Deep Generalized Method of Moments for Instrumental Variable Analysis

Andrew Bennett[*]
Cornell University
awb222@cornell.edu

Nathan Kallus[*]
Cornell University
kallus@cornell.edu

Tobias Schnabel[*]
Microsoft Research
tbs49@cornell.edu

1 Introduction

Unlike standard supervised learning that models correlations, causal inference seeks to predict the effect of counterfactual interventions not seen in the data. For example, when wanting to estimate the effect of adherence to a prescription of β-blockers on the prevention of heart disease, supervised learning may overestimate the true effect because good adherence is also strongly correlated with health consciousness and therefore with good heart health [2]. Figure 1 shows a simple example of this type and demonstrates how a standard neural network (in blue) fails to correctly estimate the true treatment response curve (in orange) in a toy example.

One approach to account for this is by adjusting for all confounding factors that cause the dependence, such as via matching [12, 13] or regression, potentially using neural networks [8, 14]. However, this requires that we actually observe *all* confounders so that treatment is as-if random after conditioning on observables. This would mean that in the β-blocker example, we would need to perfectly measure *all* latent factors that determine both an individual's adherence decision and their general healthfulness which is often not possible in practice.

Instrumental variables (IVs) provide an alternative approach to causal-effect identification. If we can find a latent experiment in another variable (the instrument) that influences the treatment (*i.e.*, is relevant) and does not directly affect the outcome (*i.e.*, satisfies exclusion), then we can use this to infer causal effects [1]. In the β-blocker example [2], the authors used co-pay cost as an IV. An important direction of research for IV analysis is to develop methods that can effectively handle complex causal relationships and complex variables like images that necessitate more flexible models like neural networks [7].

In this paper, we tackle this though a new method called DeepGMM that builds upon the optimally-weighted Generalized Method of Moments (GMM) [5], a widely popular method in econometrics that uses the moment conditions implied by the IV model to efficiently estimate causal parameters. Leveraging a new variational reformulation of the efficient GMM with optimal weights, we develop a flexible framework, DeepGMM, for doing IV estimation with neural networks. In contrast to existing approaches, DeepGMM is suited for high-dimensional treatments X and instruments Z, as well as for complex causal and interaction effects. DeepGMM is given by the solution to a smooth game between a prediction function and critic function. We prove that approximate equilibria provide consistent estimates of the true causal parameters, and provide some brief experiments demonstrating that DeepGMM's performance is state-of-the-art on standard benchmarks.

2 Setup and Notation

We assume that our data is generated by

$$Y = g_0(X) + \epsilon, \tag{1}$$

[*]Alphabetical order.

Abstract for: Smooth Games Optimization and Machine Learning Workshop (NeurIPS 2019), Vancouver, Canada.

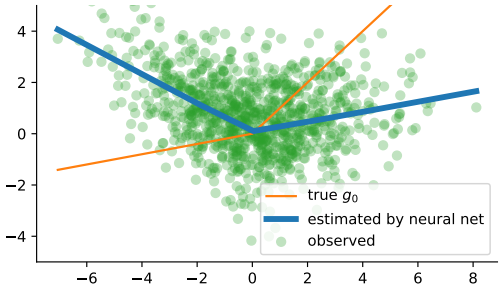

Figure 1: A toy example in which standard supervised learning fails to identify the true response function $g_0(X) = \max(\frac{X}{5}, X)$. Data was generated via $Y = g_0(X) - 2\epsilon + \eta$, $X = Z + 2\epsilon$. All other variables are standard normal.

where the residual ϵ has zero mean and finite variance, *i.e.*, $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}\left[\epsilon^2\right] < \infty$. However, different to standard supervised learning, we allow for the residual ϵ and X to be correlated, $\mathbb{E}[\epsilon \mid X] \neq 0$, *i.e.*, X can be endogenous, and therefore $g_0(X) \neq \mathbb{E}[Y \mid X]$. We also assume that we have access to an instrument Z satisfying $\mathbb{E}[\epsilon \mid Z] = 0$.

Moreover, Z should be relevant, *i.e.* $\mathbb{P}(X \mid Z) \neq \mathbb{P}(X)$. Our goal is to identify the causal response function $g_0(\cdot)$ from a parametrized family of functions $G = \{g(\cdot;\theta) \,:\, \theta \in \Theta\}$. Examples are linear functions $g(x;\theta) = \theta^T\phi(x)$, neural networks where θ represent weights, and non-parametric classes with infinite-dimensional θ. For convenience, let $\theta_0 \in \Theta$ be such that $g_0(\cdot) = g(\cdot;\theta_0)$.

