Statistical Methods for Understanding Social Issues

Casey Bradshaw

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2025

Abstract

Statistical Methods for Understanding Social Issues

Casey Bradshaw

The modern-day abundance of both data and computational resources has expanded the potential for statistical methods to improve our understanding of social phenomena. Such insights can also serve as a basis for policy decisions and resource allocation. This thesis explores pragmatic approaches to statistical inference in select applications related to social issues.

We begin by considering underreporting of sexual assault on college campuses. Because many instances of sexual assault are never reported to authorities, the number of reported assaults does not fully reflect the true total number of assaults that occurred, and could arise from many combinations of reporting rate and true incidence. We estimate these quantities via a hierarchical Bayesian model of the reported data, drawing prior information from national crime statistics to help distinguish between reporting rates and incidence.

Next, we consider the task of detecting election interference, focusing on Ghana's 2020 presidential election. Using locally-aggregated vote tabulation data, we construct a randomization test to screen for potential ballot box stuffing. After identifying regions with suspicious results, we estimate alternative vote totals under a counterfactual scenario with no election interference.

Next we turn our attention to synthetic control methods for causal inference. Using a motivating example of estimating the effects of exogenous economic shocks on support for political incumbents, we propose methods for leveraging shared information across multiple related studies.

Finally, we address the issue of data privacy. Motivated by the need to balance the utility of data

analysis with protection of individuals' privacy, we develop randomized optimization algorithms for differentially private statistical inference. These procedures allow the construction of private confidence regions, and employ a bias correction that improves empirical performance in small samples.

Table of Contents

Acknowledgments xi

Introduction 1

Chapter 1: A Bayesian Model of Underreporting for Sexual Assault on College Campuses . 4

1.1 Related Work 5

1.2 Dataset 7

1.3 Methodology 9

1.3.1 Modeling z 10

1.3.2 Modeling p 13

1.3.3 Inference 16

1.3.4 Model Assessment 18

1.3.5 Capabilities and Limitations 21

1.4 Results 23

1.4.1 Role of Covariates 23

1.4.2 Systemwide Results 25

1.4.3 School-Level Results 26

1.5 Discussion 29

Chapter 2: Detecting Voting Irregularities in Ghana's 2020 Presidential Election 31

2.1 Background . . . 31
2.2 Methodology . . . 34
2.3 Voting Patterns in the Ashanti Region . . . 37
2.3.1 Polling Place-Level Inferences . . . 38
2.3.2 Counterfactual Election Results . . . 42
2.4 Discussion . . . 48

Chapter 3: Leveraging Shared Information Across Multiple Synthetic Control Studies . . . 49
3.1 Related Work . . . 51
3.2 Methodology . . . 52
3.2.1 Conventional Synthetic Control . . . 53
3.2.2 Matrix Completion . . . 54
3.2.3 Hierarchical Bayesian Synthetic Control . . . 54
3.3 Simulation Study . . . 56
3.3.1 Unimportant Missing Units . . . 58
3.3.2 Important Missing Units . . . 59
3.4 Application: Spanish Christmas Lottery . . . 60
3.4.1 Barcelona . . . 61
3.4.2 Granada . . . 64
3.5 Discussion . . . 68

Chapter 4: Differentially Private Inference via Noisy Optimization . . . 70
4.1 Background . . . 71
4.1.1 Notation . . . 71

4.1.2 Gaussian Differential Privacy . . . 71
4.1.3 M-estimators for Parametric Models . . . 73
4.1.4 Related Work . . . 74
4.2 Randomized M-estimators via Noisy Gradient Descent . . . 77
4.2.1 Examples . . . 78
4.2.2 Convergence Analysis . . . 82
4.3 Randomized M-estimators via Noisy Newton's Method . . . 85
4.3.1 Matrix-Valued Noise for Private Hessians . . . 85
4.3.2 Differentially Private Newton's Method . . . 86
4.3.3 Initial Phase . . . 89
4.3.4 Convergence Analysis . . . 95
4.4 Inference via Noisy Variance Estimators . . . 96
4.4.1 Private Sandwich Formula . . . 96
4.4.2 Finite-Sample Correction . . . 98
4.5 Discussion . . . 100

References . . . 102

Appendix A: Supplementary Material for Chapter 1 . . . 112
A.1 Dataset Details . . . 112
A.2 Derivations . . . 114
A.3 Pooling in Hierarchical Models . . . 116
A.3.1 Complete Pooling . . . 117
A.3.2 No Pooling . . . 118

A.3.3 Comparison 119
A.4 Sensitivity to Correlated Reporting 120
A.5 Sensitivity to Prior Specification 121

Appendix B: Supplementary Material for Chapter 2 127

Appendix C: Supplementary Material for Chapter 3 129

Appendix D: Supplementary Material for Chapter 4 130
D.1 Concentration Inequalities 130
D.2 Convex Analysis 131
D.3 Proofs for Noisy Gradient Descent 132
D.3.1 Proof of Theorem 2 132
D.3.2 Proof of Proposition 1 142
D.4 Proofs for Noisy Newton's Method 146
D.5 Numerical Illustrations 153
D.5.1 Linear Regression 153
D.5.2 Logistic regression 155
D.6 Benchmark Experiments 157
D.6.1 Linear Regression 157
D.6.2 Logistic Regression 161

List of Figures

1.1 School-level reported sexual assault data is sparse, with 42% of records showing zero assaults, and a median of one assault. 8

1.2 The total number of reported assaults grew from 2014 to 2018. Total student population fluctuated relatively little, resulting in per capita reporting trends moving with the overall total number of reports. 9

1.3 Graphical model diagram 10

1.4 Prior distributions on β_0, ε, and η imply a distribution over incidence centered at 4.1 per 1000 persons, with a central 50% range of 2.3 to 7.4. 13

1.5 Junior colleges and institutions of religious instruction comprise a minority of the dataset 15

1.6 Prior distributions on α_0, γ, and δ imply a distribution over reporting probability centered at 22%, with a central 50% range of 10% to 43%. 16

1.7 Summary statistics of the held out data are plausible under the predictive distribution. Solid lines indicate the value calculated on the held out data, dashed lines indicate the 2.5th and 97.5th percentiles of the predictive distribution; (a) Proportion of reports indicating zero assaults (b) Proportion of reports indicating $\leq$ 1 assaults (c) Total number of reported assaults across all schools and all years (d) Within-school variability of reported numbers over time 20

1.8 The posterior distribution of $\lambda_0 p_0$ concentrates as the amount of observational data increases, but the posterior distribution of p_0 has some amount of uncertainty that cannot be eliminated by observational data; (a) Kernel density estimates of posterior samples of p_0, conditioned on observed datasets of size N, with a solid line representing $p(p_0|\lambda_0 p_0)$, (b) Kernel density estimates of posterior samples of $\lambda_0 p_0$, conditioned on observed datasets of size N 22

1.9 Additional observational data is more effective at reducing posterior uncertainty about the slope coefficient β than about the intercept ; (a) Kernel density estimates of posterior samples of β_0, conditioned on observed datasets of size N (b) Kernel density estimates of posterior samples of β_0, conditioned on observed datasets of size N, displayed on x-axis interval of equal width as left-hand plot 23

