AEA Papers and Proceedings 2021, 111: 621–625
https://doi.org/10.1257/pandp.20211037

Adversarial Inference Is Efficient[†]

By Tetsuya Kaji, Elena Manresa, and Guillaume A. Pouliot*

Inference methods based on simulation are popular in economics. Indeed, many theoretically motivated models can be represented as parametrized functions from which it is easy to simulate data.[1]

We consider simulation-based methods that produce as a point estimate the coefficient value at which a parametrized and identified economic model simulates data that are close, according to some notion of distance, to sampled data.

In this paper, we study properties of the *adversarial framework*, introduced in Kaji, Manresa, and Pouliot (2020). The estimators it comprises are inspired from *generative adversarial networks* (GAN), a machine learning method developed by Goodfellow et al. (2014) to generate realistic images.

The adversarial estimation framework is a minimax game between two components—a *discriminator* and a *generator*—over classification accuracy. The discriminator is an algorithm that classifies observed data from simulated data; given a data point, it delivers a probability that the point is drawn from the same distribution as the observed data, as opposed to the simulated data. Its objective is to maximize the accuracy of classification. The generator is an algorithm that simulates data; its objective is to find the parameter value for which the discriminator is least able to distinguish simulated data from real data.

We take the generative model to be derived from an economic model and the discriminator to be a classifier, a binary outcome model whose outcome variable has value one if the data are real and zero if simulated. Different choices of classifiers induce different estimators, and the framework allows us to leverage the good performance of machine learning classifiers. It raises the question: does a more accurate classifier imply more statistically efficient point estimation?

In the following, we present evidence that adversarial inference with an oracle classifier is statistically efficient. We also study the finite sample properties of the adversarial estimation framework for the autoregressive parameter of a linear dynamic fixed effects panel data model with Gaussian errors. We compare different estimators in terms of bias and root mean squared error (RMSE). Unlike maximum likelihood, but similarly as other minimum distance estimators, the adversarial estimators do not suffer from the incidental parameter bias. In addition, in our simulations, using as discriminator a one-hidden-layer neural network delivers the estimates with the smallest RMSE.

I. Methodology

Let $\mathcal{Y}_n = \{Y_i\}_{i=1}^n$ be n observed data points, which are identically and independently distributed. Assume we have a fully parametrized model in terms of a finite dimensional vector $\theta \in \Theta$, for which it is possible to obtain simulations $\mathcal{Y}_m^\theta = \{Y_i^\theta\}_{i=1}^m$. Let G be the simulation model derived from economic theory. To be sure, simulations for a candidate $\theta \in \Theta$ are generated as $Y^\theta = G(\theta, Z)$ for a random draw Z from a known distribution $\tilde{P}_0$.

Let $X = h(Y)$ be d transformations of the original data that the researcher selects, for which we give concrete examples in the dynamic panel data context in Section III.

Let $\{X_i\}_{i=1}^n$ be the n transformations of the data corresponding to the original sample, which are i.i.d. with respect to a distribution P_0, and let $\{X_i^\theta\}_{i=1}^m$ be m analogous transformations corresponding to m simulated observations, distributed with respect to the

* Kaji: Booth School of Business, University of Chicago (email: tkaji@chicagobooth.edu); Manresa: Department of Economics, New York University (email: elena.manresa@nyu.edu); Pouliot: Harris School of Public Policy, University of Chicago (email: guillaumepouliot@uchicago.edu).

[†] Go to https://doi.org/10.1257/pandp.20211037 to visit the article page for additional materials and author disclosure statement(s).

[1] Nonetheless, it is often the case that the simulation functions are not available in closed form.

distribution P_θ.[2] We assume correct specification; i.e., there exists $\theta_0 \in \Theta$ such that $P_{\theta_0} = P_0$.[3]

The classifier is a function $D : X \to [0,1]$ such that for given X, $D(X)$ is a measure of confidence that X is distributed according to the same law as the observed data; a greater value of $D(X)$ means greater confidence. Let $\mathcal{D}$ be the class of classification functions, e.g., a class of neural networks, random forests, etc.

