left(D\varepsilon_n^{(k)}\right)^{2\alpha}$$

The last inequality requires $A_1\left(D\varepsilon_n^{(k)}\right)^{2\alpha} - \sup_{\theta\in\Theta_n} \mathbb{E}\left[l^\theta\left(\lambda_*^\theta, Y\right) - l^\theta\left(\pi_n\lambda_*^\theta, Y\right)\right] - \eta_n > 0$. This is holds for $A_1 > 2$ (wlog), which follows from $\varepsilon_n^{(k)} \geq \epsilon_n$ and $\varepsilon_n^{(k)} \geq \eta_n^{1/2\alpha}$.

Thus

$$
\mathbb{P}\left(B_n^{(k)}\right) \leq \mathbb{P}\left(\sup_{\left\{D\varepsilon_n^{(k)}\leq\rho^\theta\left(\lambda,\lambda_*^\theta\right)<D\varepsilon_n^{(k-1)},\lambda\in\Lambda_n,\theta\in\Theta_n\right\}} \mathbb{E}_n\left[l^\theta(\lambda,Y)-l^\theta\left(\pi_n\lambda_*^\theta,Y\right)\right]\geq -\eta_n\right)
$$

$$
\leq \mathbb{P}\left(\sup_{\left\{D\varepsilon_n^{(k)}\leq\rho^\theta\left(\lambda,\lambda_*^\theta\right)<D\varepsilon_n^{(k-1)},\lambda\in\Lambda_n,\theta\in\Theta_n\right\}} n^{1/2}(\mathbb{E}_n-\mathbb{E})\left[l^\theta(\lambda,Y)-l^\theta\left(\pi_n\lambda_*^\theta,Y\right)\right]\geq A_1 n^{1/2}\left(D\varepsilon_n^{(k)}\right)^{2\alpha}\right)
$$

Let $v_k = \sup_{\left\{D\varepsilon_n^{(k)}\leq\rho^\theta\left(\lambda,\lambda_*^\theta\right)<D\varepsilon_n^{(k-1)},\lambda\in\Lambda_n,\theta\in\Theta_n\right\}} \mathrm{Var}\left(l^\theta\left(\pi_n\lambda_*^\theta,Y\right)-l^\theta(\lambda,Y)\right)$. By Condition C1b and $\varepsilon_n^{(k-1)}\geq\epsilon_n$ we get $v_k \leq 4A_2\left(D\varepsilon_n^{(k-1)}\right)^{2\beta}$. Since $\varepsilon_n^{(k)}$ satisfies

$$
\varepsilon_n^{(k)} \geq n^{-\min((1-2r_0)/(4\alpha)+\alpha_k-1\beta(2-r)/(4\alpha),(1-2r_0)/[\alpha(r+2)])}
$$

we know that $n^{1/2}\left(D\varepsilon_n^{(k)}\right)^{2\alpha} \geq \max\left(c_1 n^{-(2-r-8r_0)/[2(r+2)]}, c_2\left(D\varepsilon_n^{k-1}\right)^{2\beta(2-r)/4} n^{r_0}\right)$ for some constants $c_1 > 0$ and $c_2 > 0$. We can therefore apply Shen and Wong [1994]'s Lemma 1 and obtain:

$$
\mathbb{P}\left(B_n^{(k)}\right) \leq \exp\left(-\psi_1\left(A_1 n^{1/2}\left(D\varepsilon_n^{(k)}\right)^{2\alpha}, v_k, \mathcal{F}_n\right)\right)
$$

The behavior of $\psi_1(\cdot)$ can be analyzed via Shen and Wong [1994]'s Remark 12.
(i) If $\left(D\varepsilon_n^{(k)}\right)^{2\alpha} A_1 > 12\left(D\varepsilon_n^{(k-1)}\right)^{2\beta}$, then

$$
\psi_1\left(A_1 n^{1/2}\left(D\varepsilon_n^{(k)}\right)^{2\alpha}, v_k, \mathcal{F}_n\right) \geq \frac{3A_1}{4} n\left(D\varepsilon_n^{(k)}\right)^{2\alpha} \geq M_2 D^{2\alpha} n n^{-2(1-2r_0)/(r+2)} \geq M_2 D^{2\alpha} n^{(r+4r_0)/(r+2)}
$$

for some constant $M_2 > 0$. (ii) If $\left(D\varepsilon_n^{(k)}\right)^{2\alpha} A_1 \leq 12\left(D\varepsilon_n^{(k-1)}\right)^{2\beta}$, then

$$
\psi_1\left(A_1 n^{1/2}\left(D\varepsilon_n^{(k)}\right)^{2\alpha}, v_k, \mathcal{F}_n\right) \geq \frac{\left(A_1 n^{1/2}\left(D\varepsilon_n^{(k)}\right)^{2\alpha}\right)^2}{4\left(4A_2\right)\left(D\varepsilon_n^{(k-1)}\right)^{2\beta}}
$$

$$
\geq M_3 D^{4\alpha-\beta(2-r)/2}\left(\varepsilon_n^{(k-1)}\right)^{2\beta(2-r)/2} \frac{n^{r_0}}{\left(\varepsilon_n^{(k-1)}\right)^{2\beta}}
$$

$$
\geq M_3 D^{4\alpha-\beta(2-r)/2}\left(\varepsilon_n^{(1)}\right)^{-\beta r} n^{r_0} \geq M_3 D^{4\alpha-\beta(2-r)/2} n^{\beta r(1-2r_0)/(4\alpha)+r_0}
$$

for some $M_3 > 0$. Hence,

$$
\mathbb{P}\left(B_n^{(k)}\right) \leq \begin{cases} 5\exp\left(-(1-\varepsilon)M_2 D^{2\alpha} n^{(r+4r_0)/(r+2)}\right) & \text{if } \left(D\varepsilon_n^{(k)}\right)^{2\alpha} A_1 > 12\left(D\varepsilon_n^{(k-1)}\right)^{2\beta} \\ 5\exp\left(-(1-\varepsilon)M_3 D^{4\alpha-\beta(2-r)/2} n^{\beta r(1-2r_0)/(4\alpha)+r_0}\right) & \text{if } \left(D\varepsilon_n^{(k)}\right)^{2\alpha} A_1 \leq 12\left(D\varepsilon_n^{(k-1)}\right)^{2\beta} \end{cases}
$$