1.10 Posterior distribution of total incidence per 1000 persons in the college population does not exhibit a clear pattern of year-over-year change. 25

1.11 Posterior estimates of the overall reporting rate in the college population exhibit an increasing trend from 2014 to 2018. 26

1.12 Per-record posterior medians for incidence and reporting rate exhibit heterogeneity; (a) Posterior medians for incidence are centered slightly below the overall population value at 2.3 per 1000, but 25% of records have posterior median incidence greater than 3.4, and another 25% below 1.6. (b) Posterior medians for reporting probabilities are predominantly below 30%, but for some records are as high as 74%. 27

1.13 Variation in reporting probability alone is less likely to produce an increased number of reported assaults in subsequent year at (a) Randolph College than at (b) University of South Carolina- Columbia. Solid-colored bar corresponds to number of reported assaults in 2019. 28

1.14 Without a change in the true number of assaults, the number of reported assaults has a 44% chance of increasing in a subsequent year at (a) Ursinus College, and a 67% chance at (b) Temple University. 29

2.1 (a) Map of Ghana. Constituencies highlighted in pink had voting results flagged for significant positive correlation, (b) Example of a constituency whose results were flagged, (c) Example of a constituency that was not flagged 36

2.2 Constituencies in the Ashanti region. Constituencies flagged for significant positive correlation between turnout and incumbent share are highlighted in pink. . . . 37

2.3 Voter turnout and vote share for incumbent candidate per polling station are distributed differently in the flagged constituencies. 38

2.4 10%, 25%, 50%, 75%, and 95% quantiles of the estimated distribution for each cluster 40

2.5 Colors correspond to the estimated posterior probability that a polling place belongs to cluster 2 41

2.6 This mapping redistributes the data points from cluster 2 to match the approximate bivariate Gaussian distribution corresponding to cluster 1. 43

2.7 Example trajectories created by deleting votes for the incumbent at an individual polling place. As the number of deleted votes increases, the polling place's dot on the scatter plot moves from right to left along the orange curve. 45

2.8 Discarding approximately 34,000 votes for the incumbent could be sufficient to produce voting results consistent with expectations for a "fair" election. 47

3.1 In the envisioned data setting, each treatment period can be considered to demarcate a separate study. 50

3.2 Political support for the incumbent party in synthetic Barcelona closely approximates the observed data for Barcelona during most pre-treatment time periods. . . . 63

3.3 Estimated effect of lottery win on support for incumbent party in Barcelona 64

3.4 Estimates of political support for the incumbent party are less precise in Granada than in Barcelona, though most 90% posterior credible intervals for synthetic Granada cover the actual observed value during pre-treatment periods. 66

3.5 Estimated effect of lottery win on support for incumbent party in Granada 66

3.6 For both Barcelona and Granada, the estimated treatment effects are largest for the years in which the amount of money won was largest 67

4.1 Noisy gradient descent trajectories for linear regression. (a) Estimates of a single coordinate of the regression vector. (b) Gradient of the loss function evaluated at the current iterate, plotted on a log scale. 79

4.2 Gradient clipping and consistency. In the linear regression setting, both Mallows weights (plot (a)) and gradient clipping (plot (b)) approaches are consistent. 81

4.3 Gradient clipping and consistency. Clipping results in a positive bias for logistic regression (plot (b)), which does not arise using Mallows weights (plot (a)). 81

4.4 (a) Noisy Newton's method vs. gradient descent. The vertical axis records the norm of the gradient trajectory on a log scale. (b) Divergence of noisy damped Newton. When $\eta = 1$ (pure Newton), the algorithm may not converge if the initial point is far from the global optimum $\hat{\theta}$. Using smaller values of η remedies this problem. . 88

4.5 Trajectories of noisy gradient descent, pure Newton, and damped Newton algorithms. All optimizations were initialized at $\beta^{(0)} = 0$, with σ^2 assumed to be known. 90

4.6 Behavior of Newton's method with backtracking line search, compared to pure and damped Newton's method 93

4.7 Behavior of 400 repetitions of noisy Newton's method with backtracking line search, applied to data generated from a logistic regression model with true $\beta = (1, 1, 1, 1)^T$ and initialization at $\beta^{(0)} = (-4, 6, -5, 6)^T$ 94

4.8 (a) Empirical confidence interval coverage with noisy gradient descent estimators. The blue dots show the empirical coverage probabilities of the last iterate utilizing the noisy sandwich formula and the corrected formula. The gray points illustrate the coverage of the corrected formula computed using the first iterate whose gradient is smaller (in Euclidean norm) than the standard error of the noise added at each iteration. This gradient uses a non-private stopping rule. (b) 2D coverage region for a pair of parameters, showing the tighter confidence region resulting from our finite-sample correction to the sandwich formula. Nonprivate counterpart included on right. 99

4.9 Empirical confidence interval coverage for confidence intervals constructed using noisy Newton's method estimators. The blue dots show the empirical coverage probabilities of the last iterate utilizing the noisy sandwich formula and the corrected formula, for an algorithm that terminates before the budgeted number of iterations if the noisy gradient is sufficiently small in magnitude. The gray dots illustrate the coverage of the corrected formula with a stopping rule based on the true gradient. 100

A.1 Complete Pooling Model Diagram 117

A.2 No-Pooling Model Diagram 118

A.3 The partial pooling model assigns higher likelihood to held-out observational data; (a) Estimated predictive log likelihood of the held-out dataset under each model (b) Left-hand side plot rescaled to include log likelihood of the held-out dataset under the prior from the partial pooling model 119

A.4 When reporting decisions within each school are positively correlated, the model tends to overestimate the true underlying number of assaults. 121

A.5 Prior distributions on the reporting probability p under scenarios (a) - (e) (with model coefficients α_1, α_2, α_3, and α_4 set to zero). 122

A.6 Coefficients relating covariates to the total number of assaults do not display much practical difference under moderate changes to the prior distributions envisioned in scenarios a-e. Panel (i) depicts β_1, the coefficient corresponding to student population. Panel (ii) depicts β_2, the coefficient corresponding to gender imbalance of student body. 123

A.7 Estimates of coefficients relating covariates to the reporting probability remain similar under moderate changes to the prior distributions, and begin to shrink toward zero under more extreme changes to the prior distribution. Panel (i) depicts α_1, the coefficient corresponding to whether a school is a junior college. Panel (ii) depicts α_2, the coefficient corresponding to whether a school is substantially engaged in religious instruction. Panel (iii) depicts α_3, the coefficient corresponding to the proportion of a school's student body composed of women. Panel (iv) depicts α_4, the coefficient corresponding to the proportion of a school's student body receiving Pell grants. 124

A.8 Prior belief in a higher incidence of sexual assault produces higher posterior estimates of the total number of assaults occurring. 125

A.9 Prior belief in a lower probability of reporting sexual assault produces lower posterior estimates of the reporting rate. 125