A. *General Case*

The *adversarial estimator* is defined by the following minimax problem:

$$(1)\quad \hat{\theta} = \arg\min_{\theta\in\Theta} \max_{D\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n}\Big[\log\big(D(X_i)\big)\Big] + \frac{1}{m}\sum_{i=1}^{m}\Big[\log\big(1 - D\big(X_i^\theta\big)\big)\Big].$$

For the sake of intuition, let $n = m$. Then, for a fixed θ, the inner maximization is the maximum likelihood criterion of a binary outcome model, the classifier, with outcome one if X is real data and zero if it is simulated. The outer minimization in θ aims to generate synthetic data X^θ for which the best discriminator has the worst performance.

B. *Logistic Discriminator Case*

The special case of the logistic regression classifier is important. It is both easy to characterize and closely related to the simulated method of moments. Furthermore, it performs well in simulations, as detailed in Section III.

We consider the optimization problem

$$\min_\theta \max_\beta \frac{1}{n}\sum_{i=1}^{n}\log\big(D\big(X_i^T\beta\big)\big) + \frac{1}{m}\sum_{i=1}^{\tilde{n}}\log\big(1 - D\big(X_i^{\theta T}\beta\big)\big),$$

where $D(t) = 1/(1+e^{-t})$ and β is the vector of coefficients of the logistic regression. Given any value of θ, we denote by $\hat{\beta}(\theta)$ the logistic classifier likelihood estimator. Hence, we can rewrite the optimization problem as

$$(2)\quad \min_\theta \frac{1}{n}\sum_{i=1}^{n}\log\Big(D\big(X_i^T\hat{\beta}(\theta)\big)\Big) + \frac{1}{m}\sum_{i=1}^{m}\log\Big(1 - D\big(X_i^{\theta T}\hat{\beta}(\theta)\big)\Big).$$

It is natural to ask if this is an a priori principled method. We exhibit its first-order condition and asymptotic variance to argue that indeed it is.

When $\theta = \theta^0$, the first-order condition of the inner maximization problem yields

$$\frac{1}{n}\sum_{i=1}^{n}X_i = \frac{1}{m}\sum_{i=1}^{m}X_i^{\theta_0} + o_p(1),$$

where we can see that the estimator solves the moment matching problem using the features of the logistic classifier as the vector of moments.

Provided that G is differentiable with respect to θ, elementary calculations show that

$$\sqrt{n}\big(\hat{\theta} - \theta_0\big) = M^{-1}\sqrt{n}\Big(\hat{\beta}\big(\hat{\theta}\big) - 0\Big)A \to N\Big(0, \operatorname*{plim}_{n,m\to\infty}\big(A^T R^{-1} A\big)^{-1}\Big),$$

where

$$A = \left(\frac{1}{m}\sum_{i=1}^{m} D\big(X_{\theta_0,i}^T\hat{\beta}(\theta_0)\big)\frac{\partial X_{\theta_0,i}}{\partial\theta}\right),$$

and

$$R = \left(\frac{1}{n}\sum_{i=1}^{n} D'\big(X_i^T\hat{\beta}(\theta_0)\big)X_i\cdot X_i^T + \frac{1}{m}\sum_{i=1}^{m} D'\big(X_{\theta_0,i}^T\hat{\beta}(\theta_0)\big)X_{\theta_0,i}\cdot X_{\theta_0,i}^T\right),$$

which is full rank provided conditions on X_i and $X_i^{\theta_0}$.

We now inspect the form of the asymptotic variance. The Hessian of the logistic classifier, R, contributes proportionally to the variance of the adversarial estimator, while A, a measure of the sensitivity of the simulation function with respect to θ, is inversely proportional. The features X with most variability are those on which the classifier can rely least in order to distinguish

[2] The method can also accommodate conditional models. In that case, the data is an i.i.d. sample $\{(y_i,x_i)\}_{i=1}^n$, where y_i are outcome variables and x_i are conditioning variables. The model is a conditional one, also parametrized in terms of a finite-dimensional $\theta \in \Theta$, for which it is possible to obtain simulations given x as $\{(y_i^\theta,\tilde{x}_i)\}_{i=1}^m$, where $\tilde{x}_i$ terms are drawn from the sample distribution of $\{x_i\}_{i=1}^n$. Then, the vector $X = h(y,x)$ is a function of both y and x.