Take δ_*, L and $\varepsilon_n^{(k)}$ as defined in the Lemma, and we obtain $\mathbb{P}\left(B_n^{(k)}\right) \leq 5\exp\left(-Ln^{\delta_*}\right)$. This yields the convergence rate of $\widehat{\lambda}_n^\theta$. The statement about arbitrary $f(\theta, \lambda, Y)$ follows by applying analogous arguments (starting at the definition of v_k) to the expression $f(\theta, \pi_n\lambda_*^\theta, Y) - f(\theta, \lambda, Y)$ instead of $l^\theta(\pi_n\lambda_*^\theta, Y) - l^\theta(\lambda, Y)$. $\square$

C.1.2 Convergence of $\widehat{\theta}_n$

Theorem C.2 (Convergence Rate of A-Estimators). *Consider the family of M-estimators $\widehat{\lambda}_n^\theta$ defined in 3.2 and the A-estimator $\widehat{\theta}_n$ defined in 3.3. Assume that conditions C1b and C2 are satisfied with $\rho^\theta(\lambda, \lambda') := \left|\mathbb{E}[l^\theta(\lambda, Y) - l^\theta(\lambda', Y)]\right|$ (hence C1a is automatically satisfied with $\alpha = 1/2$). Let $\bar{\rho}(\cdot, \cdot)$ be some pseudo-distance on Θ. Assume that the following conditions are satisfied:*

CONDITION C1a' For some constants $\bar{A}_1 > 0$ and $\bar{\alpha} > 0$, and all small $\varepsilon > 0$:

$$\inf_{\{\bar{\rho}(\theta, \theta_*) \geq \varepsilon, \theta \in \Theta_n\}} \mathbb{E}\left[l(\theta, Y) - l(\theta_*, Y)\right] \geq 2\bar{A}_1\varepsilon^{2\bar{\alpha}}$$

CONDITION C1b'. For some constants $\bar{A}_2 > 0$ and $\bar{\beta} > 0$, and all small $\varepsilon > 0$:

$$\sup_{\{\bar{\rho}(\theta, \theta_*) \leq \varepsilon, \theta \in \Theta_n\}} \mathbb{V}\left[l(\theta, Y) - l(\theta_*, Y)\right] \leq \bar{A}_2\varepsilon^{2\bar{\beta}}$$

CONDITION C2'. Let $\bar{\mathcal{F}}_n = \left\{l(\theta, \pi_n\lambda_*^\theta, \cdot) - l\left(\pi_n\theta_*, \pi_n\lambda_*^{\pi_n\theta_*}, \cdot\right) : \theta \in \Theta_n\right\}$. For some $\bar{r}_0 < \frac{1}{2}$, $\bar{A}_3 > 0$ and all small $\varepsilon > 0$, its entropy (Def. 1) is bounded as:

$$\log\mathcal{N}\left(\varepsilon, \bar{\mathcal{F}}_n, \|\cdot\|_\infty\right) \leq \bar{A}_3 n^{2\bar{r}_0}\varepsilon^{-\bar{r}}$$

where either $\bar{r} > 0$ or $\bar{r} = 0^+$, which represents $\varepsilon^{-0^+} = \log(1/\varepsilon)$.

Let $\bar{\epsilon}_n := \rho(\pi_n\theta_*, \theta_*) \vee |\mathbb{E}l(\pi_n\theta_*, Y) - l(\theta_*, Y)|^{1/2\bar{\alpha}}$ be the approximation error of Θ_n. Then:

$$\bar{\rho}\left(\widehat{\theta}_n, \theta_*\right) = O_\mathbb{P}\left(n^{-\bar{\tau}} + \bar{\epsilon}_n + (\bar{\eta}_n + n^{-\tau} + \epsilon_n + \eta_n)^{1/2\bar{\alpha}}\right)$$

Where $\tau = \tau(1/2, \beta, r, r_0, n)$ and ϵ_n are defined as in Thm C.1 and $\bar{\tau} = \tau(\bar{\alpha}, \bar{\beta}, \bar{r}, \bar{r}_0, n)$. Also, for every $f(\theta, \lambda, Y)$ satisfying C2 and C3 when $l(\theta, \lambda, Y)$ is replaced by $f(\theta, \lambda, Y)$ (recall $l(\theta, Y) = l(\theta, \lambda_*^\theta, Y)$), we can bound the empirical process:

$$\sup_{\theta \in \widehat{\Theta}_n} (\mathbb{E} - \mathbb{E}_n)[f(\theta, \pi_n\lambda_*^\theta, Y) - f(\theta_*, \lambda_*^{\theta_*}, Y)] = O_\mathbb{P}\left(n^{-\bar{\tau}} + \bar{\epsilon}_n + (\bar{\eta}_n + n^{-\tau} + \epsilon_n + \eta_n)^{1/2\bar{\alpha}}\right)$$

Proof. The proof is similar to that of Theorem C.1. Again, we will only prove the induction step via Lemma D.1 in the Online Appendix, as the remaining arguments are analogous to the proof of the previous Theorem or that of Theorem 1

in Shen and Wong [1994]. The proof of Lemma D.1 largely mirrors that of Lemma C.1, but uses the results of Theorem C.1 to control the convergence of the adversary. The main additional complexity lies in properly switching back and forth between the sieve spaces and the target function spaces, when bounding the empirical process terms and variances respectively. $\square$

C.1.3 Proofs of Theorem 3.1 and Lemma B.5

Proof. To see that Theorem 3.1 and Lemma B.5 follow from the previous results, simply choose $\rho^\theta(\lambda, \lambda') = |\mathbb{E}[l^\theta(\lambda, Y) - l^\theta(\lambda', Y)]|$ and $\bar{\rho}(\theta, \theta') = |\mathbb{E}[l(\theta, Y) - l(\theta', Y)]|$, such that Conditions C1a and C1a' are automatically satisfied with $\alpha = 1/2$. Further, substitute $\gamma = 2\beta \vee 2\bar{\beta}$ such that Conditions C1b and C1b' hold by assumptions 3.7 and 3.6 for all $\varepsilon < 1 \wedge C$. Conditions C2 and C2' directly follow from 3.8, substituting $s = 2r_0 = 2\bar{r}_0$ and fixing $r = \bar{r}$. This yields Theorem 3.1, and Lemma B.5 with $e_n = 0$.

For a proof of Lemma B.5 with $e_n \neq 0$, note that Condition C1b in Theorem C.1 is only needed to verify $v_k \leq 4A_2 \left(D\varepsilon_n^{(k-1)}\right)^{2\beta}$ in the proof of Lemma C.1. Hence we can re-define $\epsilon_n \leftarrow \epsilon_n + e_n$ such that the definition of $\varepsilon_n^{(k-1)} \geq \epsilon_n$ automatically ensures $v_k \prec \left(D\varepsilon_n^{(k-1)}\right)^{2\beta}$. This change in constants does not affect the result. Analogous arguments can be applied to Lemma D.1 and thus Theorem C.2. $\square$

C.2 Theorem 3.3

Proof. The approximate Nash conditions 1.2 and 1.3 imply

$$
\begin{aligned}
O_\mathbb{P}(e_n^2) &\geq \mathbb{E}_n l(\widehat{\theta}_n, \pi_n \bar{\lambda}_n^{\widehat{\theta}_n}(\widehat{\lambda}_n), Y) - l(\pi_n \bar{\theta}(\widehat{\theta}_n), \widehat{\lambda}_n, Y) \\
&= \mathbb{E}_n l'(\widehat{\theta}_n, \widehat{\lambda}_n, Y)[\widehat{\theta}_n - \pi_n \bar{\theta}_n(\widehat{\theta}_n), \pi_n \bar{\lambda}_n^{\widehat{\theta}_n}(\widehat{\lambda}_n) - \widehat{\lambda}_n] + O_\mathbb{P}(e_n^2) \\
&= \mathbb{E}_n l'(\widehat{\theta}_n, \widehat{\lambda}_n, Y)[e_n v, e_n \lambda_*'^{\widehat{\theta}_n}[v]] + O_\mathbb{P}(e_n^2)
\end{aligned}
\tag{C.2}
$$

The second line uses Taylor's theorem. The third line substitutes the definitions of $\bar{\theta}_n, \bar{\lambda}_n^{\widehat{\theta}_n}$ and applies Condition N3. Since the signs of $v, \lambda_*'^{\widehat{\theta}_n}[v]$ are arbitrary, we may replace the inequality with an equality, which yields $O_\mathbb{P}(e_n) = \mathbb{E}_n l'(\widehat{\theta}_n, \widehat{\lambda}_n, Y)[v, \lambda_*'^{\widehat{\theta}_n}[v]]$. Adding and subtracting a few terms, we get:

$$
\begin{aligned}
\mathbb{E}_n l'(\theta_*, Y)[v] =& (\mathbb{E}_n - \mathbb{E})[l'(\widehat{\theta}_n, \widehat{\lambda}_n, Y)[v, \lambda_*'^{\widehat{\theta}_n}[v]] - l'(\theta_*, Y)[v]] \\
&+ \mathbb{E}[l'(\widehat{\theta}_n, \widehat{\lambda}_n, Y)[v, \lambda_*'^{\widehat{\theta}_n}[v]] - l'(\theta_*, Y)[v]] + O_\mathbb{P}(e_n) \\
=& \langle \widehat{\theta}_n - \theta_*, v \rangle + O_\mathbb{P}(e_n)
\end{aligned}
\tag{C.3}
$$

Where the last line uses Conditions N1 and N2. Substituting $v = v_*$, we get:

$$
\sqrt{n}\left(F(\widehat{\theta}_n) - F(\theta_*)\right) = \sqrt{n}\langle \widehat{\theta}_n - \theta_*, v_* \rangle + o_\mathbb{P}(1) = \sqrt{n}\mathbb{E}_n[l'(\theta_*, Y)[v]] + o_\mathbb{P}(1) \xrightarrow{d} \mathcal{N}(0, V)
$$

46

with $V = \mathbb{V}\left(l'(\theta_*, Y)[v]\right)$ by the standard central limit theorem. $\square$