A.10 Under scenario (e), posterior estimates of the incidence rate do not exhibit a clear increasing or decreasing trend, while posterior estimates of the reporting rate are more suggestive of an increase over 2014-2018. Panel (i) depicts the posterior distribution of the incidence rate; panel (ii) depicts the posterior distribution of the reporting rate. 126

D.1 Frequency with which the hypothesis β_j was rejected, for each of $j = 1, \ldots, 8$, under (a) differentially private gradient descent and inference, and (b) the corresponding non-private estimation and inference 154

D.2 Subsampled estimates of $\mathbb{E}(\|\hat{\beta} - \beta\|_2)$. 155

D.3 Linear regression with bounded covariates and responses 157

D.4 Linear regression with unbounded covariates and responses 160

D.5 (a) Scenario 1: Many data points near $\sup_i \|x_i\|_2$. (b) Scenario 2: Few data points near $\sup_i \|x_i\|_2$. 162

D.6 Logistic regression with unbounded covariates 164

List of Tables

1.1 Sexual Assault Incidence per 1000 persons in the US, by age group 13

1.2 Percentage of Sexual Assaults Reported to Police, by age group 16

1.3 Posterior means and quartiles of latent variables 24

2.1 Actual and hypothetical aggregated vote totals for flagged constituencies in the Ashanti region 43

3.1 When only unimportant units have missing data, SCM-complete assigns the highest predictive likelihood to the held-out data. 58

3.2 When important units have missing data, HB assigns the highest predictive likelihood to the held-out data. 59

3.3 Posterior means of synthetic control weights constituting synthetic Barcelona. Includes all provinces with a weight ≥ 0.02 in any of the six years. 62

3.4 Posterior means of synthetic control weights constituting synthetic Granada. Includes all provinces with a weight ≥ 0.001 in either year. 65

D.1 Coefficients estimated via noisy gradient descent ($\mu = 0.25$) 153

Acknowledgements

Many wonderful people have supported me through the course of this PhD, and I am immensely grateful to them for making this work possible.

To my advisor, David Blei- thank you for your invaluable support and patience. You have taught me so much, and your intellectual curiosity and enthusiasm are a constant source of inspiration. Every single conversation with you boosted my morale. I also thank the professors on my defense committee- Marco Avella Medina, Tian Zheng, Donald Green, and Mark Hansen- for many enlightening research discussions, and for offering guidance and connecting me with opportunities over the years. Thanks are also in order for Wesley Tansey at Memorial Sloan Kettering, who graciously hosted me in his lab and shared many useful lessons about scholarship and academia.

I am also grateful to my classmates in the statistics department for years of friendship and camaraderie. Likewise, thank you to the students and postdocs in Dave's lab, whose exceptional kindness and sense of humor made this journey much more enjoyable.

Thank you to my brother, parents, grandmother, and in-laws for always being so genuinely excited for me. I am profoundly fortunate to have you all in my corner. Finally, thank you to my husband Phil, who is my biggest cheerleader and makes me want to be a better version of myself.

Introduction

The modern-day abundance of both data and computational resources has expanded the potential for statistical methods to improve our understanding of social phenomena. Researchers are equipped to explore many detailed questions in this arena. For instance, what community characteristics are associated with steep increases in rates of homelessness?[1] How effective are power plant emission interventions at reducing the number of pediatric asthma emergencies?[2] What role does the US university system play in intergenerational socioeconomic mobility?[3] Chapters 1-3 of this thesis investigate pragmatic approaches to statistical inference in applications related to social issues, each with its own distinct challenges.

Chapter 1 considers the underreporting of sexual assault on college campuses. In an effort to quantify and combat sexual assault, US colleges and universities are required to disclose the number of reported sexual assaults on their campuses each year. However, many instances of sexual assault are never reported to authorities, and this underreporting clouds the interpretation of the reported data; the reported values could arise from many combinations of reporting rate and true incidence. National crime statistics, however, provide a source of information about overall reporting rates and incidence of assault. Using these national crime statistics as the basis for priors in a hierarchical Bayesian model of the number of reported assaults, we can estimate the underlying incidence and reporting rates at individual colleges. Results suggest an increasing trend in reporting rates for the overall college population over the 2014-2019 time frame. However,

[1] Glynn et al. (2021)
[2] Wikle and Zigler (2024)
[3] Chetty et al. (2020)

the extent of underreporting varies widely across schools. That variation has implications for how individual schools should interpret their reported crime statistics. This chapter is based on joint work with David Blei.

Chapter 2 addresses the task of detecting election interference from locally-aggregated per-party vote totals. Ghana's 2020 presidential election drew allegations of fraud, namely in the form of ballot box stuffing. One of the central challenges in quantitative approaches to detecting election fraud is the absence of labeled data, particularly for fraudulent cases. Consequently, the methods in this chapter rely on prevailing theories of how ballot box stuffing manifests in vote tabulation data. A randomization test based on those ideas identifies several regions whose voting patterns display hallmarks of ballot box stuffing. However, a closer examination of regions with suspicious voting patterns suggests that even if ballot box stuffing were solely responsible for those voting irregularities, it could potentially be explained by an amount of fraud far too small to have materially impacted the outcome of the election. This chapter is based on joint work with Sesinam Dagadu and Talay Cheema.

Chapter 3 focuses on synthetic control methods for causal inference. Pioneered by Abadie et al. (2010), synthetic control methods provide a strategy for estimating the causal effects of large-scale interventions, which may or may not be randomly assigned, from observational panel data. In this chapter, we propose a hierarchical Bayesian approach to sharing information across a collection of related synthetic control studies. The proposed method is designed to accommodate multiple studies for which the set of units that are *not* exposed to the intervention varies across studies. The motivating application concerns economic voting behavior: we aim to estimate the effects of exogenous economic shocks on support for incumbent politicians, using data from a large national lottery in Spain. This work is based on discussions with David Blei, Achille Nazaret, Andrew Jesson, and Donald Green.

In Chapter 4, we shift our focus inward and consider data analysis itself as a social issue. While the proliferation of publicly available data and statistical tools create many potential benefits, when data sets contain the records of individuals, researchers and data curators must also take

care to avoid compromising those individuals' privacy. To that end, differential privacy is a popular paradigm for defining and quantifying the privacy loss incurred in data analysis. Chapter 4 develops an optimization-based approach to computing differentially private versions of M-estimators from the robust statistics literature. The resulting private estimators converge with high probability to a small neighborhood around their non-private counterparts, and due to asymptotic normality, can also be extended to provide parametric confidence intervals. This chapter is based on joint work with Marco Avella Medina and Po-Ling Loh.

Chapter 1: A Bayesian Model of Underreporting for Sexual Assault on College Campuses

Sexual assault on college campuses is a pressing public health concern in the United States, prompting government action such as the 2014 formation of the White House Task Force to Protect Students from Sexual Assault (Obama, 2014) and the "It's on us" awareness campaign (Somanader, 2014). Per the Clery Act of 1990, US colleges and universities are required to report annual crime statistics for their campuses and adjacent areas. These crime statistics include the number of reported sexual assaults, but because sexual assault is widely believed to be underreported, these figures do not tell the whole story. The number of reported assaults in the Clery data could arise from many different combinations of reporting rates and true underlying numbers of assaults. Without further assumptions, it is equally plausible that a given school has a large number of assaults and a low reporting rate, or few assaults and a high reporting rate.