2.1 Existing methods for IV estimation

Two-stage methods. One strategy to identifying g_0 is based on noting that our instrument satisfies

$$\mathbb{E}[Y \mid Z] = \mathbb{E}[g_0(X) \mid Z] = \int g_0(x) d\mathbb{P}(X = x \mid Z)\,. \tag{2}$$

If we let $g(x;\theta) = \theta^T\phi(x)$ this becomes $\mathbb{E}[Y \mid Z] = \theta_0^T\mathbb{E}[\phi(X) \mid Z]$. The two-stage least squares (2SLS) method [1, §4.1.1] first fits $\mathbb{E}[\phi(X) \mid Z]$ by least-squares regression of $\phi(X)$ on Z (with Z possibly transformed) and then estimates $\hat{\theta}^{\text{2SLS}}$ as the coefficient in the regression of Y on $\mathbb{E}[\phi(X) \mid Z]$. This, however, fails when one does not know a sufficient basis $\phi(x)$ for $g(x,\theta_0)$. [3, 11] propose non-parametric methods for expanding such a basis but such approaches are limited to low-dimensional settings. [7] instead propose DeepIV, which estimates the conditional density $\mathbb{P}(X = x \mid Z)$ by flexible neural-network-parametrized Gaussian mixtures. This may be limited in settings with high-dimensional X and can suffer from the non-orthogonality of MLE under any misspecification, known as the "forbidden regression" issue [1, §4.6.1]

Moment methods. The generalized method of moments (GMM) instead leverages the moment conditions satisfied by θ_0. Given functions $f_1, \ldots, f_m$ we have $\mathbb{E}[f_j(Z)\epsilon] = 0$, giving us

$$\psi(f_1;\theta_0) = \cdots = \psi(f_m;\theta_0) = 0, \quad \text{where} \quad \psi(f;\theta) = \mathbb{E}[f(Z)(Y - g(X;\theta))]\,. \tag{3}$$

A usual assumption when using GMM is that the m moment conditions in Eq. (3) are sufficient to uniquely pin down (identify) θ_0. To estimate θ_0, GMM considers these moments' empirical counterparts, $\psi_n(f;\theta) = \frac{1}{n}\sum_{i=1}^n f(Z_i)(Y_i - g(X_i;\theta))$, and seeks to make all of them small simultaneously, measured by their Euclidean norm $\|v\|^2 = v^Tv$:

$$\hat{\theta}^{\text{GMM}} \in \operatorname*{argmin}_{\theta\in\Theta} \|(\psi_n(f_1;\theta), \ldots, \psi_n(f_m;\theta))\|^2\,. \tag{4}$$

Other vector norms are possible. In particular, a celebrated result [5] shows that (with finitely-many moments), using the following norm in Eq. (4) will yield *minimal* asymptotic variance (efficiency) for any consistent estimate $\tilde{\theta}$ of θ_0:

$$\|v\|^2 = v^TC_{\tilde{\theta}}^{-1}v, \quad \text{where} \quad [C_\theta]_{jk} = \frac{1}{n}\sum_{i=1}^n f_j(Z_i)f_k(Z_i)(Y_i - g(X_i;\theta))^2. \tag{5}$$

Examples of this are the two-step, iterative, and continuously updating GMM estimators [6]. We generically refer to the GMM estimator given in Eq. (4) using the norm given in Eq. (5) as *optimally-weighted GMM* (OWGMM), or $\hat{\theta}^{\text{OWGMM}}$.

3 Methodology

3.1 Reformulating OWGMM

Let us start by reinterpreting OWGMM. Consider the following objective function:

$$\Psi_n(\theta; \mathcal{F}, \tilde{\theta}) = \sup_{f \in \mathcal{F}} \psi_n(f; \theta) - \frac{1}{4n} \sum_{i=1}^{n} f(Z_i) h(Z_i)(Y_i - g(X_i; \theta))^2 \tag{6}$$

Lemma 1. *Let $\|v\|$ be the GMM optimally-weighted norm, given by weighting the empirical moments by the inverse square root of the covariance matrix, and let $\mathcal{F} = \operatorname{span}(f_1, \ldots, f_m)$. Then*

$$\|(\psi_n(f_1; \theta), \ldots, \psi_n(f_m; \theta))\|^2 = \Psi_n(\theta; \mathcal{F}, \tilde{\theta}).$$

In other words, Lemma 1 provides a variational formulation of the objective function of OWGMM.

3.2 DeepGMM

Given our reformulation above in Lemma 1, our approach is to simply replace the set $\mathcal{F}$ with a more flexible set of functions. Namely we let $\mathcal{F} = \{f(z; \tau) : \tau \in \mathcal{T}\}$ be the class of all neural networks of a given architecture with varying weights τ (but *not* their span). Using a rich class of moment conditions allows us to learn correspondingly a rich g_0. We therefore similarly let $\mathcal{G} = \{g(x; \theta) : \theta \in \Theta\}$ be the class of all neural networks of a given architecture with varying weights θ.

Given these choices, we let $\hat{\theta}^{\text{DeepGMM}}$ be the minimizer in Θ of $\Psi_n(\theta; \mathcal{F}, \tilde{\theta})$ for any, potentially data-driven, choice $\tilde{\theta}$.[2] Since this is no longer closed form, we formulate our algorithm in terms of solving a smooth zero-sum game. Formally, our estimator is defined as:

$$\hat{\theta}^{\text{DeepGMM}} \in \operatorname*{argmin}_{\theta \in \Theta} \sup_{\tau \in \mathcal{T}} U_{\tilde{\theta}}(\theta, \tau) \tag{7}$$

$$\text{where} \quad U_{\tilde{\theta}}(\theta, \tau) = \frac{1}{n} \sum_{i=1}^{n} f(Z_i; \tau)(Y_i - g(X_i; \theta)) - \frac{1}{4n} \sum_{i=1}^{n} f^2(Z_i; \tau)(Y_i - g(X_i; \tilde{\theta}))^2.$$

Since evaluation is linear, for any $\tilde{\theta}$, the game's payoff function $U_{\tilde{\theta}}(\theta, \tau)$ is convex-concave in the functions $g(\cdot; \theta)$ and $f(\cdot; \tau)$, although it may not be convex-concave in θ and τ as is usually the case when we parametrize functions using neural networks.[3]

3.3 Consistency

Our first main result is a proof that DeepGMM provides consistent estimation of the paramters θ_0. We prove this consistency theorem based on some generic assumptions based on some generic bounded-complexity classes $\mathcal{F}, \mathcal{G}$; not necessarily neural networks.

Assumption 1 (Identification)**.** θ_0 is the unique $\theta \in \Theta$ satisfying $\psi(f; \theta) = 0$ for all $f \in \mathcal{F}$.

Assumption 2 (Bounded complexity)**.** $\mathcal{F}$ and $\mathcal{G}$ have vanishing Rademacher complexities.

Assumption 3 (Absolutely star shaped)**.** For every $f \in \mathcal{F}$ and $|\lambda| \leq 1$, we have $\lambda f \in \mathcal{F}$.

Assumption 4 (Continuity)**.** For any x, $g(x; \theta), f(x; \tau)$ are continuous in θ, τ, respectively.

Assumption 5 (Boundededness)**.** Y, $\sup_{\theta \in \Theta} |g(X; \theta)|$, $\sup_{\tau \in \mathcal{T}} |f(Z; \tau)|$ are all bounded RVs.

[2] In practice we use $\tilde{\theta} = \hat{\theta}$, since this is a consistent estimate for θ_0 as discussed below, but we treat it as a constant when computing gradients.

[3] Solving Eq. (7) can be done with any of a variety of smooth game playing algorithms. In practice we use OAdam [4], a variant of Adam designed for optimizing smooth game problems.

Scenario	DirectNN	Vanilla2SLS	Poly2SLS	GMM+NN	AGMM	DeepIV	Our Method
sin	$.26 \pm .00$	$.09 \pm .00$	$.04 \pm .00$	$.08 \pm .00$	$.11 \pm .01$	$.06 \pm .00$	$.02 \pm .00$
step	$.21 \pm .00$	$.03 \pm .00$	$.03 \pm .00$	$.06 \pm .00$	$.06 \pm .01$	$.03 \pm .00$	$.01 \pm .00$
abs	$.21 \pm .00$	$.23 \pm .00$	$.04 \pm .00$	$.14 \pm .02$	$.17 \pm .03$	$.10 \pm .00$	$.03 \pm .01$
linear	$.09 \pm .00$	$.00 \pm .00$	$.00 \pm .00$	$.06 \pm .01$	$.03 \pm .00$	$.04 \pm .00$	$.01 \pm .00$

Table 1: Experiment results: Test MSE averaged across ten runs with standard errors.

Theorem 1. *Suppose Assumptions 1 to 5 hold. Let $\tilde{\theta}_n$ by any data-dependent sequence with a limit in probability. Let $\hat{\theta}_n, \hat{\tau}_n$ be any approximate equilibrium in the game Eq.* (7)*,* i.e.*,*

$$\sup_{\tau \in \mathcal{T}} U_{\tilde{\theta}_n}(\hat{\theta}_n, \tau) - o_p(1) \leq U_{\tilde{\theta}_n}(\hat{\theta}_n, \hat{\tau}_n) \leq \inf_{\theta} U_{\tilde{\theta}_n}(\theta, \hat{\tau}_n) + o_p(1).$$

Then $\hat{\theta}_n \to \theta_0$ in probability.

4 Experiments

Finally we present some brief experimental results for DeepGMM[4] against a set of baselines. We evaluate performance of an estimated $\hat{g}$ by MSE against the true g_0. We use the following baselines:

1. DirectNN: Predicts Y from X using a neural network with standard least squares loss.
2. Vanilla2SLS: Standard two-stage least squares on raw X, Z.
3. Poly2SLS: Both X and Z are expanded via polynomial features, and then 2SLS is done via ridge regressions at each stage, with hyperparameters picked via cross-validation at each stage.
4. GMM+NN: Here, we combine OWGMM with a neural network $g(x; \theta)$. We solve Eq. (4) over network weights θ using Adam. We employ optimal weighting, Eq. (5), by iterated GMM [6].
5. AGMM [10]: Uses the publicly available implementation[5] of the Adversarial Generalized Method of Moments, which performs no-regret learning on the one-step GMM objective Eq. (4) with norm $\|\cdot\|_\infty$ and an additional jitter step on the moment conditions after each epoch.
6. DeepIV [7]: We use the latest implementation that was released as part of the econML package.[6]

Note that GMM+NN relies on being provided moment conditions. When we follow AGMM [10] and expand Z via RBF kernels around 10 centroids returned from a Gaussian Mixture model applied to the Z data.

Similar to [10], we generated data via the following process:

$$\begin{array}{ll} Y = g_0(X) + e + \delta & X = 0.5\ Z_1 + 0.5\ e + \gamma \\ Z \sim \text{Uniform}([-3, 3]^2) & e \sim \mathcal{N}(0, 1),\ \ \gamma, \delta \sim\ \mathcal{N}(0, 0.1) \end{array}$$

In other words, only the first instrument has an effect on X, and e is the confounder breaking independence of X and the residual $Y - g_0(X)$. We keep this data generating process fixed, but vary the true response function g_0 between the following cases:

sin: $g_0(x) = \sin(x)$ **step**: $g_0(x) = \mathbb{1}_{\{x \geq 0\}}$ **abs**: $g_0(x) = |x|$ **linear**: $g_0(x) = x$

We sample $n = 2000$ points for train, validation, and test sets each. Table 1 shows the corresponding MSE over the test set. We note that our method appears to obtain the best performance in non-linear scenarios. We also obtained results that our method obtains the strongest performance on additional scenarios where the X and/or Z values are high-dimensional images, which are witheld from this extended abstract for space purposes.

Conclusions. We presented DeepGMM as a way to deal with IV analysis with high-dimensional variables and complex relationships. The method was based on a new variational reformulation of GMM with optimal weights with the aim of handling many moments and was formulated as the solution to a smooth zero-sum game. Our empirical experiments showed that the method is very adaptable, competing with the best tuned existing method in standard benchmark scenarios.

[4] Our implementation of DeepGMM is publicly available at `https://github.com/CausalML/DeepGMM`.

[5] `https://github.com/vsyrgkanis/adversarial_gmm`

[6] `https://github.com/microsoft/EconML`

References

[1] J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press, 2008.

[2] J. A. Cole, H. Norman, L. B. Weatherby, and A. M. Walker. Drug copayment and adherence in chronic heart failure: effect on cost and outcomes. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 26(8):1157–1164, 2006.

[3] S. Darolles, Y. Fan, J.-P. Florens, and E. Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

[4] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

[5] L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, pages 1029–1054, 1982.

[6] L. P. Hansen, J. Heaton, and A. Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, 1996.

[7] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Deep iv: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1414–1423. JMLR. org, 2017.

[8] F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.

[9] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[10] G. Lewis and V. Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018.

[11] W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

[12] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[13] D. B. Rubin. Matching to remove bias in observational studies. *Biometrics*, pages 159–183, 1973.

[14] U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085, 2017.

A Omitted Proofs

Proof of Lemma 1. First note that since $\|v\|^2 = v^T C_{\tilde{\theta}}^{-1} v$, the associated dual norm is $\|v\|_*^2 = v^T C_{\tilde{\theta}} v$. Next, define as shorthand $\boldsymbol{\psi}$ as shorthand for $(\psi_n(f_1;\theta),\ldots,\psi_n(f_m;\theta))$. It follows from the definition of the dual norm that $\|\boldsymbol{\psi}\| = \sup_{\|v\|_* \leq 1} v^T \boldsymbol{\psi}$. Therefore we have:

$$\begin{aligned}\|\boldsymbol{\psi}\|^2 &= \sup_{\|v\|_* \leq \|\boldsymbol{\psi}\|} v^T \boldsymbol{\psi} \\ &= \sup_{v^T C_{\tilde{\theta}} v \leq \|\boldsymbol{\psi}\|^2} v^T \boldsymbol{\psi}\end{aligned}$$

The Lagrangian of this optimization problem is given by:

$$\mathcal{L}(v,\lambda) = v^T \boldsymbol{\psi} + \lambda(v^T C_{\tilde{\theta}} v - \|\boldsymbol{\psi}\|^2)$$

Taking the derivative of this with respect to v shows us that when $\lambda < 0$, this quantity is maximized by $v = -\frac{1}{2\lambda} C_{\tilde{\theta}}^{-1} \boldsymbol{\psi}$. In addition we clearly have strong duality for this problem by Slater's condition whenever $\|\boldsymbol{\psi}\| > 0$ (since in this case $v = 0$ is a feasible interior point). This therefore gives us the following dual formulation for $\|\boldsymbol{\psi}\|^2$:

$$\begin{aligned}\|\boldsymbol{\psi}\|^2 &= \inf_{\lambda<0} -\frac{1}{2\lambda}\|\boldsymbol{\psi}\|^2 + \lambda(\frac{1}{4\lambda^2}\|\boldsymbol{\psi}\|^2 - \|\boldsymbol{\psi}\|^2) \\ &= \inf_{\lambda<0} -\frac{1}{4\lambda}\|\boldsymbol{\psi}\|^2 - \lambda\|\boldsymbol{\psi}\|^2\end{aligned}$$

Taking derivative with respect to λ we can see that this is minimized by setting $\lambda = -\frac{1}{2}$. Given this and strong duality, we know it must be the case that $\|\boldsymbol{\psi}\|^2 = \sup_v \mathcal{L}(v,-\frac{1}{2}) = \sup_v v^T\psi - \frac{1}{2}v^T C_{\tilde{\theta}} v + \frac{1}{2}\|\boldsymbol{\psi}\|^2$. Rearranging terms and doing a change of variables $v \leftarrow 2v$ gives us the identity:

$$\|\boldsymbol{\psi}\|^2 = \sup_v v^T \boldsymbol{\psi} + \frac{1}{4} v^T C_{\tilde{\theta}} v$$

Finally, we can note that any vector $v \in \mathbb{R}^m$ corresponds to some $f \in \text{span}(\mathcal{F})$, such that $f = \sum_i v_i f_i$, and according to this notation we have $v^T\boldsymbol{\psi} = \psi_n(f;\theta)$ and $v^T C_{\tilde{\theta}} v = \mathcal{C}_{\tilde{\theta}}(f,f)$. Therefore our required result follows directly from the previous identity. □

Proof of Theorem 1. Define $m(\theta,\tau,\tilde{\theta}) = f(Z;\tau)(Y - g(X;\theta) - \frac{1}{4}f(Z;\tau)^2(Y - g(X;\tilde{\theta}))^2$, $M(\theta) = \sup_{\tau\in\mathcal{T}} \mathbb{E}[m(\theta,\tau,\tilde{\theta}]$, and $M_n(\theta) = \sup_{\tau\in\mathcal{T}} \mathbb{E}_n[m(\theta,\tau,\tilde{\theta}_n]$, where $\mathbb{E}_n$ refers to the empirical measure (average over the n data points) and $\tilde{\theta}_n \to_p \tilde{\theta}$. We will proceed by proving the following three conditions, and then proving our results in terms of these conditions:

1. $\sup_\theta |M_n(\theta) - M(\theta)| \to_p 0$
2. for every $\delta > 0$ we have $\inf_{d(\theta,\theta_0)\geq\delta} M(\theta) > M(\theta_0)$
3. $M_n(\hat{\theta}_n) \leq M_n(\theta_0) + o_p(1)$

We will proceed by proving these conditions one by one. For the first, we can derive the inequality:

$$\begin{aligned}&\sup_\theta |M_n(\theta) - M(\theta)| \\ &\quad= \sup_\theta \left|\sup_\tau \mathbb{E}_n[m(\theta,\tau,\tilde{\theta}_n)] - \sup_\tau \mathbb{E}[m(\theta,\tau,\tilde{\theta})]\right| \\ &\quad\leq \sup_{\theta,\tau} \left|\mathbb{E}_n[m(\theta,\tau,\tilde{\theta}_n)] - \mathbb{E}[m(\theta,\tau,\tilde{\theta})]\right| \\ &\quad\leq \sup_{\theta,\tau} \left|\mathbb{E}_n[m(\theta,\tau,\tilde{\theta}_n)] - \mathbb{E}[m(\theta,\tau,\tilde{\theta}_n)]\right| + \sup_{\theta,\tau} \left|\mathbb{E}[m(\theta,\tau,\tilde{\theta}_n)] - \mathbb{E}[m(\theta,\tau,\tilde{\theta})]\right| \\ &\quad\leq \sup_{\theta_1,\theta_2,\tau} |\mathbb{E}_n[m(\theta_1,\tau,\theta_2)] - \mathbb{E}[m(\theta_1,\tau,\theta_2)]| + \sup_{\theta,\tau} \left|\mathbb{E}[m(\theta,\tau,\tilde{\theta}_n)] - \mathbb{E}[m(\theta,\tau,\tilde{\theta})]\right|\end{aligned}$$

Next we will bound these two terms separately, which we will term B_1 and B_2. For the first term, we can derive the following bound, where ϵ_i are iid Rademacher random variables, $m_i(\theta, \tau, \tilde{\theta}_n) = f(Z_i;\tau)(Y_i - g(X_i;\theta) - \frac{1}{4}f(Z_i;\tau)^2(Y_i - g(X_i;\tilde{\theta}))^2$, and $m_i'(\theta, \tau, \tilde{\theta}_n')$ are shadow variables:

$$\begin{aligned}
\mathbb{E}[B_1] &= \mathbb{E}\left[\sup_{\theta_1,\theta_2,\tau}\left|\frac{1}{n}\sum_i m_i(\theta_1,\tau,\theta_2) - \mathbb{E}[m_i'(\theta_1,\tau,\theta_2')]\right|\right] \\
&\leq \mathbb{E}\left[\sup_{\theta_1,\theta_2,\tau}\left|\frac{1}{n}\sum_i m_i(\theta_1,\tau,\theta_2) - m_i'(\theta_1,\tau,\theta_2')\right|\right] \\
&= \mathbb{E}\left[\sup_{\theta_1,\theta_2,\tau}\left|\frac{1}{n}\sum_i \epsilon_i(m_i(\theta_1,\tau,\theta_2) - m_i'(\theta_1,\tau,\theta_2'))\right|\right] \\
&\leq 2\mathbb{E}\left[\sup_{\theta_1,\theta_2,\tau}\left|\frac{1}{n}\sum_i \epsilon_i m_i(\theta_1,\tau,\theta_2)\right|\right] \\
&\leq 2\mathbb{E}\left[\sup_{\theta,\tau}\left|\frac{1}{n}\sum_i \epsilon_i f(Z_i;\tau)(Y_i - g(X_i;\theta)\right|\right] \\
&\qquad + \frac{1}{2}\mathbb{E}\left[\sup_{\theta,\tau}\left|\frac{1}{n}\sum_i \epsilon_i f(Z_i;\tau)^2(Y_i - g(X_i;\theta))^2\right|\right] \\
&\leq \mathbb{E}\left[\sup_{\theta,\tau}\left|\frac{1}{n}\sum_i \epsilon_i f(Z_i;\tau)^2\right|\right] + \mathbb{E}\left[\sup_{\theta,\tau}\left|\frac{1}{n}\sum_i \epsilon_i (Y_i - g(X_i;\theta)^2\right|\right] \\
&\qquad + \frac{1}{4}\mathbb{E}\left[\sup_{\theta,\tau}\left|\frac{1}{n}\sum_i \epsilon_i f(Z_i;\tau)^4\right|\right] + \frac{1}{4}\mathbb{E}\left[\sup_{\theta,\tau}\left|\frac{1}{n}\sum_i \epsilon_i (Y_i - g(X_i;\theta))^4\right|\right]
\end{aligned}$$

Note that in the final inequality we apply the inequality $xy \leq 0.5(x^2 + y^2)$. Now given Assumption 5, the functions that map ($f(Z_i;\tau)$ and $g(X_i;\theta)$) to the summands in each term are Lipschitz. Now for any function class $\mathcal{F}$ and L-Lipschitz function ϕ we have $\mathcal{R}_n(\phi \circ \mathcal{F}) \leq L\mathcal{R}_n(\mathcal{F})$, where $\mathcal{R}_n(\mathcal{F})$ is the Rademacher complexity of class $\mathcal{F}$ [9, Thm. 4.12]. Therefore we have

$$\mathbb{E}[B_1] \leq L(\mathcal{R}_n(\mathcal{G}) + \mathcal{R}_n(\mathcal{F})),$$

for some constant L. Thus given Assumption 2 it must be case that $\mathbb{E}[B_1] \to 0$. Now let B_1' be some recalculation of B_1 where we are allowed to edit the i'th X, Z, and Y values. Then given Assumption 5 we can derive the following bounded differences inequality:

$$\begin{aligned}
\sup_{X_{1:n},Z_{1:n},Y_{1:n},X_i',Z_i',Y_i'} |B_1 - B_1'| &\leq \sup_{\theta_1,\theta_2,\tau,X_{1:n},Z_{1:n},Y_{1:n},X_i',Z_i',Y_i'} \left|\frac{1}{n}\left(m_i(\theta_1,\tau,\theta_2) - m_i'(\theta_1,\tau,\theta_2)\right)\right| \\
&\leq \frac{c}{n}
\end{aligned}$$

for some constant c. Therefore from McDiarmid's Inequality we have $P(|B_1 - \mathbb{E}[B_1]| \geq \epsilon) \leq 2\exp\left(-\frac{2n\epsilon^2}{c^2}\right)$. Putting this and the previous result for $\mathbb{E}[B_1]$ together we get $B_1 \to_p 0$.

Next, define $\omega_n = \left|(Y - g(X;\tilde{\theta}_n))^2 - (Y - g(X;\tilde{\theta}))^2\right|$. Recall that from the premise of the theorem we have $\tilde{\theta}_n \to_p \tilde{\theta}$. Then by Slutsky's Theorem, the Continuous Mapping Theorem, and Assumption 4 we have $\omega_n = o_p(1)$. Given this we can bound B_2 as follows:

$$\begin{aligned}
B_2 &= \sup_{\theta,\tau}\left|\mathbb{E}[m(\theta,\tau,\tilde{\theta}_n)] - \mathbb{E}[m(\theta,\tau,\tilde{\theta})]\right| \\
&= \frac{1}{4}\sup_{\theta,\tau}\left|\mathbb{E}[f(Z;\tau)^2(Y - g(X;\tilde{\theta}_n))^2] - \mathbb{E}[f(Z;\tau)^2(Y - g(X;\tilde{\theta}))^2]\right| \\
&\leq \frac{1}{4}\sup_{\theta,\tau}\left|\mathbb{E}[f(Z;\tau)^2(Y - g(X;\tilde{\theta}))^2] - \mathbb{E}[f(Z;\tau)^2(Y - g(X;\tilde{\theta}))^2]\right| + \frac{1}{4}\sup_{\tau}\left|\mathbb{E}[f(Z;\tau)^2\omega_n]\right| \\
&= \frac{1}{4}\sup_{\tau}\left|\mathbb{E}[f(Z;\tau)^2\omega_n]\right|
\end{aligned}$$

Now we know from Assumption 5 that $f(Z;\tau)$ is uniformly bounded, so it follows that $B_2 \leq \frac{b}{4}\mathbb{E}[|\omega_n|]$ for some constant b. Next we can note, again based on our boundedness assumption, that ω_n is uniformly bounded. Therefore it follows from the Lebesgue Dominated Convergence Theorem that $\mathbb{E}[|\omega_n|] \to 0$. Thus we know that both B_1 and B_2 converge, so we have proven the first of the three conditions, that $\sup_\theta |M_n(\theta) - M(\theta)|$ converges in probability to zero.

For the second condition we will first prove that $M(\theta_0)$ is the unique minimizer of $M(\theta)$. Clearly by Assumptions 1 and 3 we have that θ_0 is the unique minimizer of $\sup_\tau \mathbb{E}[f(Z;\tau)(Y - g(X;\theta)]$, since it sets this quantity to zero, and by these assumptions any other value of θ must have at least one τ that can be played in response that makes this expectation strictly positive. Now we can see that $M(\theta_0) = 0$ also, since $M(\theta_0) = \sup_\tau -\frac{1}{4}f(Z;\tau)^2(Y - g(X;\tilde{\theta}))^2$, and the inside of the supremum is clearly non-positive but can be set to zero using the zero function for f, which is allowed given Assumption 3. Furthermore, for any other $\theta' \neq \theta_0$, let f' be some function in $\mathcal{F}$ such that $\mathbb{E}[f(Z)(Y - g(X;\theta')] > 0$. If we have $\mathbb{E}[f'(Z)^2(Y - g(X;\tilde{\theta}))^2] = 0$ then it follows immediately that $M(\theta') > 0$. Otherwise, consider the function $\lambda f'$ for arbitrary $0 < \lambda < 1$. Since by Assumption 3 this function is also contained in $\mathcal{F}$, it follows that:

$$\begin{aligned} M(\theta') &= \sup_{f \in \mathcal{F}} \mathbb{E}[f(Z)(Y - g(X;\theta')) - \frac{1}{4}f(Z)^2(Y - g(X;\tilde{\theta}))^2] \\ &\geq \lambda \mathbb{E}[f'(Z)(Y - g(X;\theta'))] - \frac{\lambda^2}{4}\mathbb{E}[f'(Z)^2(Y - g(X;\tilde{\theta}))^2] \end{aligned}$$

This expression is a quadratic in λ that is clearly positive when λ is sufficiently small, so therefore it still follows that $M(\theta') > 0$.

Given this, we will prove the second condition by contradiction. If this were false, then for some $\delta > 0$ we would have that $\inf_{\theta \in B(\theta_0,\delta)} M(\theta) = M(\theta_0)$, where $B(\theta_0, \delta) = \{\theta \mid d(\theta, \theta_0) \geq \delta\}$. This is because from Assumption 1 we know θ_0 is the unique minimizer of $M(\theta)$. Given this there must exist some sequence $(\theta_1, \theta_2, \ldots)$ in $B(\theta_0, \delta)$ satisfying $M(\theta_n) \to M(\theta_0)$. Now by construction $B(\theta_0, \delta)$ is closed, and the corresponding limit parameters $\theta^* = \lim_{n\to\infty} \theta_n \in B(\theta_0, \delta)$ must satisfy $M(\theta^*) = M(\theta_0)$, since given Assumption 4 $M(\theta)$ is clearly a continuous function of θ so we can swap function application and limit. However $d(\theta^*, \theta_0) \geq \delta > 0$, so $\theta^* \neq \theta_0$. This contradicts the fact that θ_0 is the unique minimizer of $M(\theta)$, so we have proven the second condition.

Finally, for the third condition we will use the fact that by assumption $\hat{\theta}_n$ satisfies the approximate equilibrium condition:

$$\sup_{\tau \in \mathcal{T}} \mathbb{E}_n m(\hat{\theta}_n, \tau, \tilde{\theta}_n) - o_p(1) \leq \mathbb{E}_n m(\hat{\theta}_n, \hat{\tau}_n, \tilde{\theta}_n) \leq \inf_\theta \mathbb{E}_n m(\theta, \hat{\tau}_n, \tilde{\theta}_n) + o_p(1)$$

Now by definition $M_n(\hat{\theta}_n) = \sup_{\tau \in \mathcal{T}} \mathbb{E}_n m(\hat{\theta}_n, \tau, \tilde{\theta}_n)$. Therefore,

$$\inf_\theta \mathbb{E}_n m(\theta, \hat{\tau}_n, \tilde{\theta}_n) \leq \inf_\theta \sup_\tau \mathbb{E}_n m(\theta, \tau, \tilde{\theta}_n) = \inf_\theta M_n(\theta) \leq M_n(\theta_0).$$

Thus we have

$$M_n(\hat{\theta}_n) - o_p(1) \leq \mathbb{E}_n m(\hat{\theta}_n, \hat{\tau}_n, \tilde{\theta}_n) \leq M_n(\theta_0) + o_p(1).$$

At this point we have proven all three conditions stated at the start of the proof. For the final part we can first note that from the first and third conditions it easily follows that $M_n(\hat{\theta}_n) \leq M(\theta_0) + o_p(1)$, since $|M_n(\theta_0) - M(\theta_0)| \to_p 0$. Therefore we have:

$$\begin{aligned} M(\hat{\theta}_n) - M(\theta_0) &\leq M(\hat{\theta}_n) - M_n(\hat{\theta}_n) + o_p(1) \\ &\leq \sup_\theta \left| M(\hat{\theta}) - M_n(\hat{\theta}) \right| + o_p(1) \\ &\leq o_p(1) \end{aligned}$$

Next, define $\eta(\delta) = \inf_{d(\theta,\theta_0) \geq \delta} M(\theta) - M(\theta_0)$. Now by definition of η we know that whenever $d(\hat{\theta}_n, \theta) \geq \delta$ we have $M(\hat{\theta}_n) - M(\theta_0) \geq \eta(\delta)$. Therefore $\mathbb{P}[d(\hat{\theta}_n, \theta) \geq \delta] \leq \mathbb{P}[M(\hat{\theta}_n) - M(\theta_0) \geq \eta(\delta)]$. Now since for every $\delta > 0$ we have $\eta(\delta) > 0$ from the second condition, and we know $M(\hat{\theta}_n) - M(\theta_0) = o_p(1)$, we have that for every $\delta > 0$ the RHS probability converges to zero. Thus $d(\hat{\theta}_n, \theta_0) = o_p(1)$, so we can conclude that $\hat{\theta}_n \to_p \theta_0$. □

B Additional Methodology Details

B.1 Hyperparameter Optimization Procedure

We provide more details here about the hyperparameter optimization procedure described in **??**. Let m be the total number of hyperparameter choices under consider. Then for each candidate set of hyperparameters $\gamma_i \in \{\gamma_1, \ldots, \gamma_m\}$ we run our learning algorithm for a fixed number of epochs using γ_i, training it on our train partition. Every k_{eval} epochs we save the current parameters $\hat{\tau}$ and $\hat{\theta}$ at that epoch. This gives, for each hyperparameter choice γ_i, a finite set of f functions $\hat{\mathcal{F}}_i$, and a finite set of θ values $\hat{\Theta}_i$.

Now, define the set of functions $\hat{\mathcal{F}} = \cup_{i=1}^{m} \hat{\mathcal{F}}_i$. We define our approximation of our variational objective as

$$\hat{\Psi}_n(\theta) = \Psi_n(\theta; \hat{\mathcal{F}}, \theta),$$

where Ψ_n is as defined in Eq. (6). Note that this means for every θ we wish to evaluate we choose to approximate $\tilde{\theta}$ using that value of θ.

Given this, we finally choose the set of hyperparameters γ_i whose corresponding trajectory of parameter values $\hat{\Theta}_i$ minimizes the objective function $\min_{\theta \in \hat{\Theta}_i} \hat{\Psi}_n(\theta)$, calculated on the validation data. Note that this objective function is meant to approximate the value of the variational objective we would have obtained if we performed learning with that set of hyperparameters using early stopping.

Note that in practice, since we only ever calculate $\hat{\Psi}_n$ on the validation data, and we only ever use Θ_i in optimizing the above objective function on the validation data, instead of saving the actual parameter values $\hat{\tau}$ and $\hat{\theta}$ we can instead save vectors $f(Z_{\text{val}}, \tau)$ and $g(X_{\text{val}}, \theta)$, where Z_{val} and X_{val} are the vectors of Z and X values respectively in our validation data. This makes our methodology tractable when working with very complex deep neural networks.