C.3   Theorem 3.4

Proof of Theorem 3.4. Note that the regularization does not interfere with Theorem 3.2: the approximation power relative to Θ_*, Λ_* is not reduced as we remove only elements form the sieves Θ_n, Λ_n that are far away from Θ_*, Λ_*, and the regularization is sufficiently slow to guarantee that the sieves are nonempty for small enough $\epsilon > 0$. Hence Theorem 3.2 holds with $\bar{r} = 2$, yielding rates $o_\mathbb{P}(n^{-2/(2+\bar{r})}) = o_\mathbb{P}(n^{-1/2})$. Also note that the assumption $\frac{d^*}{p} \vee \frac{\bar{d}^*}{\bar{p}} < 1/4$ along with the lower bound $\underline{r} > 2/3$ ensures $\sup_{\theta \in \Theta_*} \|\theta - \pi_n\theta\|_{\bar{\mathcal{X}}} = o(n^{-1})$ and $\sup_{\lambda \in \Lambda_*} \|\lambda - \pi_n\lambda\|_{\mathcal{X}} = o(n^{-1})$. We first verify condition N1, decomposing it into two parts by adding and subtracting $l'(\theta, \lambda_*'^\theta, Y)[v_*, \lambda_*'^\theta[v_*]] = l'(\theta, Y)[v_*]$. First, we show

$$\sup_{\theta \in \widehat{\Theta}_n, \lambda \in \widehat{\Lambda}_n(\theta)} (\mathbb{E}_n - \mathbb{E})l'(\theta, Y)[v_*] - l'(\theta_*, Y)[v_*] = o_\mathbb{P}(n^{-1/2})$$

If A7ii) holds with $\mathbb{V}[l'(\theta_*, Y)[v] - l'(\theta, Y)[v]] \prec \mathbb{E}[l(\theta, Y) - l(\theta_*, Y)]$, then this can be established with Lemma B.5 for $\gamma = 1$, using the Lipschitz condition A4 and analogous arguments to those in the proof of Theorem 3.2. Lemma B.5 then yields the same $o_\mathbb{P}(n^{-1/2})$ rates as Theorem 3.2. If A7ii) instead asserts that Θ_* is $\mathbb{P}$-Donsker, the same result follows from Lemma B.4, which can be applied because the Lipschitz continuity A4 implies the L2 continuity required by the Lemma. This implies condition N1, together with:

$$\sup_{\theta \in \widehat{\Theta}_n, \lambda \in \widehat{\Lambda}_n(\theta)} (\mathbb{E}_n - \mathbb{E})l'(\theta, \lambda, Y)[v_*, \lambda_*'^\theta[v_*]] - l'(\theta, \lambda_*'^\theta, Y)[v_*, \lambda_*'^\theta[v_*]] = o_\mathbb{P}(n^{-1/2})$$

which can be established via analogous arguments and A7i). We proceed to verify condition N2. Using a similar decomposition, we note that

$$\sup_{\theta \in \widehat{\Theta}_n, \lambda \in \widehat{\Lambda}_n(\theta)} \mathbb{E}l'(\theta, \lambda, Y)[v_*, \lambda_*'^\theta[v_*]] - l'(\theta, \lambda_*'^\theta, Y)[v_*, \lambda_*'^\theta[v_*]] = O_\mathbb{P}(e_n)$$

which follows from A6i) and the $o_\mathbb{P}(n^{-1/2})$ rates of Theorem 3.2. Similarly, A6ii) implies $\sup_{\theta \in \widehat{\Theta}_n, \lambda \in \widehat{\Lambda}_n(\theta)} \mathbb{E}l'(\theta, Y)[v_*] - l'(\theta_*, Y)[v_*] - \langle \theta - \theta_*, v_* \rangle = O_\mathbb{P}(e_n)$ hence condition N2 holds. Finally, we verify condition N3. Define $\pi_*\theta := \arg\inf_{\theta' \in \Theta_*} \|\theta' - \theta\|_\infty$ as the projection onto Θ_*. Similarly, $\pi_*\lambda := \arg\inf_{\lambda' \in \Lambda_*} \|\lambda' - \lambda\|_\infty$. Due to the reguarlization, we know $\|\theta - \pi_*\theta\|_\infty = o(n^{-1})$. Therefore $\|\bar{\theta}_n(\theta) - (\pi_*\theta - e_n v_*)\|_\infty = o(n^{-1})$. By convexity of Θ_*, we have $(\pi_*\theta - e_n v_*) \in \Theta_*$ for n large enough. Given that $\sup_{\theta' \in \Theta_*} \|\theta' - \pi_n\theta'\|_\infty = o(n^{-1})$ due to $\frac{d^*}{p} \vee \frac{\bar{d}^*}{\bar{p}} < 1/4$ and our choice of $\underline{r}$, we get $\|(\pi_*\theta - e_n v_*) - \theta_n(\pi_*\theta - e_n v_*)\|_\infty = o(n^{-1})$. Taken together, these statements imply $\|\bar{\theta}_n(\theta) - \pi_n\bar{\theta}_n(\theta)\|_\infty = o(n^{-1})$. Analogous arguments yield $\|\bar{\lambda}_n^\theta(\lambda) - \pi_n\bar{\lambda}_n^\theta(\lambda)\|_\infty = o(n^{-1})$. Hence N3 holds. $\square$

D   Online Appendix

D.1   Proof of Theorem C.2

The induction step for the proof of Theorem C.2 is given by the following Lemma.

Lemma D.1 (Induction Step for Theorem C.2).
Suppose the Conditions of Theorem C.2 hold. If at Step $k-1$ we have a rate

$$\varepsilon_n^{(k-1)} = n^{-\bar{\alpha}_k - 1} > \max\left(n^{-(1-2\bar{r}_0)/[\bar{\alpha}(\bar{r}+2)]}, \bar{\epsilon}_n, \epsilon_n\right)$$

so that

$$\mathbb{P}\left(\bar{\rho}\left(\widehat{\theta}_n, \theta_*\right) \geq D\varepsilon_n^{(k-1)}\right) \leq 5\left[\exp\left(-(1-\varepsilon)\max\left(D^{4\bar{\alpha}}, D^{2\bar{\alpha}}\right)M_1 n^{2\bar{r}_0}\right) + (k-1)\exp\left(-Ln^{\delta_*}\right)\right]$$

*where $\delta_* = \min\left(\frac{r+4\bar{r}_0}{r+2}, \frac{\bar{\beta}\bar{r}(1-2\bar{r}_0)}{4\bar{\alpha}} + \bar{r}_0\right)$ and $L = (1-\varepsilon)\min\left(M_2 D^{2\bar{\alpha}}, M_3 D^{4\bar{\alpha}-\bar{\beta}(2-\bar{r})/2}\right)$,
then at Step k, we can find an improved rate*

$$\varepsilon_n^{(k)} = \max\left(n^{-\bar{\alpha}_k}, n^{-(1-2\bar{r}_0)/[\bar{\alpha}(\bar{r}+2)]}, \bar{\epsilon}_n, (\bar{\eta}_n + r_n)^{1/2\bar{\alpha}}\right)$$

where $\bar{\alpha}_k = (1-2\bar{r}_0)/(4\bar{\alpha}) + \bar{\alpha}_{k-1}\bar{\beta}(2-\bar{r})/(4\bar{\alpha})$, so that

$$\mathbb{P}\left(\bar{\rho}\left(\widehat{\theta}_n, \theta_*\right) \geq D\varepsilon_n^{(k)}\right) \leq 5\left[\exp\left(-(1-\varepsilon)\max\left(D^{4\bar{\alpha}}, D^{2\bar{\alpha}}\right)M_1 n^{2\bar{r}_0}\right) + k\exp\left(-Ln^{\delta_*}\right)\right]$$

*Furthermore, for every $f(\theta, \lambda, Y)$ satisfying Conditions C1b and C2 when $l(\theta, \lambda, Y)$
is replaced by $f(\theta, \lambda, Y)$ (recall $l(\theta, Y) = l(\theta, \lambda_*^\theta, Y)$), the same bound applies to:*

$$\mathbb{P}\left(\sup_{\theta \in \widehat{\Theta}_n}(\mathbb{E} - \mathbb{E}_n)[f(\theta, \pi_n \lambda_*^\theta, Y) - f(\theta_*, \lambda_*^{\theta_*}, Y)] \geq D\varepsilon_n^{(k)}\right)$$

Proof. As in the proof of Lemma C.1 we assume $D > 1$ (wlog) and we only prove the case of $4\bar{\alpha} \geq \bar{\beta}(2-\bar{r})/2$. Let $B_n^{(i)} = \left\{D\varepsilon_n^{(i)} \leq \bar{\rho}\left(\widehat{\theta}_n, \theta_*\right) < D\varepsilon_n^{(i-1)}\right\}$ for $i = 2, \ldots, k$. As before, we only need to bound $\mathbb{P}\left(B_n^{(k)}\right)$. To this end, it will be useful to define

$$r_n := \sup_{\theta \in \Theta_n} \mathbb{E}[l(\theta, \lambda_*^\theta, Y) - l(\theta, \widehat{\lambda}_n^\theta, Y)] \vee \sup_{\theta \in \Theta_n}(\mathbb{E} - \mathbb{E}_n)[l(\theta, \widehat{\lambda}_n^\theta, Y) - l(\theta, \pi_n \lambda_*^\theta, Y)]$$

such that Theorem C.1 implies $r_n = O_{\mathbb{P}}(n^\tau + \epsilon_n + \eta_n)$, which also implies, by definition of r_n and ϵ_n: $\sup_{\theta \in \Theta_n}\left|\mathbb{E}_n[l(\theta, \widehat{\lambda}_n^\theta, Y) - l(\theta, \pi_n \lambda_*^\theta, Y)]\right| \leq r_n + \epsilon_n \leq 2r_n$. Together with 3.3 this yields:

$$\mathbb{E}_n[l(\widehat{\theta}_n, \pi_n \lambda_*^{\widehat{\theta}_n}, Y) - l(\theta, \pi_n \lambda_*^\theta, Y)] \leq 2r_n + \bar{\eta}_n \quad \forall \theta \in \Theta_n \tag{D.1}$$

48

By Condition C1a', we can therefore bound:

$$\inf_{\theta\in\Theta_n:\bar\rho(\theta,\theta_*)\geq D\varepsilon_n^{(k)}}\mathbb{E}\left[l(\theta,\pi_n\lambda_*^\theta,Y)-l(\pi_n\theta_*,\pi_n\lambda_*^{\pi_n\theta_*},Y)\right]-\bar\eta_n-2r_n$$

$$=\inf_{\theta\in\Theta_n:\bar\rho(\theta,\theta_*)\geq D\varepsilon_n^{(k)}}\mathbb{E}[l(\theta,Y)-l(\theta_*,Y)]+\mathbb{E}[l(\theta_*,Y)-l(\pi_n\theta_*,Y)]-\bar\eta_n-2r_n$$

$$+\mathbb{E}[l(\theta,\pi_n\lambda_*^\theta,Y)-l\,(\theta,Y)]+\mathbb{E}[l(\pi_n\theta_*,Y)-l(\pi_n\theta_*,\pi_n\lambda_*^{\pi_n\theta_*},Y)]$$

$$\geq 2\bar A_1\left(D\varepsilon_n^{(k)}\right)^{2\bar\alpha}-\bar\epsilon_n^{2\bar\alpha}-\bar\eta_n-r_n-\epsilon_n-\epsilon_n\geq\bar A_1\left(D\varepsilon_n^{(k)}\right)^{2\bar\alpha}$$

$$\text{(D.2)}$$

Where the last line used C1a', the definition of the approximation errors, assumed large enough $\bar A_1$ (wlog) and used various lower-bounds implied by the definition of $\varepsilon_n^{(k)}$. Together with D.1, this yields:

$$\mathbb{P}\left(B_n^{(k)}\right)\leq\mathbb{P}\left(\sup_{\left\{D\varepsilon_n^{(k)}\leq\bar\rho(\theta,\theta_*)<D\varepsilon_n^{(k-1)},\theta\in\Theta_n\right\}}\mathbb{E}_n\left[l(\pi_n\theta_*,\pi_n\lambda_*^{\pi_n\theta_*},Y)-l\,(\theta,\pi_n\lambda_*^\theta,Y)\right]\geq-\bar\eta_n-2r_n\right)$$

$$\leq\mathbb{P}\left(\sup_{\substack{\theta\in\Theta_n\\D\varepsilon_n^{(k)}\leq\bar\rho(\theta,\theta_*)<D\varepsilon_n^{(k-1)}}}n^{1/2}(\mathbb{E}_n-\mathbb{E})\left[l(\pi_n\theta_*,\pi_n\lambda_*^{\pi_n\theta_*},Y)-l\,(\theta,\pi_n\lambda_*^\theta,Y)\right]\geq A_1n^{1/2}\left(D\varepsilon_n^{(k)}\right)^{2\bar\alpha}\right)$$

Let $v_k=\sup_{\left\{D\varepsilon_n^{(k)}\leq\bar\rho(\theta,\theta_*)<D\varepsilon_n^{(k-1)},\theta\in\Theta_n\right\}}\mathbb{V}\left[l(\pi_n\theta_*,\pi_n\lambda_*^{\pi_n\theta_*},Y)-l\,\left(\theta,\pi_n\lambda_*^\theta,Y\right)\right]$. To bound v_k, we add and subtract terms and apply the Cauchy-Schwartz inequality:

$$\mathbb{V}[l(\pi_n\theta_*,\pi_n\lambda_*^{\pi_n\theta_*},Y)-l(\theta,\pi_n\lambda_*^\theta,Y)]$$

$$\leq 3\mathbb{V}[l(\pi_n\theta_*,\lambda_*^{\pi_n\theta_*},Y)-l(\theta,\lambda_*^\theta,Y)]+3\mathbb{V}[l(\pi_n\theta_*,\pi_n\lambda_*^{\pi_n\theta_*},Y)-l(\theta,\pi_n\lambda_*^\theta,Y)]$$

$$\prec\mathbb{V}[l(\theta,Y)-l(\theta_*,Y)]+\mathbb{V}[l(\pi_n\theta_*,Y)-l(\theta_*,Y)]+\mathbb{V}[l^{\pi_n\theta_*}(\pi_n\lambda_*^{\pi_n\theta_*},Y)-l^{\pi_n\theta_*}(\lambda_*^{\pi_n\theta_*},Y)]$$

$$+\mathbb{V}[l^\theta(\lambda_*^\theta,Y)-l^\theta(\pi_n\lambda_*^\theta,Y)]$$

By Conditions C1b and C1b', and since $\varepsilon_n^{(k-1)}\geq\bar\epsilon_n$, we obtain $v_k\leq 4\bar A_2\left(D\varepsilon_n^{(k-1)}\right)^{2\bar\beta}$, assuming $\bar A_2$ is large enough (wlog). The remaining arguments are unchanged from the proof of Lemma C.1, which eventually yields: $\mathbb{P}\left(B_n^{(k)}\right)\leq 5\exp\left(-Ln^{\delta_*}\right)$. This completes the proof for the convergence rate of $\widehat\theta_n$. To prove the statement about arbitrary $f(\theta,\lambda,Y)$ satisfying C1b and C2, simply repeat the arguments of the previous proof (starting at the definition of v_k) with $l(\theta,\lambda,Y)$ replaced by $f(\theta,\lambda,Y)$. $\qquad\square$

D.2 Proof of Proposition 4.1

Proof. The first order conditions for λ in 2.1 yield the optimal population adversary $\lambda_*^\theta(y)$:

$$\lambda_*^\theta(y)\mathrm{d}\mathbb{P}_\theta(y) - f_*'(\lambda_*^\theta(y))\mathrm{d}\mathbb{P}(y) \overset{!}{=} 0 \iff \lambda_*^\theta(y) = f'\left(\frac{\mathrm{d}\mathbb{P}_\theta(y)}{\mathrm{d}\mathbb{P}(y)}\right) \qquad \text{(D.3)}$$

We verify the conditions of Theorem 3.2. The Lipschitz condition A1 can be verified by writing $l(\theta, \lambda, Y) = \int \lambda(y)\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}}(y)\mathrm{d}\mathbb{P}(y) - f_*(\lambda(Y))$ and using the Lipschitzness of $\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}}$ in θ and that of f_* (which follows from boundedness of Λ and differentiability of f). Towards A2, apply a 2nd order Taylor expansion (with mean value reminder) of $D_f(\mathbb{P}_\theta\|\mathbb{P})$ at $\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}} = \frac{\mathrm{d}\mathbb{P}_{\theta*}}{\mathrm{d}\mathbb{P}}$ in direction of $\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}}$, which yields for some $\widetilde\theta \in \Theta$ on a path from θ_* to θ:

$$D_f(\mathbb{P}_\theta\|\mathbb{P}) = \int f\left(\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}}\right)\mathrm{d}\mathbb{P} = \frac{1}{2}\int \left(\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}} - \frac{\mathrm{d}\mathbb{P}_{\theta*}}{\mathrm{d}\mathbb{P}}\right)^2 f''\left(\frac{\mathrm{d}\mathbb{P}_{\widetilde\theta}}{\mathrm{d}\mathbb{P}}\right)\mathrm{d}\mathbb{P} \asymp \left\|\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}} - \frac{\mathrm{d}\mathbb{P}_{\theta*}}{\mathrm{d}\mathbb{P}}\right\|_{L^2(\mathcal{Y})}^2$$

where the last step follows from strict positivity and boundedness of $f''\left(\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}}\right)$ wp1. Further note that

$$\mathbb{V}[l(\theta, Y) - l(\theta_*, Y)] = \mathbb{V}\left[f_*\left(\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}}(Y)\right) - f_*\left(\frac{\mathrm{d}\mathbb{P}_{\theta*}}{\mathrm{d}\mathbb{P}}(Y)\right)\right] \prec \left\|\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}} - \frac{\mathrm{d}\mathbb{P}_{\theta*}}{\mathrm{d}\mathbb{P}}\right\|_{L^2(\mathcal{Y})}^2$$

due to Lipschitzness of f_*. Also note that Lipschitzness of $\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}}$ in θ implies

$$\left\|\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}} - \frac{\mathrm{d}\mathbb{P}_{\theta*}}{\mathrm{d}\mathbb{P}}\right\|_{L^2(\mathcal{Y})}^2 \prec \left\|\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}} - \frac{\mathrm{d}\mathbb{P}_{\theta*}}{\mathrm{d}\mathbb{P}}\right\|_{\widetilde{\mathcal{Y}}}^2 + \mathbb{P}(y \in \widetilde{\mathcal{Y}}) \prec \|\theta - \theta_*\|_\infty^2 + \mathbb{P}(y \in \widetilde{\mathcal{Y}})$$

Taken together, this implies A2. A3 can be verified analogously, starting with a Taylor expansion yielding $\mathbb{E}[l(\theta, \lambda, Y) - l(\theta, \lambda_*^\theta, Y)] = -\int f_*''(\widetilde\lambda)(\lambda - \lambda_*^\theta)^2\mathrm{d}\mathbb{P} \asymp -\|\lambda - \lambda_*^\theta\|_{L^2(\mathcal{Y})}^2$ for some $\widetilde\lambda \in \Lambda$ on a path from λ_*^θ to λ. $\qquad\square$

D.3 Proof of Proposition 4.2

Proof. First, we verify the conditions of Theorem 3.4. Note that

$$l'(\theta, \lambda, Y_i)[v, w] = \int \lambda(y)\nabla_{\theta\to v}\frac{\mathrm{d}\mathbb{P}_\theta(y)}{\mathrm{d}\mathbb{P}(y)}\mathrm{d}\mathbb{P}(y) + \int w(y)\frac{\mathrm{d}\mathbb{P}_\theta(y)}{\mathrm{d}\mathbb{P}(y)}\mathrm{d}\mathbb{P}(y) - f_*'(\lambda(Y_i))w(Y_i)$$

The Lipschitz condition A4 is therefore satisfied by the Lipschitzness of f_*', $\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}}$ and that of $\nabla_{\theta\to v}\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}}$. Towards A6 i), we apply a Taylor expansion with 2nd order mean-

value reminder around $\lambda = \lambda_*^\theta$, yielding for some $\widetilde{\lambda}^\theta$ on a path between λ_*^θ and λ:

$$\mathbb{E}[l'(\theta, \lambda, Y_i)[v, \lambda_*'^\theta[v]] - l'(\theta, \lambda_*^\theta, Y_i)[v, \lambda_*'^\theta[v]]]$$

$$= \nabla_{\theta \to v} \nabla_{\lambda_*^\theta \to \lambda - \lambda_*^\theta} \mathbb{E}[l(\theta, \lambda_*^\theta, Y)] + \mathbb{E}[f_*''(\widetilde{\lambda}^\theta(Y))(\lambda - \lambda_*^\theta)^2 w(Y)] \prec \mathbb{E}[(\lambda - \lambda_*^\theta)^2]$$

where we used $\nabla_{\lambda_*^\theta \to \lambda - \lambda_*^\theta} \mathbb{E}[l(\theta, \lambda_*^\theta, Y)] \equiv 0$ to get rid of the first-order term. The last line follows from the boundedness (in absolute value) of $f_*''(\cdot)$ and $w(\cdot)$ on their compact support. Hence A6i) is satisfied. A7i) follows from:

$$\mathbb{V}[l'(\theta, \lambda, Y_i)[v, w] - l'(\theta, \lambda_*^\theta, Y_i)[v, w]] = \mathbb{V}[(f_*'(\lambda(Y)) - f_*'(\lambda_*^\theta(Y)))w(Y)] \prec \mathbb{E}[(\lambda - \lambda_*^\theta)^2]$$

where the last step used the Lipschitzness of f_*' and again the boundedness of $w(\cdot)$. Assumption A7 ii) is satisfied since $\times_*$ is Donsker by assumption. Finally, we verify A6 ii). Applying the mean value theorem twice, we get that for some $\widetilde{\theta}, \widetilde{\theta}'$ on a path between θ_* and θ, $\mathbb{E}[l'(\theta, Y)[v_*] - l'(\theta_*, Y)[v_*] - \langle \theta - \theta_*, v_* \rangle = \nabla_{\widetilde{\theta} \to \widetilde{\theta}' - \theta_*} \nabla_{\widetilde{\theta} \to \theta - \theta_*} \mathbb{E}[l'(\widetilde{\theta}, Y)[v_*]]$ which is dominated by $\mathbb{E}[l(\theta, Y)] = D_f(\mathbb{P}_\theta \| \mathbb{P}_{\theta_*})$ via the last assumption stated in the proposition. Therefore A6ii) is satisfied, and Theorem 2.4 applies. $\square$

D.4 Proof of Proposition 4.6

Proof. Note that $V = \nabla_{\theta_*} \nabla_{\theta_*'} \mathbb{E}[l(\theta_*, Y)] = \mathbb{V}[\nabla_{\theta_*} l(\theta_*, Y)]$. We verify the conditions of Theorem 3.4, for $v_* = V^{-1} \zeta$, such that its conclusion becomes $\sqrt{n} \langle \theta - \theta_*, v_* \rangle = \sqrt{n}(\theta - \theta_*)' \zeta \to \mathcal{N}(0, \zeta' V^{-1} \zeta)$, which yields the Proposition via the Cramér-Wold device. Note that $l'(\theta, \lambda, Y)[v, w] = v'd(X, \theta)'\lambda(Z) + m(X, \theta)'w(Z) - \frac{1}{2}\lambda(Z)'w(Z)$ Hence assumption A4 follows from boundedness and Lipschitzness of $d(X, \cdot), m(X, \cdot)$. A5 holds by assumption. To verify A6i), notice that:

$$\mathbb{E}[l'(\theta, \lambda, Y)[v_*, \lambda_*'^\theta[v_*]] - l'(\theta, Y)] = \mathbb{E}\left[\left(v_*'d(X, \theta) - \frac{1}{2}\lambda_*'^\theta[v_*](Z)\right)(\lambda - \lambda_*^\theta)(Z)\right] = 0$$

where the last equality used the fact that $\mathbb{E}[v_*'d(X, \theta)|Z] = \frac{1}{2}\lambda_*'^\theta[v_*](Z)$. Towards Assumption A6ii), note that we can apply a first-order Taylor expansion with mean-value reminder twice, which yields for some $\bar{\theta}, \widetilde{\theta}$ on a path between θ_* and θ:

$$\mathbb{E}[l'(\theta, Y)[v_*] - l'(\theta_*, Y)[v_*] - \langle \theta - \theta_*, v_* \rangle] = (\bar{\theta} - \theta_*)' \nabla_{\widetilde{\theta}} \nabla_{\widetilde{\theta}'} \mathbb{E}[l'(\widetilde{\theta}, Y)][v_*](\theta - \theta_*) \prec \|\theta - \theta_*\|_2^2$$

Given the identification assumption 2.5, we can use a second-order Taylor expansion with mean-value reminder to show that $\|\theta - \theta_*\|_2^2 \asymp \mathbb{E}[l(\theta, Y) - l(\theta_*, Y)]$, which then yields A6ii). Similarly, we can show A7ii) via a mean-value reminder:

$$\mathbb{E}[(l'(\theta, Y)[v] - l'(\theta_*, Y)[v])^2] = \mathbb{E}\left[\left((\theta - \theta_*)' \nabla_{\widetilde{\theta}} l'(\widetilde{\theta}, Y)[v]\right)^2\right] \prec \|\theta - \theta_*\|_2^2$$

We similarly can establish A7i) via

$$\mathbb{V}[l'(\theta, \lambda, Y)[v, \lambda'^\theta_*[v_*]] - l'(\theta, \lambda^\theta_*, Y)[v, \lambda'^\theta_*[v_*]]] \prec \mathbb{E}[(\lambda(Z) - \lambda^\theta_*(Z))^2] \asymp \mathbb{E}[l(\theta, \lambda, Y) - l(\theta, \lambda^\theta_*, Y)]$$

$\square$

D.5 Proof of Proposition 4.4

Proof. A0 holds by assumption, and condition A1 is implied by Lipschitzness of V_θ, P_θ in θ. Continuity of V_θ, P_θ, compactness of Θ and A0 further imply that there is some $0 \le M < \infty$ such that

$$\sup_{s,a,s_+,\theta} |R_\theta(s,a) + \beta V(s^+) - V(s) - \log P_\theta(a|s)| \le M$$

Given that $\lambda^{\theta_*}_* \equiv 0 \implies l(\theta_*, Y) = 0$, condition A2 then follows from

$$\mathbb{V}[l(\theta, Y) - l(\theta_*, Y)] \prec \mathbb{E}[l(\theta, Y)^2] \le 2M^2 \mathbb{E}[\lambda^\theta_*(s,a)^2] \prec \mathbb{E}[l(\theta, Y)] = \mathbb{E}[l(\theta, Y) - l(\theta_*, Y)]$$

as well as $\mathbb{E}[l(\theta, Y) - l(\theta_*, Y)] \prec \mathbb{E}[(\lambda^\theta_*(s,a) - \lambda^{\theta_*}_*(s,a))^2 | (s,a) \in \widetilde{\mathcal{X}}] + \mathbb{P}((s,a) \notin \widetilde{\mathcal{X}}) \prec \|\theta - \theta_*\|^2_{\widetilde{\mathcal{X}}} + \mathbb{P}((s,a) \notin \widetilde{\mathcal{X}})$, $\forall \widetilde{\mathcal{X}} \subset \bar{\mathcal{X}}$. A3 can be established analogously, hence the conclusions of Theorem 3.2 hold. $\square$

D.6 Proof of Proposition 4.7

Proof. Note that $\lambda^\theta_*(x) = \theta(x) - \theta_*(x)$ follows from the first order conditions of the adversary. Condition A0 is satisfied by assumption, and A1 follows from Lipschitzness of $m(Y, \cdot)$ and boundedness. Lipschitzness of $m(\theta, \lambda(x))$ in $\lambda(x)$ and boundedness imply that $l(\theta, \lambda, Y) = m(Y, \lambda) - \theta(x)\lambda(x) - \lambda(x)^2/2$ is Lipschitz in $\lambda(x)$. This implies

$$\mathbb{V}[l(\theta, \lambda, Y) - l(\theta, \lambda^\theta_*, Y)] \prec \mathbb{E}[(\lambda(x) - \lambda^\theta_*(x))^2]$$

and together with $\lambda^\theta_*(x) = \theta(x) - \theta_*(x)$, it yields:

$$\mathbb{V}[l(\theta_*, \lambda^{\theta_*}_*, Y) - l(\theta, \lambda^\theta_*, Y)] \prec \mathbb{E}[(\theta(x) - \theta_*(x))^2]$$

Both bounds imply A3 and A2 respectively. The result follows because the loss $\mathbb{E}[l(\theta, Y)]$ is proportional to the squared L2 norm of $\theta - \theta_*$. $\square$