To estimate the true incidence and reporting rates of sexual assault in the US college population, we construct a hierarchical Bayesian model of the reported data, together with a Markov Chain Monte Carlo (MCMC) sampling scheme for posterior inference. More generally, underreported count data arises in many disciplines, with applications including criminology (Moreno and Giron, 1998; Fernández-Fontelo et al., 2019) as well as epidemiology (Bailey et al., 2005; Bracher and Held, 2021), traffic safety (Kumara and Chin, 2005; Ma and Li, 2010), and economics (Winkelmann, 1996; Fader and Hardie, 2000). The central challenge in such problems is to disentangle the per-school reporting rate from the true latent counts. This issue can be remedied either through the use of validation data not subject to underreporting, or by using domain knowledge to assign more informative prior distributions to latent variables. In the case of campus sexual assault, fully observed validation data is unavailable, and so we use national crime statistics as a

source of external information. The absence of fully observed validation data also has implications for model assessment. Without knowledge of the true total number of assaults, we use predictive checks applied to held-out data to demonstrate the fitted model's compatibility with the data.

In general, we find that inference for the latent total number of assaults and reporting rate at a given school offer a clearer picture of the campus environment than the reported figures alone. In particular, these estimates can suggest the relative contributions of reporting rate and incidence to year-over-year changes in the reported number of assaults. This is a relevant line of inquiry, given that ex ante an increase in the reported number of assaults could equally be explained by an increase in the reporting rate (a positive outcome, indicating that students are aware of campus support resources) or an increase in the true total number of assaults (a bad outcome).

The remainder of this chapter is structured as follows. Section 1.1 reviews related work, followed by an overview of the dataset in Section 1.2. The proposed modeling methodology is outlined in Section 1.3. Section 1.4 discusses modeling results and insights, and Section 1.5 offers concluding remarks and directions for future work.

1.1 Related Work

There are various perspectives on underreported count data. Some studies consider underreporting as a type of censoring. In this formulation, each observed value is possibly right-censored and has an accompanying latent censoring indicator. de Oliveira et al. (2017) takes this view and assigns a covariate-dependent prior distribution to the latent censoring indicators, including the censoring indicators as a target for posterior inference.

Underreporting also has conceptual overlap with positive-unlabeled (PU) learning. In PU learning, data consist of labeled positive cases and unlabeled cases which may be positive or negative (Letouzey et al., 2000). The recorded number of positive cases is thus a lower bound for the true number of positive cases, which can be viewed as an example of underreporting. The application of PU learning techniques to underreported data is explored in Shanmugam and Pierson (2021) and Wu et al. (2023).

Within the broader literature on underreported count data models, one particular line of work is most relevant to modeling reports of sexual assault. These modeling approaches begin with an assumption that the latent true number of events, z, follows a Poisson distribution with rate parameter λ. It is common to further suppose that each of these z-many events may or may not be reported, independently of all others, with reporting probability p. The number of reported events x, where $x \leq z$, is the data value we observe.

Moreno and Giron (1998) articulates the identifiability issue in this setup: marginally, the observed data $x_1, \ldots, x_n$ follow a Poisson(λp) distribution, so without further constraints or assumptions, one cannot distinguish among the possible pairs (λ, p) with the same product. Moreno and Giron (1998) further demonstrate that placing uninformative prior distributions on λ and p is insufficient to produce useful inference. For instance, the choice of a uniform prior on p and a Jaynes (Jaynes, 1968) or Jeffreys (Jeffreys, 1946) prior on λ results in infinite posterior expectations for λ and z given the observed data. In some research applications this identifiability issue is remedied through the use of validation data not subject to underreporting (Powers et al., 2010; Dvorzak and Wagner, 2016). An alternative approach is to use external information to assign informative prior distributions, as Schmertmann and Gonzaga (2018) and Stoner et al. (2019) do for the reporting probability. Because external information is available regarding both the incidence and reporting rate of sexual assault, we take this latter approach, choosing informative priors for both components of the model.

Earlier work on underreporting models explored the somewhat simpler case of i.i.d. observations governed by a single reporting probability and incidence parameter λ (Moreno and Giron, 1998; Fader and Hardie, 2000). In this chapter we adopt the more recent extension of works such as Dvorzak and Wagner (2016) and Stoner et al. (2019), allowing p and λ to vary across units. When modeling λ_i and p_i, we introduce school features as covariates to account for demographic differences among schools, as detailed in Section 1.3.

Posterior distributions for latent variables in this type of underreporting model are generally intractable. However, choosing conditionally conjugate priors for λ and p (Gamma and Beta

distributions, respectively) leads to analytic expressions for their marginal posterior distributions (Moreno and Giron, 1998; Fader and Hardie, 2000). Alternatively, MCMC approaches to posterior inference allow greater flexibility in choice of priors in exchange for a greater computational burden. Dvorzak and Wagner (2016) use a bespoke implementation of a Gibbs-style sampler, while Stoner et al. (2019) use slice sampling. Our research extends this line of work by demonstrating the viability of gradient-based samplers such as HMC for posterior inference in this model setting.

With respect to the application context, the most relevant work comes from Fernández-Fontelo et al. (2019), who model underreporting of domestic violence complaints across districts in Spain. Unlike the campus sexual assault data, observations in Fernández-Fontelo et al. (2019)'s dataset are dense in the time domain (quarterly for 10 years), motivating the authors to adopt time series methodology.

1.2 Dataset

The dataset contains 11,369 records for 1,973 US colleges and universities over the 2014-2019 time frame [1], where a "record" consists of the total number of assaults reported at a particular school in one particular calendar year. Schools located in the 50 US states, the District of Columbia, and Puerto Rico are included in this analysis. Annual campus crime statistics are furnished by the US Department of Education's Office of Postsecondary Education. All schools that receive federal financial aid funding under Title IV of the Higher Education Act of 1965 are required to submit annual campus crime reports. Most schools have records for all six years in this time frame, though 8% appear in fewer years due to schools opening/closing or beginning/ ceasing to participate in Title IV.

School characteristics and demographic data are available through the National Center for Education Statistics (NCES), which publishes comprehensive institutional profile information through its IPEDS database [2]. Additional dataset details are included in Appendix A.1.

[1] Data from 2020 and 2021 are excluded. Because the COVID-19 pandemic significantly disrupted patterns of social interaction, campus sexual assault data from this time are not well described by the same model used for other years.

[2] `https://nces.ed.gov/ipeds/`

We exclude institutions with no degree-granting programs and institutions with no programs classified as "academic" in nature. We also exclude institutions with no residential housing facilities. For such schools, the school does not represent the students' community, and we would not reasonably expect students to report sexual assaults to the school.

Institutions in this dataset range in size from 5 to 79,500 in-person students, with a median of 2,350. Public institutions comprise 41% of the dataset, religiously affiliated institutions 36%, and private, non-religiously affiliated institutions 22%. The reported sexual assault data is notably sparse: 42% of records show zero reported assaults, and 22% of institutions show no assaults across all six years. The distribution of reported values, shown in Figure 1.1, has a long right tail, with a handful of schools reporting more than 100 assaults in a given year. This sparsity pattern is not explained by student population size alone: schools with as many as 36,000 students had zero reported assaults, while assaults *were* reported at schools with as few as several hundred students.

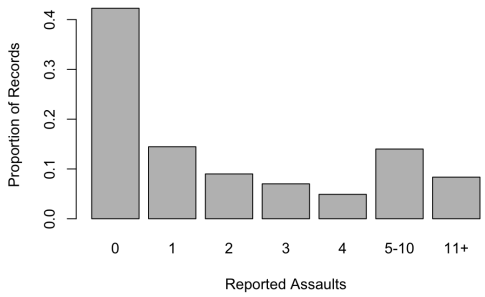


Figure 1.1: School-level reported sexual assault data is sparse, with 42% of records showing zero assaults, and a median of one assault.

The total number of assaults reported across schools increased steadily from 2014 to 2018, growing by nearly 50% before falling in 2019 (see Figure 1.2). The total in-person student population was relatively stable over this time period, remaining close to 10.6 million.

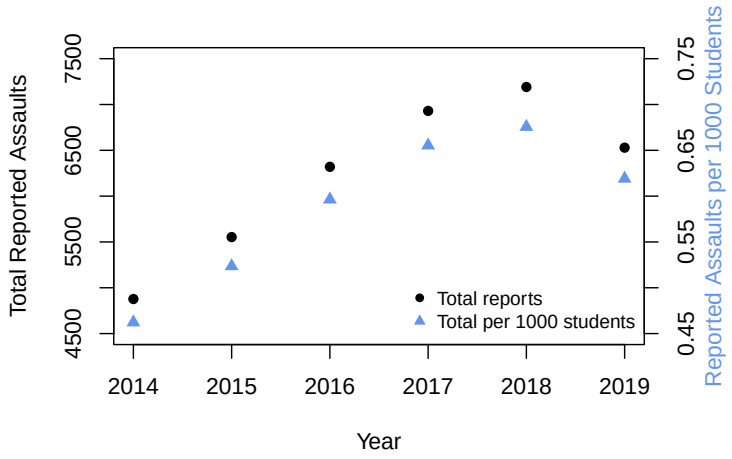


Figure 1.2: The total number of reported assaults grew from 2014 to 2018. Total student population fluctuated relatively little, resulting in per capita reporting trends moving with the overall total number of reports.

Campus crime numbers in this dataset are a direct tally of the individual assaults for which a student voluntarily made a formal report to campus authorities. Note that those reports do not arise from a survey process or broad inquiry about student experiences; on its own, this dataset contains no information about underreporting. For complementary information about reporting rates, we consider the National Crime Victimization Survey (NCVS). The NCVS is a large-scale annual survey conducted on behalf of the behalf of the US Bureau of Justice Statistics to determine incidence rates of personal and property crimes, including sexual assault. This survey dates back 50 years, and one of its primary objectives is to quantify the extent of crime not reported to authorities. By attempting to measure crimes not reported to authorities, the NCVS provides additional context which is absent from the college campus Clery Report data.

1.3 Methodology

For a particular school i in a particular year j, suppose that the true, unknown number of assaults z_{ij} comes from a Poisson distribution with rate parameter λ_{ij}. Further suppose that each of those z_{ij} assaults is independently reported, or not reported, with probability of reporting equal

to p_{ij}. This produces the reported value x_{ij} that we observe, according to a binomial thinning process:

$$x_{ij}|z_{ij}, p_{ij} \sim \text{Binom}(z_{ij}, p_{ij}). \tag{1.1}$$

With this likelihood model, we construct a hierarchical generative model for the reported campus sexual assault data. The proposed hierarchical model, diagrammed in Figure 1.3, entails a component describing the reporting probability, and a component describing the Poisson rate governing the true number of assaults. A common approach in the underreporting literature is to model $\log(\lambda_{ij})$ and/or $\log\left(\frac{p_{ij}}{1-p_{ij}}\right)$ as deterministically equal to a linear function of covariates (Dvorzak and Wagner, 2016; Stoner et al., 2019; de Oliveira et al., 2021). To provide additional flexibility, we instead model the Poisson rate and reporting probability as random functions of their respective covariates. Full details of each model component are discussed below.

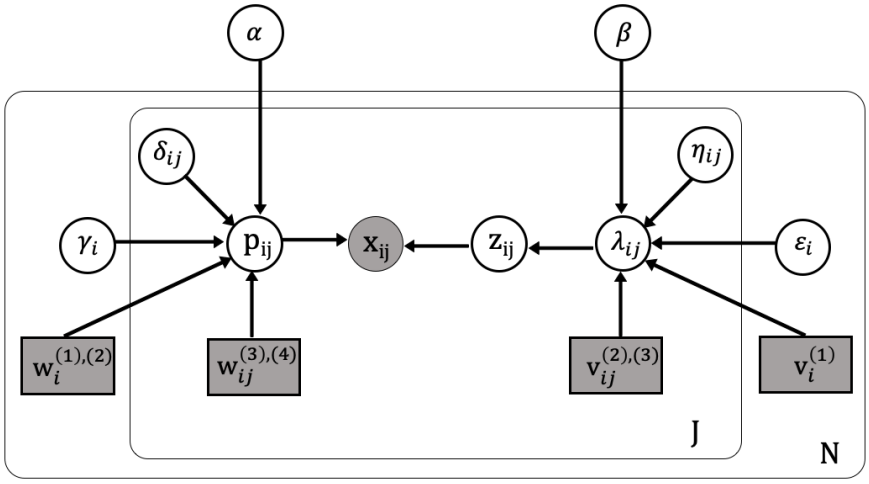


Figure 1.3: Graphical model diagram

1.3.1 Modeling z

For each record in the dataset, we suppose the true number of assaults arises from a Poisson distribution parameterized by λ_{ij}, where λ_{ij} depends on observed characteristics $\mathbf{v}$ of school i in year j as follows:

$$
\begin{aligned}
z_{ij}|\lambda_{ij} &\sim \text{Poisson}(\lambda_{ij}) && (1.2)\\
\log(\lambda_{ij}) &= \beta_{0,\mathrm{v}_i^{(1)}} + \beta_1 \mathrm{v}_{ij}^{(2)} + \beta_2 \mathrm{v}_{ij}^{(3)} + \varepsilon_i + \eta_{ij} && (1.3)\\
\beta_{0,1}, \beta_{0,2}, \beta_{0,3} &\sim N(-5.5, 0.5^2)\\
\beta_1 &\sim N(1, 0.1^2)\\
\beta_2 &\sim N(0, 4^2)\\
\varepsilon_i &\sim N(0, 0.75^2)\\
\eta_{ij} &\sim N(0, 0.1^2)
\end{aligned}
$$

The covariates associated with λ are:

$\mathrm{v}_i^{(1)}$: degree of urbanization of school i's campus
1 = urban
2 = suburban
3 = rural

$\mathrm{v}_{ij}^{(2)}$: log(number of in-person students) for school i in year j

$\mathrm{v}_{ij}^{(3)}$: (fraction of women in student body at school i in year j - 0.5)^2

This generative model considers in-person student population, degree of campus urbanization, and student body gender composition[3] as covariates of interest. The functional form of equations (1.2) - (1.3) implies that for the expected number of assaults we have:

$$
\begin{aligned}
\mathbb{E}(z_{ij}|\beta, \mathbf{v}) &\propto \exp(\beta_{0,\mathrm{v}_i^{(1)}} + \beta_2 \mathrm{v}_{ij}^{(3)}) \exp(\beta_1 \log(\text{student population}_{ij}))\\
&\propto (\text{student population}_{ij})^{\beta_1}. && (1.4)
\end{aligned}
$$

This modeling choice allows for the possibility that the number of assaults does not scale linearly with the number of in-person students. Because the number of interpersonal interactions does not always scale linearly with community size, this type of "power law" dynamic has been observed for various social phenomena, including crime (Bettencourt et al., 2007; Chang et al.,

[3]The gender categories reported by NCES are currently "male" and "female".

2018).

Incorporating separate intercepts in (1.3) allows for different incidence patterns at urban, suburban, and rural schools. This decision is motivated by the national differences in violent crime victimization rates across locales of varying urbanization (Truman and Langton, 2014, 2015; Truman and Morgan, 2016; Morgan and Kena, 2018; Morgan and Truman, 2018; Morgan and Thompson, 2020).

Gender composition of the student body is included as a covariate because most perpetrators of sexual assault are men, while most victims are women (Sinozich and Langton, 2014). We center the proportion of women in the student body by subtracting 0.5, and square this value, to allow for the possibility that schools with more extreme gender imbalances may have different rates of assault.

Per-school intercepts ε_i allow for between-school differences in assault incidence patterns which are not otherwise captured by covariates. Per-record noise terms η_{ij} allow for within-school differences across years. Supposing that $\beta_1 = 1$ (linear relationship between student population and number of assaults) and $\beta_2 = 0$ (no relationship between student body gender composition and expected number of assaults), priors on ε, η, and intercepts β_0 are set so that the resulting distribution over assaults per 1000 persons roughly corresponds with external estimates from the National Crime Victimization Survey (NCVS). The resulting distribution is displayed in Figure 1.4. Note that, conditional on covariates, the Poisson rate λ_{ij} follows a lognormal distribution, leading to a number of assaults z_{ij} whose marginal distribution is overdispersed relative to a Poisson.

	18-20	21-24	12+
2014	1.2*	3.8	1.1
2015	3.2	1.6*	1.6
2016	3.8	1.8	1.1
2017	8.6	2.4	1.4
2018	10.1	6.0	2.7
2019	7.1	6.4	1.7

Table 1.1: Sexual Assault Incidence per 1000 persons in the US, by age group

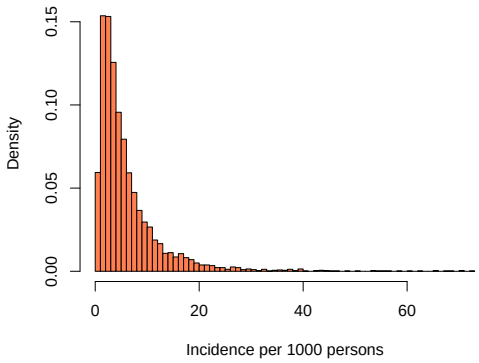


Figure 1.4: Prior distributions on β_0, ε, and η imply a distribution over incidence centered at 4.1 per 1000 persons, with a central 50% range of 2.3 to 7.4.

The NCVS is a large sample survey conducted annually on behalf of the US Bureau of Justice Statistics to determine incidence rates of personal and property crimes. This survey includes crimes reported and not reported the police. Over the 2014-2019 time period, the estimated incidence of sexual assault in the overall population aged 12 and older ranged from 1.1 to 2.7 per 1000 persons. As shown in Table 1.1, estimated incidence skewed higher in the 18-20 and 21-24 age groups, which are particularly relevant to the college student population.

1.3.2 Modeling p

As formulated in Equation 1.1, the reported number of assaults arises from the latent true number of assaults z and the reporting probability p. Having discussed a model for z, we now turn to modeling the reporting probability p. Of the total z_{ij} assaults occurring at school i in year j, suppose that each individual assault is reported, independently, with probability p_{ij}, where p_{ij}

* Per BJS: "[estimate is] based on 10 or fewer sample cases, or coefficient of variation is greater than 50%"

depends on observed characteristics $\mathbf{w}$ of school i in year j as follows:

$$\begin{aligned}
\log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \alpha_0 + \alpha_1 \mathrm{w}_i^{(1)} + \alpha_2 \mathrm{w}_i^{(2)} + \alpha_3 \mathrm{w}_{ij}^{(3)} + \alpha_4 \mathrm{w}_{ij}^{(4)} + \gamma_i + \delta_{ij} \qquad (1.5)\\
\alpha_0 &\sim N(-1.25, 0.5^2)\\
\alpha_1, \alpha_2, &\sim N(0, 2^2)\\
\alpha_3, \alpha_4 &\sim N(0, 4^2)\\
\gamma_i &\sim N(0, 1.25^2)\\
\delta_{ij} &\sim N(0, 0.5^2).
\end{aligned}$$

The covariates associated with p are:

$\mathrm{w}_i^{(1)}$: school i issues associate degrees only ($\in \{0, 1\}$)
$\mathrm{w}_i^{(2)}$: school i substantially engaged in religious instruction ($\in \{0, 1\}$)
$\mathrm{w}_{ij}^{(3)}$: (women as fraction of student body at school i in year j) - 0.5
$\mathrm{w}_{ij}^{(4)}$: Pell grant recipients as fraction of student body at school i in year j, median centered

The reporting probability at school i in year j accounts for two school-level covariates that are fixed across years, and two that vary. The first fixed school-level covariate is $\mathrm{w}_i^{(1)}$, an indicator for whether school i is a junior college, issuing only associate degrees as opposed to bachelor degrees or higher. The reporting probability also incorporates an indicator $w_i^{(2)}$ for whether school i is substantially engaged in religious instruction. A theological seminary, for instance, would satisfy this definition, whereas a religiously affiliated liberal arts college would not. The inclusion of these school-level random effects encourages the probabilities for an individual school to exhibit similarity across years. Junior colleges comprise roughly 13% of institutions in the dataset, and account for 7% of total in-person students. Institutions of religious instruction make up a relatively smaller portion of the dataset, and generally have small student populations. These two categories of institution have negligible overlap, as detailed in Figure 1.5.

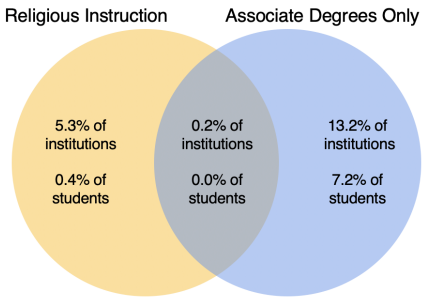


Figure 1.5: Junior colleges and institutions of religious instruction comprise a minority of the dataset

Gender composition of a school's student body also contributes to the reporting probability, and varies across years. Many institutions in this dataset have approximately gender-balanced student populations. Across records, the proportion of female students is centered around 57%, with smaller peaks at each end of the spectrum corresponding to all-male and all-female schools. The final covariate incorporated in the reporting rate is the fraction of Pell grant recipients in the student body. Pell grants are federal financial aid awarded to students on the basis of exceptional financial need. This variable is included as a proxy for socioeconomic status, which research indicates may be correlated with reporting rates (Fisher et al., 2003; Sabina and Ho, 2014). The median percentage of undergraduates receiving Pell grants is 36%, and values are more widely spread over the unit interval compared to the gender composition composition covariate.

As in the model of Poisson rate λ_{ij}, per-school intercepts γ_i in Equation 1.5 allow for between-school differences in assault incidence patterns not captured by covariates, while per-record noise terms δ_{ij} allow for within-school differences across years. Supposing that $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$, priors on ε, η, and intercepts β_0 are chosen to be broadly consistent with NCVS estimates of the percentage of sexual assaults reported to police. Over the 2014-2019 time period, the estimated rate of reporting to police ranged from 23% to 40% in the overall 12+ population, while reporting rates for the 18-20 and 21-24 age groups tended to be lower (see Table 1.2). The induced prior distribution over values of p, when $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$, is depicted in Figure 1.6 below.

	18-20	21-24	12+
2014	0.0*	16.48	33.6
2015	10.6*	26.0*	32.5
2016	10.2*	31.0*	23.2
2017	45.3	38.3*	40.4
2018	11.0*	10.2*	24.9
2019	17.5*	24.7*	33.9

Table 1.2: Percentage of Sexual Assaults Reported to Police, by age group

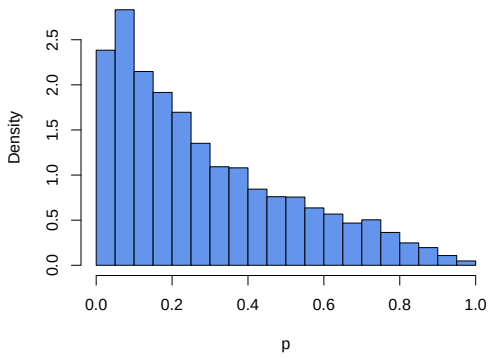


Figure 1.6: Prior distributions on α_0, γ, and δ imply a distribution over reporting probability centered at 22%, with a central 50% range of 10% to 43%.

In these models of z and p, global coefficients α and β encourage the sharing of information across schools, while school-level terms γ and ε capture local properties of individual schools. This partial pooling model structure is a compromise between complete pooling, in which no school-level differences are permitted (γ and ε omitted), and no pooling, in which each school i is modeled separately from all others, with its own coefficient vectors α_i and β_i. Compared to the complete-pooling and no-pooling alternatives, the partial pooling model assigns higher likelihood to held-out observational data- refer to Appendix A.3 for details.

1.3.3 Inference

The model does not admit closed form expressions for posterior quantities of interest. We draw samples from the joint posterior distribution over latent variables via MCMC. The true number of assaults z_{ij} is a discrete latent variable, which presents an obstacle to gradient-based sampling schemes such as Hamiltonian Monte Carlo (HMC). However, this can be circumvented by marginalizing out z_{ij}. Instead of working with z_{ij}, one can take advantage of the fact that conditional on the event rate λ_{ij} and reporting probability p_{ij}, but not the true number of assaults z_{ij},

the reported number of assaults follows a Poisson distribution:

$$x_{ij}|\lambda_{ij}, p_{ij}, \mathbf{w}_{ij}, \mathbf{v}_{ij} \sim \text{Poisson}(\lambda_{ij} p_{ij}).$$

See Appendix A.2 for derivation. This marginalizes out the discrete latent variable z_{ij}, allowing for gradient-based MCMC sampling from the posterior $p(\boldsymbol{\lambda}, \mathbf{p}|\mathbf{x}, \mathbf{w}, \mathbf{v})$. However, the latent true number of assaults z_{ij} is also a quantity of interest. To address this, first note the conditional distribution of the number of unreported assaults, u_{ij} (where $u_{ij} \stackrel{\text{def}}{=} z_{ij} - x_{ij}$):

$$u_{ij}|x_{ij}, \lambda_{ij}, p_{ij} \sim \text{Poisson}(\lambda_{ij}(1 - p_{ij})).$$

Because u_{ij} is conditionally independent of the covariates $\mathbf{v}_{ij}$ and $\mathbf{w}_{ij}$ given latent parameters λ_{ij} and p_{ij}, it is equivalently true that

$$u_{ij}|x_{ij}, \lambda_{ij}, p_{ij}, \mathbf{w}_{ij}, \mathbf{v}_{ij} \sim \text{Poisson}(\lambda_{ij}(1 - p_{ij})).$$

Each MCMC sample $(\lambda_{ij}^{(s)}, p_{ij}^{(s)})$ drawn from (approximately) $p(\lambda_{ij}, p_{ij}|x_{ij}, \mathbf{w}_{ij}, \mathbf{v}_{ij})$ can be augmented by sampling a corresponding $u_{ij}^{(s)}$ from $p(u|x_{ij}, \lambda_{ij}, p_{ij})$. The sampled value $u_{ij}^{(s)}$ can then be added to the observed count x_{ij} to produce a sample for the latent count $z_{ij}^{(s)}$. Altogether, the sample $(\boldsymbol{\lambda}^{(s)}, \mathbf{p}^{(s)}, \mathbf{z}^{(s)})$ constitutes an MCMC sample from $p(\boldsymbol{\lambda}, \mathbf{p}, \mathbf{z}|\mathbf{x})$. This sampling procedure is outlined in Algorithm 1, and can be conveniently implemented in a probabilistic programming language. Posterior inference for this work was conducted using HMC in Stan (Stan Development Team, 2023).

Algorithm 1: Posterior Sampling Scheme

Input:

- Data $\mathbf{x}$
- Covariates $\mathbf{w}, \mathbf{v}$
- Index set $\mathcal{I}$ (pairs (i,j) for which the dataset contains a record x_{ij}
- MCMC algorithm `SAMPLER` (e.g. HMC)
- Markov chain length S

Output: samples from the posterior $p(\mathbf{z}, \boldsymbol{\lambda}, \mathbf{p}|\mathbf{x}, \mathbf{w}, \mathbf{v})$;

for *s in 1:S* **do**

  Draw $(\boldsymbol{\lambda}^{(s)}, \mathbf{p}^{(s)})$ from `SAMPLER` approximating $p(\boldsymbol{\lambda}, \mathbf{p}|\mathbf{x}, \mathbf{w}, \mathbf{v})$;

  for (i, j) *in* $\mathcal{I}$ **do**

    Draw $u_{ij} \sim \text{Poisson}((1 - p_{ij}^{(s)})\lambda_{ij}^{(s)})$;

    Set $z_{ij} = u_{ij} + x_{ij}$;

  end

end

return $\left\{(z^{(s)}, \lambda^{(s)}, p^{(s)})\right\}_{s=1}^{S}$

1.3.4 Model Assessment

Fully-observed data including the true total number of assaults is not available, and thus cannot be used for model validation. Instead, the predictive distribution provides information about the quality of the model fit. The posterior distribution of the latent variables implies a predictive distribution over future observations $p(x^{new}|x)$ and, accordingly, over summary statistics of future observations. This motivates the notion of posterior predictive checks (Guttman, 1967; Box, 1980; Rubin, 1984; Gelman et al., 1996): that a satisfactory model will yield predictive distributions over key summary statistics under which the actual values observed in the dataset are not extreme. As detailed in Moran et al. (2024) and Li and Huggins (2022), such posterior predictive checks can be overly optimistic about model fit due to double use of the data (both to fit the model and to evaluate it). The aforementioned works propose randomly reserving a portion of the original dataset to use exclusively for model assessment, an approach that demonstrably achieves a more accurate gauge of model performance by avoiding the pitfall of data reuse. In the spirit of this suggestion, roughly 20% of entries in the campus sexual assault dataset are held out, with the remaining 80% used for posterior inference. Held out data points are sampled at the level of (school, year) pairs. The held out data set includes records from

all years in scope (2014-2019), and does not contain data from any additional schools not seen in the main sample. Predictive samples for the held out data points are generated as follows.

Algorithm 2: Predictive Distribution Sampling

Input:

- Index set $\mathcal{I}$ (pairs (i,j) for which the held out dataset contains a record x_{ij}^{new})
- Covariates $\mathbf{w}^{new}, \mathbf{v}^{new}$ for held out dataset
- Posterior distribution $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\varepsilon} | \mathbf{x}, \mathbf{w}, \mathbf{v})$

Output: Sample dataset from the predictive distribution $p(\mathbf{x}^{new} | \mathbf{x}, \mathbf{w}, \mathbf{v}, \mathbf{w}^{new}, \mathbf{v}^{new})$;

Draw $(\boldsymbol{\alpha}^s, \boldsymbol{\beta}^s, \boldsymbol{\gamma}^s, \boldsymbol{\varepsilon}^s)$;
for (i, j) *in* $\mathcal{I}$ **do**
  Draw $\delta_{ij} \sim N(0, 0.5)$;
  Draw $\eta_{ij} \sim N(0, 0.2)$;
  if *school* $i \in$ *training data* **then**
    $\gamma_i \leftarrow \gamma_i^s$;
    $\varepsilon_i \leftarrow \varepsilon_i^s$;
  else
    Draw $\gamma_i \sim N(0, 1)$;
    Draw $\varepsilon_i \sim N(0, 0.5)$;
  end
  $p_{ij} \leftarrow \text{invlogit}(\alpha_0^s + \langle \boldsymbol{\alpha}^s, \mathbf{w_{ij}}^{new} \rangle + \gamma_i + \delta_{ij})$;
  $\lambda_{ij} \leftarrow \exp(\beta_{0,v_i}^s + \langle \boldsymbol{\beta}^s, \mathrm{v}_{ij}^{new} \rangle + \varepsilon_i + \eta_{ij})$;
  Draw $z_{ij} \sim \text{Poisson}(\lambda_{ij})$;
  Draw $x_{ij}^{new} \sim \text{Binom}(z_{ij}, p_{ij})$;
end
return $\mathbf{x}^{new}$

As noted in Section 1.2, each year many schools had zero reported assaults, and the median number of reports was 1. Generating 10,000 resampled datasets according to Algorithm 2 and calculating the proportion of records with zero reported assaults produces the distribution shown below in Figure 1.7(a). Within the held out dataset 42.1% of records had zero reported assaults, while values ranged from roughly 38% to 44% in datasets sampled from the predictive distribution.

The distribution of reported assault numbers has been pulled slightly away from the extreme end, such that the observed proportion of zeroes falls toward the high end of the predictive distribution. However, the median number of assaults was equal to 1 in all datasets sampled from the predictive distribution, and the true proportion of held-out records x_{ij} with ≤ 1 assaults sits comfortably within its respective predictive distribution (Figure 1.7(b)). Overall, the model adequately captures the bottom-heaviness of the distribution of reported assault numbers.

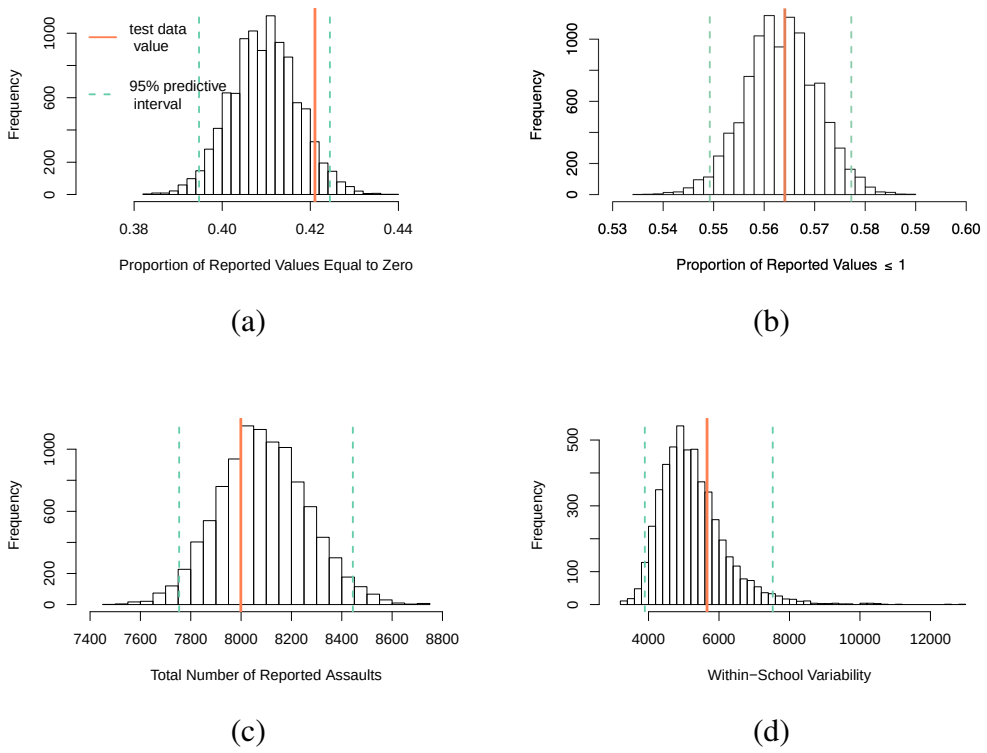


Figure 1.7: Summary statistics of the held out data are plausible under the predictive distribution. Solid lines indicate the value calculated on the held out data, dashed lines indicate the 2.5th and 97.5th percentiles of the predictive distribution; (a) Proportion of reports indicating zero assaults (b) