The University of Chicago

Differential Privacy and Econometric Methods

By

Caleb Rollins

August 2022

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts Program in the Social Sciences

Preceptor: Christopher Dobronyi
Faculty Advisor: Christopher Dobronyi

Differential Privacy and Econometric Methods

Caleb Rollins

June 2022

Introduction

Differential privacy is an emerging topic in Economics that has just recently started to gain more attention. In my thesis, I am trying to answer can econometric estimates be preserved when estimation is performed on differentially private synthetic data? While most current approaches to differential privacy in causal inference focus on making specific estimators differentially private, my approach will make synthetic data that is differentially private, so any estimator can be applied to this synthetic data. I will do this using generative adversarial networks (GANs). In doing so, I will also explore other differentially private methods for econometric estimators, as well as explore ways to achieve consistent standard errors when reporting these estimators.

We will begin with a review of the literature related to differential privacy, generative adversarial networks, and how these topics relate to economics. I will follow this up with a section on calculating standard errors under differential privacy. I will give a small example to demonstrate how to adjust standard errors, along with some proofs. I will then finish with experiments and discussion of drawbacks and future directions.

Literature Review

The literature on differential privacy in economics is relatively new. With only a few applications in economics, my work should add a considerable amount to this literature. In this section I will explain what differential privacy is. I will then explain what GANs are and how both of these topics relate to economics.

Differential Privacy

As more data is collected, the desire to maintain privacy has become increasingly important. However, the increase in data has also made this goal more difficult to achieve. There have been high profile examples where privacy of individuals has been compromised because of auxilary data that has been leveraged to reveal the underlying data using only anonymity data or in some cases released statistics [1]. To solve this problem, differential privacy was invented.

Differential privacy is a method to release the output of an algorithm in a way that will protect the data that this algorithm was run on [2]. Formally, an algorithm A that takes data as input is differentially private if for two data sets D_1 and D_2 and any set of possible outcomes from K, then

$$Pr(A(D_1) \in K) \leq e^{\epsilon} Pr(A(D_2) \in K)$$

The intuition for why this a reasonable notion of privacy is that it preserves privacy even for two data sets that only differ by one entry. As an example, lets say I am in some data set that holds sensitive data–such as medical data–and I am worried about the output of A revealing whether or not I have some condition. If we consider D_1 as the data set that claims I have the condition, while D_2 claims I don't, then I can guarantee that regardless of what output K is observed from A, that

$$\frac{Pr(A(D_1) \in K)}{Pr(A(D_2) \in K)} \leq e^{\epsilon}$$

This means that there is a limit to how certain an observer can be about whether D_1 or D_2 was used, so there is always some level of plausible deniability as to whether or not I have the condition in question.

In this equation, ϵ is the privacy parameter and an algorithm for which this condition holds is called ϵ-differntially private. The larger ϵ is, the less privacy is preserved with $\epsilon = \infty$ providing no privacy and $\epsilon = 0$ providing perfect privacy but the noise necessary to provide this privacy drowning out any signal.

There is no procedure for the selection of ϵ. This is usually chosen ad hoc, but some work has looked at the tradeoff of signal to privacy when making this decision [3]. This work uses common ideas in economics for its analysis. This is one of the earliest applications of economics to the differential privacy literature, but will not be the focus of this paper. Instead, I'll focus on differential privacy applied to econometric estimators.

Several papers have looked at differential privacy's application to causal inference. Work by Chetty and Friedman [4] has already been devoted to differential privacy for OLS regression. This paper aims to obtain differentially private estimates of OLS regressors for small sample sizes of U.S. census data. However, they are unable to fully achieve this goal. In their paper, they take the approach of adding noise to their estimate to ensure privacy, which is a common approach to achieving differential privacy. However, to get the correct magnitude of noise, you must know the maximum amount you estimator can change from removing one data point from your data. This change is theoretically unbounded for OLS, so Chetty and Friedman instead use the maximum change possible over the whole data set and applied this for each sample. While this isn't formally differentially private, it does provide significant privacy. My approach would extend this work by allowing for OLS estimation that is actually differentially private. Because my approach would generate data that is itself differentially private, any functions of that data would also be differentially private, which includes OLS estimates.

Some work has already achieved differential privacy for causal inference[5]. This paper gives a differentially private approach to the additive noise model framework. In their paper, the authors show that this model successfully preserves causal estimates even under privacy restrictions. This is a good sign for my paper, as it shows the possibility of successful causal inference under the requirements of differential privacy. While this paper does give an approach to differentially private causal inference, the additive noise model is not commonly used in econometrics. My paper will extend this work by looking specifically at the most common econometric methods of causal inference (IV, RDD, etc.). My approach would also be more general, as their work looks at a specific framework of causal inference, while mine would allow for any framework to be applied.

The previous paper talks about differential privacy for causal inference, with application one being medical data. Medical data will be a large focus of my paper as well. Medical data is one of the most important use cases of differential privacy. Privacy concerns are a big hurdle to even accessing medical data, so having synthetic data for preliminary analysis could help researchers determine fruitful paths to take before the strenuous process of actually getting access to this data. However, in addition to simpler methods discussed earlier in this review, I would also like for more complex methods (such as debiased machine learning) to be able to be applied to the synthetic data I create. Since these approaches rely on machine learning, it is important that synthetic data can be produced that behaves similarly to the underlying data on machine learning tasks. Recent work [6] suggests that this is possible with medical data. In this paper, the authors generate synthetic medical data. They then attempt several classification tasks on this data, training their models on the synthetic data and testing on the real data. They then compared this to the performance of classifiers both trained and tested on the real data. The results for each training method were very close, showing that the necessary information for machine learning was preserved. This is promising for my work, as my work also depends on complex relationships in the data being preserved. If my approach proves successful then this will be further evidence that differentially private synthetic data can preserve complex relationships.

In addition to the body of work promoting methods of causal inference that respect differential privacy, there are also papers looking at the limits of differential privacy in causal frameworks. A paper by Komarova and Nekipelov [7] looks at the theoretical limitations of differentially private regression discontinuity estimates. However, this work is largely theoretical. My approach will be able to test these theoretical limits in a practical application. As the authors of this paper acknowledge "While differential privacy does provide formal non-disclosure guarantees, its impact on the identification of empirical economic models as well as its impact on the performance of estimators in nonlinear empirical Econometric models has not been sufficiently studied." Through my research, I will be able to begin to fill in this gap in knowledge.

Generative Adversarial Networks

One of the most significant advances in deep learning in the past decade was the advent of Generative Adversarial Networks (GANs) [8]. The goal of a GAN is to generate samples that resemble some underlying dataset. These were first used to imitate images, but can be used to imitate any arbitrary type of data.

GANs are made up of two models: A generator and a discriminator. The generator takes in random noise as an input, and outputs a sample that is meant to resemble the underlying data. The discriminator in turn takes as input samples from out data and outputs from the generator and is trained to classify the inputs as real or fake. Thus, as the generator gets better (its outputs look more like the underlying data) then the discriminator must also improve to discriminate between the increasingly similar samples. As the discriminator improves, the generator must also improve. This results in generators that can produce samples that look very similar to the underlying data.

One interesting application to economics is using GANs to train structural models [9]. By using a structural model as the generator, it is possible to tune the parameters of the model until a discriminator can't tell the difference between data generated by the structural model and the real data used to train the model.

GANs are especially useful when generating high dimensional data, which is the focus of this paper. There are certain statistical guarantees that are given by GANs [10]. In tests, we can see that GANs perform very well [11]. Both of these papers show that the distribution learned by a GAN will converge to the actual distribution under a variety of distance metrics.

Confidence Intervals

One of the most important aspects of an econometric estimator is consistent confidence intervals. Such confidence intervals allow the researcher to assess the plausibility that the actual value underlying an estimator is in a certain range. Thus, the researcher can determine whether the estimator is "significant", meaning the estimator is likely not zero and actually affects the outcome in question.

Because of the importance of confidence intervals, it would be very useful if we could find consistent confidence intervals on our synthetic data.

For a differentially private estimator to have consistent confidence intervals, these intervals need to be increased to account for the increased uncertainty added by the differential privacy. The propery we want to hold is that

$$Pr(\theta \in CI) = 1 - \alpha$$

for an α confidence interval. This is easy to achieve when our estimator is asymptotically normal and we use additive noise. In this case, we know that

$$\sqrt{n}(\hat{\theta} - \theta) = N(0, \Sigma)$$

. It then follows that if we add noise L to $th\hat{e}ta$ with mean 0, then we tend towards a distribution with a mean of zero and variance of

$$\sigma^2 = 1/n * \sigma + var(L)$$

.

Below I will explain 3 ways to achieve consistence confidence intervals. I will then run some simple experiments to demonstrate the effectiveness of these methods. The experiments will be as follows:

- Draw n samples from a distribution with support on [0,1] and mean 0.5
- Calculate the mean and standard error (using a t statistic)
- Add differentially private noise to the mean
- Repeat 100 times and report how many time 0.5 lands in the adjusted 95 percent confidence interval

By limiting the distribution to a support of [0,1], we can guarantee that the mean doesn't change by more than 1/1000 when changing one data point, which will allow us to apply differential privacy. Since the mean is asymptotically normally distributed, we would expect that, with no noise, that 0.5 is in the reported confidence interval 95 percent of the time. In each table, we will report both with and without adjustment for a range of ϵ values and for both a uniform and beta distribution as the distribution we're drawing from.

Gaussian Noise

As will be seen throughout this paper, there are certain ways to relax differential privacy in a way that doesn't provide the same privacy guarantees, but can still be useful in some instances. One such relaxation is (δ, ϵ) differential privacy [2], which defines a mechanism as (ϵ, δ) differentially private if it meets the condition

$$Pr(A(D_1) \in K) \leq e^{\epsilon} Pr(A(D_2) \in K)$$

By relaxing the definition in this way, it is possible to achieve differential privacy by adding gaussian noise with standard deviation $\sigma = 2ln(1.25/\delta)/\epsilon$ [12]. This only holds for $\epsilon \in (0, 1)$. In the literuature it is common to set δ to a small value, so for our analysis I'll use $\delta = 0.01$ anytime Gaussian noise is used.

Gaussian noise is particularly useful because many econometric estimators are asymptotically normal. This means that as n increases,

$$\sqrt{n}(\hat{\theta} - \theta) = N(0, \Sigma)$$

For an estimator $\hat{\theta}$ of θ. If L is normally distributed with privacy (δ, ϵ), then the resulting distribution of $\hat{\theta}$ approaches

$$N(\theta, \frac{1}{n}\Sigma + (2ln(1.25/\delta)/\epsilon)^2 I)$$

because the sum of two normally distributed variables is also normally distributed. Below are experiments using such a normal distribution for gaussian noise.

ϵ	n	Distribution	Coverage (no adjustment)	Coverage (Adjusted)
1	100	uniform	0.73	0.91
		beta	0.76	0.94
	1000	uniform	0.79	0.95
		beta	0.8	0.96
	10,000	uniform	0.8	0.94
		beta	0.8	0.95
8	100	uniform	0.82	0.92
		beta	0.85	0.92
	1000	uniform	0.87	0.95
		beta	0.87	0.95
	10,000	uniform	0.86	0.97
		beta	0.87	0.95

Laplace Noise

Even though we know the variance of our resulting asymptotic distribution, without knowing the actual functional form of this distribution we can't determine the necessary confidence intervals. For instance, using confidence intervals produced by gaussian noise when laplace noise is added gives the following results.

As can be seen, this doesn't give the correct coverage because the sum of a gaussian and laplace distribution is not normally distributed. Thus, to get the actual desired confidence interval, we need to do a convolution of these two distributions. Let $X = N(\mu, \sigma)$ and $Y = Laplace(0, b)$. We then need to solve for $Pr(X+Y \leq k)$ by integrating over the space where $X+Y \leq k$. Since X and Y are independent, there joint distribution is just the product of each marginal distribution, so the integral we need to solve is

$$\int_{x=-\infty}^{x=\infty}\int_{y=-\infty}^{y=k-x}\left(\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{\sigma^2}}\right)\left(\frac{1}{2}e^{-|\frac{y}{b}|}\right)dxdy$$

Since X and Y are independent, we can switch the order of integration pull the density of X out of the first integral to get

$$\int_{x=-\infty}^{x=\infty}\left(\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{\sigma^2}}\right)\int_{y=-\infty}^{y=k-x}\left(\frac{1}{2}e^{-|\frac{y}{b}|}\right)dydx$$

$$\int_{x=-\infty}^{x=\infty}\left(\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{\sigma^2}}\right)F_y(k-x)dx$$

where $F_y(y)$ is the CDF of Y. Since Y is a laplace distribution, the CDF is a piecewise function with $F_y(y) = (1/2)e^{y/b}$ for $y \leq 0$ and $F_y(y) = 1-(1/2)e^{-y/b}$

for $y \geq 0$. Thus, to solve our integral we can split our integral into a sum over the different pieces of the function. Using this, we get the following equation for our integral.

$$\int_{x=-\infty}^{x=\infty} (\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}}) F_y(k-x)dx + \int_{x=-\infty}^{x=\infty} (\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}}) F_y(k-x)dx$$

$$= \int_{x=-\infty}^{x=\infty} (\frac{1}{2 * \sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2} + \frac{k-x}{b}})dx + \int_{x=-\infty}^{x=\infty} (\frac{1}{2 * \sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}})dx$$

$$- \int_{x=-\infty}^{x=\infty} (\frac{1}{2 * \sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2} - \frac{k-x}{b}})dx$$

We can shift our focus to the integral at the beginning of the equation. To solve this we'll convert it back into the form of a gaussian variable with mean μ_1 and standard deviation σ_1 so we can write it as a function of an error function. Doing this will leave some leftover term which we'll call c_1.

$$\int_{x=-\infty}^{x=\infty} (\frac{1}{2 * \sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2} + \frac{k-x}{b}})dx$$

$$= \int_{x=-\infty}^{x=\infty} (\frac{1}{2 * \sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{\sigma_1^2} + c_1})dx$$

$$= \frac{\sigma_1 e^{c_1}}{2\sigma} \int_{x=-\infty}^{x=\infty} (\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{\sigma_1^2}})dx$$

$$= \frac{\sigma_1 e^{c_1}}{2\sigma} \Phi(\frac{k-\mu_1}{\sigma_1})$$

Summing the three addends of this equation together we get the following solution

$$\frac{\sigma_1 e^{c_1}}{2\sigma} \Phi(\frac{k-\mu_1}{\sigma_1}) - \frac{\sigma_2 e^{c_2}}{2\sigma} \Phi(\frac{k-\mu_2}{\sigma_2}) + \Phi(\frac{k-\mu}{\sigma})$$

Once we have a form for the the CDF of the resulting distribution, all we have left to do is solve for

$$Pr(x < k) = 1 - \alpha$$

. While this has no anlytical solution, we can quickly solve this using newton raphson. In practice, I solved this using an automatic solver from the scipy package.

ϵ	n	Distribution	Coverage (no adjustment)	Coverage (Adjusted)
1	100	uniform	0.65	0.89
		beta	0.69	0.93
	1000	uniform	0.71	0.95
		beta	0.69	0.95
	10,000	uniform	0.70	0.96
		beta	0.72	0.95
8	100	uniform	0.75	0.93
		beta	0.77	0.93
	1000	uniform	0.80	0.94
		beta	0.76	0.95
	10,000	uniform	0.79	0.96
		beta	0.75	0.95

Bootstrap

One final way to obtain confidence intervals is using the bootstrap. This is a popular statistical approach that has useful properties when looking for properties of a distribution underlying a sample. This method is especially useful when dealing with difficult to calculate standard errors, which is the case here with the added noise from our differentially private estimators.

To calculate these standard errors, we'll use the following algorithm. Assuming that there is some function $D(\hat{\theta}, \epsilon)$ that returns a differentially private estimate of $\hat{\theta}$, we do the following, which is outlined in [13]. This method involves drawing bootstrapped samples from the data, calculating the desired estimator, and then adding differentially private noise to the estimator. We can then get a quantile of the estimator by finding the quantile of our bootstrapped estimators.

In order for this process to be useful, it needs to returns consistent confidence interval. That means that, at least asymptotically, the probability that the true θ is in a $1 - \alpha$ confidence interval is α. This is shown to be true in [13]. For add noise, the simulation below use Laplacian noise.

Synthetic Data Experiment

We already how standard errors can be adjusted when differential privacy is achieved via additive noise. We looked at two different forms of additive noise: Gaussian and Laplace. We also showed how bootstrapping can be used to get standard errors.

One popular way to generate differentially private synthetic data is histogram based methods. If we take a histogram of our data, we can add differentially private noise to each bin because a single data point change can only change the bin by at most 1. We can then use these bins with added noise to draw from and get new data that is differentially private. We will do this using the method outlined in [14].

ϵ	n	Distribution	Coverage (no adjustment)	Coverage (Adjusted)
1	100	uniform	0.66	0.92
		beta	0.68	0.95
	1000	uniform	0.74	0.95
		beta	0.72	0.96
	10,000	uniform	0.71	0.95
		beta	0.72	0.94
8	100	uniform	0.74	0.95
		beta	0.76	0.94
	1000	uniform	0.81	0.95
		beta	0.78	0.95
	10,000	uniform	0.76	0.97
		beta	0.73	0.95

ϵ	n	Distribution	Gaussian Adjustment	Laplacian Adjustment	Bootstrap Adjustment
1	100	uniform	0.82	0.84	0.86
		beta	0.84	0.84	0.85
	1000	uniform	0.85	0.84	0.85
		beta	0.86	0.86	0.86
	10,000	uniform	0.85	0.84	0.87
		beta	0.84	0.83	0.86
8	100	uniform	0.87	0.88	0.90
		beta	0.86	0.86	0.87
	1000	uniform	0.87	0.9	0.90
		beta	0.85	0.88	0.89
	10,000	uniform	0.85	0.87	0.90
		beta	0.86	0.90	0.91

We will now use a very simple histogram based method to generate synthetic data. From here, we can use the three methods proposed previously and see how it improves coverage of confidence intervals.

In the remainder of this paper, I'll use the three methods described previously to adjust standard errors. Because of the black box nature of GANs, it is hard to say how each will work, so I'll use experimental results to judge each method.

Linear Models

One problem that frequently crops up in the differential privacy literature is the difficulty of reporting differentially private estimators for a range of estimators. This was acknowledged by Chetty in [4] when looking at linear regression. This can be easily illustrated using a simple example. Consider a regression of one variable with no intercept. Thus, we are fitting

$$Y_i = X_i\beta$$

, which, if a squared error is used, gives a solution of

$$\beta = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

If we let $Y_1 = X_1^2$ and we let $N = \sum_{i=2}^{n} X_i Y_i$ and $D = \sum_{i=2}^{n} X_i^2$, then we can rewrite β as

$$\beta = \frac{X_1^3 + N}{X_1^2 + D}$$

Thus, choosing large enough or small enough X_i and Y_i, we can make arbitrarily large changes to β, so no amount of noise can distinguish the β when X_1 is included and β when X_1 is not, so we can't guarantee differential privacy because we would always know if X_1 was present.

Despite this limitation, we can relax differential privacy slightly to get estimators that preserve privacy. In Chetty's paper, this limitation is overcome by putting bounds on X_i and Y_i. Building out these models will be useful for two reasons. First, as will be demonstrated, my model of inference will give better results in simulation, so it is useful in its own right. Secondly, it will give us another econometric estimator to test on our synthetic data. This is important because sampling a wide array of econometric estimators will help to see how useful the synthetic data produced actually is.

While this can be useful in situations where many regressions are being done, it might not be appropriate for situations in which only a few regressions are being done. I'll relax differential privacy in a different way for these instances. In many cases, the range a variable is able to take is naturally bounded. For instance, something like a test score might be bounded between 0 and 100. Even if there isn't an explicit range, we can usually put reasonable ranges on variables. For instance, if we're looking at heart rate, which is technically unbounded, we can put a large upper limit like 1000 that couldn't be reasonably achieved. By doing this, we can now define the maximum difference between two data sets that differ by one input, so we can use additive noise on our linear regression.

In order to do this, we will assume that the maximum change (absolute value of the difference between the smallest and largest possible value) across any element of X or Y is f. We could take maximum differences element by element to get smaller privacy bounds, but for this paper we'll just bound all variables between 0 and 1 in simulations for simplicity of calculations.

We'll use the classical setup for linear regression where $Y = \beta X$. Since $\hat{\beta} = (X'X)^{-1}X'Y$. By the rules of composition, if we use two $\epsilon/2$ differentially private mechanisms and combine them using some mathematical operation, we get an ϵ differentially private mechanism. We'll use this to get $\epsilon/2$ private estimates of $X'X$ and $X'Y$ by adding laplace noise L, and then matrix multiply $(X'X + L)^{-1}(X'Y + L)$ to get an ϵ differentially private estimate of β

To determine the appropriate noise, we need to determine what the maximum sensitivity is for $X'X$ and for $X'Y$.

Since linear regression is asymptotically normal, we can use the method above to solve for confidence intervals when Laplace noise is added. However, there is the issue of bias that could effect the validity of our confidence intervals. We can also use the bootstrap method. A similar method for linear regression is proposed in [13]. Both of these will be presented in the simulations below.

To generate our data for the simulations, we'll use $Y = \beta' X + u$ where u is standard normal (when using chetty's approximate method) and beta(0.5,0.5) when using the bounding method from above. Also, X will be standard normal or beta(0.5,0.5) for those respective methods. I will use β of all 1s and calculate coverage experiments and MSE using the first element of β

Debiased Machine Learning

An emerging method that has gained popularity in econometrics is debiased machine learning [15]. This is a semiparametric method used to estimate treatment effects. I will use the auto-DML implementation used in [15]. This method involves using some machine learning model γ to estimate outcome Y given treatment D and covariates X. We can then create a biased estimate of the ATE as the average of $\gamma(1, X_i) - \gamma(0, X_i)$ over all X_i in the test set. We can then use the method in [15] to remove the bias from this estimate.

One advantage of this method is its flexibility. Any estimator can be used for γ and we can still get unbiased estimates. This is useful because it allows us to use an estimator trained with differential privacy and still get unbiased estimates. For this, we will use differentially private neural nets as presented in (source) and differentially private random forests as presented in [16].

To generate data for my experiments, I will use the following data generating process

$$Y_i = D_i + D_i * X_i + X_i - X_i^3 + u_i$$

where u_i is standard normal noise, D_i is treatment assigned with selection bias, such that $D_i = 1$ if $X_i + v_i > 0$ where v_i is standard normal, and x_i is drawn from a standard normal. This process has an ATE of 1.

Binary Choice Logit Model

One of the most important methods in econometrics is the binary choice logit model. This model is particularly important in industrial organization. This model relates a binary variable Y_i to some set of regressors X_i with the following conditional probability.

$$Pr(Y_i = 1 | X_i, \beta) = \frac{e^{X_i'\beta}}{1 + e^{X_i'\beta}}$$

When we assume that β is constant, we can estimate this model using standard maximum likelihood estimation. Let $\phi(x) = \frac{e^x}{1+e^x}$ For observations $(Y_1, X_1), ...(Y_n, X_n)$, we can write the log likelihood as

$$L = Y_i ln(\phi(X_i'\beta)) + (1 - Y_i) ln(1 - \phi(X_i'\beta))$$

. While this cannot be solved analytically, as the equation setting the derivative to zero isn't analytically tractable, we can still use gradient based methods to solve this. While most approaches to logit model estimation use newton ralphson, we use gradient descent. For my experiments, I will use a β of all 1s with X drawn from a standard normal.

To give a baseline, I will build a differentially private estimator of the binary choice logit model. I will start by constructing an estimator that is able to preserve privacy without any relaxation of the definition of differential privacy. I will then present a second method that will slightly relax the definition of differential privacy. I will argue why this relaxation may be useful in some cases. I will finally construct methods to get standard errors, and then run some experiments confirming the effectiveness of my approach. To begin, I will use a method similar to the one used to train differentially private neural nets [17].

Up until this point, all the methods we have seen have added noise at the end to guarantee privacy. However, in training a differentially private neural net, the noise is added during the training process [18]. Specifically, this noise is added to the gradients, which are clipped after the noise is added. By clipping the gradients (only allowing them in a certain range), it is possible to set a bound on the maximum amount that they can differ when a data point is added or removed from the data set. This is important because the noise that is added depends on the maximum amount that a single point can effect the output.

As mentioned earlier, binary logit models are usually solved using newton raphson. However, as shown in [18] this can sometimes be unstable because the hessian matrix might, under added noise, fail to be positive semidefinite. To overcome this problem in [18], the authors use a public and private dataset, pulling the hessian from the public data. While this is an interesting idea, we're interested in only having private data, so this method won't work for our purposes. Instead, I choose to use gradient descent, as the nueral net literature does, to ensure privacy. This gives the following algorithm.

- Calculate gradient
- Add Laplacian noise to the gradient
- Clip gradient between -f and f
- Repeat until privacy budget has been expended

Methodology

In the previous sections, we developed differentially private econometric estimators for a number of common econometrics problems. We also showed how histrogram based methods for generating differentially private synthetic data can be used effectively for small data sizes. We will now expand to a higher dimensional setting. In this setting, we will generate the differentially private data using GANs (as implemented in) and a histogram based method outlined in (source).

Using debiased machine learning and binary choice logit have significant advantages for testing our synthetic data. First, each method is nonlinear, so we can see if the complex patterns we want to see preserved by the GANs are actually preserved. Secondly, both can have differential privacy applied directly to the method itself (as outlined in each methods respective section) so we can treat this as a ”best case” or using differential privacy. Since the differential privacy is applied only to the method itself and not the entire data set, it is likely impossible to better performance. For the synthetic data, privacy is preserved for every type of analysis, not just one specific analysis.

For our simulations, we will use the data generating processes outlined in the linear models, debiased machine learning, and binary choice logit model sections above. We will then compare the mean squared errors and coverage rates of the respective econometric estimators using the differentially private estimator on the underlying data and using the non differentially private estimator on the synthetic data. We will do this over a number of different values of ϵ and dimensions. We will also choose one dimension and ϵ to test different methods of adjusting standard errors on the synthetic data. We will look at the resulting coverage

Results

Linear Model

ϵ	n	MSE	Synthetic Data Coverage	Adjusted Estimator Coverage
1	1000	0.835	0.85	0.94
	10,000	0.841	0.84	0.93
8	1000	0.506	0.88	0.92
	10,000	0.492	0.89	0.95
25	1000	0.252	0.88	0.95
	10,000	0.244	0.90	0.96

Debiased Machine Learning

ϵ	n	MSE	Synthetic Data Coverage	Adjusted Estimator Coverage
1	1000	0.932	0.82	0.93
	10,000	0.906	0.83	0.95
8	1000	0.677	0.85	0.96
	10,000	0.655	0.87	0.95
25	1000	0.310	0.90	0.96
	10,000	0.324	0.89	0.95

Binary Choice Logit Model

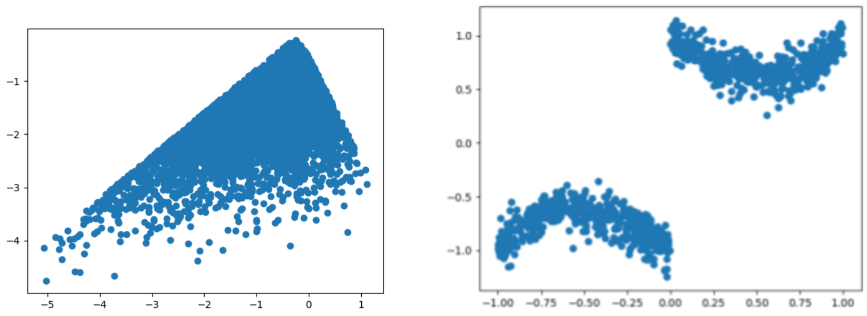
ϵ	n	MSE	Synthetic Data Coverage	Adjusted Estimator Coverage
1	1000	0.678	0.79	0.95
	10,000	0.699	0.81	0.96
8	1000	0.456	0.82	0.94
	10,000	0.407	0.85	0.96
25	1000	0.105	0.85	0.97
	10,000	0.098	0.86	0.95

Discussion

Drawbacks

As can be seen above, the results on the synthetic data are not as good as with the actual estimator. There is thus a tradeoff between the versatility of having differentially private synthetic data (which can work on many different estimators) and a tailored estimator that will have better performance and, crucially, consistent confidence intervals.

Another problem we ran into was that there are certain estimators that the synthetic data couldn't capture at all. This was the biggest problem with the regression discontinuity design. I used data generated from the following data generating process.



(a) Synthetic Data with $\epsilon = 1$

(b) Underlying Data

Figure 1: RDD Demonstration

Using the RDD, this should give a LATE at 0 of 1. This can be demonstrated in the graph below. This graph shows differentially private data generated on the right, generated from the underlying data on the left.

0.1 Graphic

However, as can be seen in one sample of generated data from a GAN, this is not the case. Across 1000 samples, there was no tendency towards a LATE of 1, with estimates varying wildly. This is consistent with the theory presented in [7].

Future Work

This work is a very preliminary look at a topic with much potential. I looked at only a handful of methods and only a couple of data generating methods. It would thus be a natural extension to look at more econometric methods and data generating methods. It could also be helpful to tailor data generating methods to specific econometric estimators. For instance, if we're working with consumer data, it would be helpful to have synthetic data that performs well on logit and probit models.

Conclusion

From the results, we can see that there is still work that needs to be done to improve the usefulness of synthetic data for econometric methods. However, there were promising results both in terms of coverage and mean squared error for the estimators studied. Also, we were able to get excellent performance from the differentially private estimators for all three methods studied. This suggests that, at least for the time being, the best approach to differential privacy in econometrics is using specifically designed differentially private estimators for each task, rather than using synthetic data.

References

[1] B. Lubarsky, "Re-identification of 'anonymized' data," *Georgetown Law Technology Review*, 2017.

[2] M. Aitsam, "Differential privacy made easy," 2022.

[3] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth, "Differential privacy: An economic method for choosing epsilon," *2014 IEEE Computer Security Foundations Symposium*, 2014.

[4] R. Chetty and J. Friedman, "A practical method to reduce privacy loss when disclosing statistics based on small samples," *National Bureau of Economic Research Working Paper Series*, 2019.

[5] M. Kusner, Y. Sun, K. Sridharan, and K. Weinberger, "Private causal inference," *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016.

[6] A. Torfi, E. Fox, and C. Reddy, "Differentially private synthetic medical data generation using convolutional gans," *Information Sciences*, 2021.

[7] T. Komarova and D. Nekipelov, "Identification and formal privacy guarantees," *https://arxiv.org/abs/2006.14732*, 2021.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Advances in Neural Information Processing Systems 27*, 2014.

[9] T. Kaji, E. Manresa, and G. Pouliot, "An adversarial approach to structural estimation," *BFI Working Paper Series*, 2020.

[10] M. Chen, W. Liao, H. Zha, and T. Zhao, "Statistical guarantees of generative adversarial networks for distribution estimation," *Proceedings of Machine Learning Research*, 2021.

[11] T. Liang, "On how well generative adversarial networks learn densities: Nonparametric and parametric results," 2018.

[12] F. Liu, "Generalized gaussian mechanism for differential privacy," *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[13] C. Ferrando, S. Wang, and D. Sheldon, "Parametric bootstrap for differentially private confidence intervals," *Proceedings of Machine Learning Research*, 2022.

[14] A. T. Suresh, "Differentially private anonymized histograms," *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[15] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and structural parameters," *The Econometric Journal*, 2018.

[16] S. Fletchera and M. Z. Islama

[17] M. Knolle, D. Usynin, A. Ziller, M. R. Makowski, D. Rueckert, and G. Kaissis, "Neuraldp differentially private neural networks by design," 2021.

[18] I. A. Adjei and R. Karim, "An application of bootstrapping in logistic regression model," *Open Access Library Journal*, 2016.