[3] The misspecified case is treated in Kaji, Manresa, and Pouliot (2020).

synthetic from true data. Correspondingly, these are the features that contribute most to the variability of the adversarial estimator.

Under appropriate regularity conditions, the minimizer of (2) is asymptotically equivalent, to first order, to the simulated method of moments estimator using as moments $(1/n)\sum_{i=1}^{n} X_i$, with optimal weighting, as characterized in Gourieroux, Monfort, and Renault (1993). The proof of this equivalence result can be found in the supplementary Appendix of Kaji, Manresa, and Pouliot (2020).

II. Theory

Heuristically speaking, the method elects as its point estimate the $\hat{\theta}$, which "confuses" the classifier D the most by presenting it with data $\{X_i^{\theta}\}_{i=1}^{m}$ whose empirical distribution is seemingly indistinguishable from that of $\{X_i\}_{i=1}^{n}$. This begs the question: does a good classifier, one that is hard to "confuse," make for a more accurate point estimate $\hat{\theta}$? In the following, we show that among all consistent estimators of θ_0, adversarial inference using the classifier solving the population classification problem in D achieves the greatest precision. Specifically, it has the smallest possible asymptotic variance among consistent estimators.

The population counterpart of (1) is

$$\arg\min_{\theta\in\Theta} \max_{D\in\mathcal{D}} E_{X\sim P_0}\left[\log\left(D(X)\right)\right] + E_{X^{\theta}\sim P_{\theta}}\left[\log\left(1-D\left(X^{\theta}\right)\right)\right].$$

If the feasible set $\mathcal{D}$ is unrestricted, the optimum classification function for the inner maximization is known to be

$$D_{\theta}^{*}(X) = \frac{p_0(X)}{p_0(X)+p_{\theta}(X)},$$

where p_0 and p_{θ} are densities corresponding to P_0 and P_{θ}, respectively (Goodfellow et al. 2014, Proposition 1). Note that the objective function with this discriminator is the Jensen-Shannon divergence between P_0 and P_{θ}. The discriminator D_{θ}^{*} implements Bayes' Rule, and this optimal discriminator delivers efficient point estimation.

PROPOSITION 1: *Suppose P_0 and P_{θ} have densities p_0 and p_{θ}, respectively, for all θ. Assume that $p_{\theta}(x)$ and $G(\theta,z)$ are both twice differentiable in both their arguments. Suppose that $D = D_{\theta}^{*}$. Suppose that $n/m \to 0$ and that $\hat{\theta}$ converges in probability to θ_0. Further suppose that the Hessian is strictly positive definite, hence*

$$V \equiv E\left[\left.\frac{\partial^2 l_{\theta}}{\partial\theta^2}\right|_{\theta_0}\right]^{-1} > 0, \quad \text{where } l_{\theta} = \log p_{\theta}.$$

Then,

$$\sqrt{n}\left(\hat{\theta}_{GAN}-\theta_0\right) \xrightarrow{d} N(0,V).$$

The proof is deferred to the Appendix.

Proposition 1 pertains to the idealized case in which the discriminator is optimal. The choice of the optimal yet infeasible population classifier as a plug-in is purposeful; it distills the analysis and focuses the exposition on efficiency as opposed to convergence and consistency.

The result may be extended to accommodate feasible discriminators. Kaji, Manresa, and Pouliot (2020) consider a growing sequence of classes of discriminators and provide sufficient conditions under which the efficiency result carries through.

We now offer some intuition for Proposition 1. One can verify that the first-order optimality condition of the inner maximization problem, at $\theta = \theta_0$, may be given as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_{\theta}}{\partial\theta}(X_i) = \frac{1}{m}\sum_{i=1}^{n}\frac{\partial l_{\theta}}{\partial\theta}\left(X_i^{\theta}\right) + o_p(1).$$

Hence, adversarial inference with classifier D_{θ}^{*} is implicitly matching the score functions evaluated at the true and synthetic data, implying that θ^{*} is selected such that the true and synthetic datasets are indistinguishable according to the likelihood principle (Casella and Berger 2002).

III. Finite Sample Performance

We report the results of some simulation experiments examining the relative performance of different estimators for the linear fixed effects dynamic panel data model:

$$y_{it} = \alpha_i + \rho\, y_{it-1} + \varepsilon_{it}, \tag{3}$$

where $\varepsilon_{it} \overset{\text{iid}}{\sim} N(0,1)$. The initial condition is drawn independently as

$$y_{i0} \mid \alpha_i \sim N\left(\frac{\alpha_i}{1-\rho}, \frac{1}{\sqrt{1-\rho^2}}\right). \tag{4}$$

We consider $n \in \{100, 200\}$, and $T \in \{5, 10\}$. For each combination of n and T, we show results for five different estimators of ρ, two of which are adversarial. Importantly, none of the estimators require specifying the distribution of α_i, as they only depend on within-group variation.

We denote the observed data as $\mathcal{Y}_n = (y_1', \ldots, y_n')'$, where $y_i = (y_{i1}, \ldots, y_{iT})'$, and the simulated data as $\tilde{y}^{\ell}(\rho, \alpha) = (\tilde{y}_1^{\ell}(\rho, \alpha_1)', \ldots, y_n^{\ell}(\rho, \alpha_n)')'$ for path $\ell = 1, \ldots, H$, where $\tilde{y}_i^{\ell}(\rho, \alpha_i) = (\tilde{y}_{i1}^{\ell}(\rho, \alpha_i), \ldots, \tilde{y}_{iT}^{\ell}(\rho, \alpha_i))'$, for given choices of $\alpha = (\alpha_1, \ldots, \alpha_n)$ and ρ. In each replication, we draw $H = 20$ paths. Hence, $m = H \cdot n$. In addition, we define the lagged outcome $y^- = (y_1^{-\prime}, \ldots, y_n^{-\prime})'$, where $y_i^- = (y_{i0}, \ldots, y_{iT-1})'$. Further denote $A = I_n \otimes A_T$, where $A_T = I_T - \iota'\iota$, with $\iota = (1, \ldots, 1)'$ of dimension T, the deviation from the mean operator. To simulate data, we use $\rho = 0.8$ and $\alpha_i = 1$ for all i.

We consider the following estimators. *Maximum likelihood* (ML) coincides with the within-group estimator, $\hat{\rho}_{ML} = (y^{-\prime} A y^-)^{-1} y^{-\prime} A y$. As is well known, the ML estimator of (3) is biased for fixed T, even as N tends to infinity (e.g., Nickell 1981). We consider the *indirect inference* (II) method proposed by Gouriéroux, Phillips, and Yu (2010), which computes the ML estimator over each simulated path and is defined as $\hat{\rho}_{II} = \arg\min_{\rho} \| \hat{\rho}_{ML} - (1/H) \sum_l^H \tilde{\rho}_{ML}^l(\rho) \|^2$, where $\tilde{\rho}_{ML}^l$ is the ML estimator using the ℓ simulated path $\tilde{y}^{\ell}(\rho, \alpha)$. We consider the *adversarial estimator with logistic discriminator* (A-logit). We take $\mathcal{D}$ to be the family of logistic regression discriminators with $X_i = (T^{-1} y_i' A_T y_i^-, T^{-1} y_i' A y_i)'$, that is, the first-order autocovariance in deviations from the mean and the individual variance in deviations from the mean. For comparison, we also compute the *optimally weighted simulated method of moments* (SMM) *estimator* with the same individual moments. That is, $\hat{\rho}_{SMM} = \arg\min_{\rho} \big((1/n)\sum_{i=1}^n X_i - (1/m)\sum_{i=1}^m X_i^{\rho}\big)' \Omega \big((1/n)\sum_{i=1}^n X_i - (1/m)\sum_{i=1}^m X_i^{\rho}\big)$, with Ω the optimal weighting matrix. Finally, we consider the *adversarial estimator with neural network discriminator* (A-NN). We take $\mathcal{D}$ to be the family of one hidden layer neural network with five nodes, and $X_i = (y_{i1}, \ldots, y_{iT})$.

TABLE 1—RESULTS FOR $n = 100$

	$T = 5$		$T = 10$	
	Bias	RMSE	Bias	RMSE
ML	−0.443	0.445	−0.160	0.162
II	−0.040	0.116	0.001	0.036
SMM	−0.202	0.286	−0.030	0.104
A-logit	−0.020	0.075	0.005	0.038
A-NN	−0.008	0.031	−0.009	0.030

TABLE 2—RESULTS FOR $n = 200$

	$T = 5$		$T = 10$	
	Bias	RMSE	Bias	RMSE
ML	−0.442	0.443	−0.159	0.160
II	−0.033	0.104	0.002	0.027
SMM	−0.191	0.278	−0.019	0.086
A-logit	0.004	0.037	0.004	0.029
A-NN	−0.006	0.023	−0.009	0.023

Notes: Bias and RMSE for 500 replications. **ML** indicates maximum likelihood estimator, **II** indicates indirect inference estimator, **SMM** stands for simulated method of moments, **A-logit** indicates adversarial estimator with logistic discriminator, and **A-NN** indicates adversarial estimator with neural network discriminator. The first two columns correspond to $T = 5$, while the last two columns correspond to $T = 10$.

Tables 1 and 2 report bias and RMSE of all 5 estimators from 500 Monte Carlo replications for $n = 100$ and $n = 200$, respectively.

When $n = 100$, ML displays large bias for $T = 5$, which shrinks significantly when $T = 10$. This is in line with the asymptotic theory, as $\rho = 0.8$ is close to unity. The II estimator dramatically improves upon ML in terms of both bias and RMSE but is still moderately biased for small T. Note that the optimally weighted SMM estimator suffers from significant finite sample bias for $T = 5$. Finite sample bias in SMM is well known and documented in Altonji and Segal (1996). Instead, when $T = 5$, the A-logit estimator, while asymptotically equivalent to SMM, displays bias 10 times smaller than that of SMM and half than that of II. When $T = 10$, the performances of II and A-logit are comparable in terms of both bias and RMSE and are superior to that of SMM. Finally, the A-NN shows the

smallest bias and the smallest RMSE across all estimators for both $T = 5$ and $T = 10$. We interpret this result as evidence that the neural network provides the closest approximation to the oracle discriminator.

Table 2 provides simulation results when $n = 200$. ML behaves similarly to when $n = 100$ for both $T = 5$ and 10. Likewise, for $T = 5$, II and SMM do not improve or deteriorate. However, when $T = 10$, both II and SMM show smaller RMSE relative to $n = 100$. Finally, both adversarial estimators improve their performance when n increases. Our interpretation of this fact is that when n increases, the discriminator provides more accurate predictions, which in turn translates into smaller RMSE on $\hat{\theta}$. The A-NN again displays the smallest RMSE across the table.

IV. Conclusion

We show that the adversarial method has interesting properties, both from the theoretical point of view as well as in finite samples. In addition, we showcase implementation and performance in a limited MC experiment in a dynamic linear fixed effects panel data model. Notably, compared to other proposals in the literature, the adversarial estimator with a flexible class of discriminators yields unbiased estimates with the smallest RMSE.

REFERENCES

Altonji, Joseph G., and Lewis M. Segal. 1996. "Small-Sample Bias in GMM Estimation of Covariance Structures." *Journal of Business & Economic Statistics* 14 (3): 353–66.

Casella, George, and Roger L. Berger. 2002. *Statistical Inference*. 2nd ed. Duxbury Pacific Grove, CA: Duxbury.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville et al. 2014. "Generative Adversarial Nets." In *NIPS '14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, Vol. 2, edited by Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberge, 2672–80. Cambridge, MA: MIT Press.

Gourieroux, Christian, Alain Monfort, and Eric Renault. 1993. "Indirect Inference." *Journal of Applied Econometrics* 8 (S1): S85–118.

Gourieroux, Christian, Peter C.B. Phillips, and Jun Yu. 2010. "Indirect Inference for Dynamic Panel Models." *Journal of Econometrics* 157 (1): 68–77.

Kaji, Tetsuya, Elena Manresa, and Guillaume Pouliot. 2020. "An Adversarial Approach to Structural Estimation." https://arxiv.org/abs/2007.06169.

Nickell, Stephen. 1981. "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49 (6): 1417–26.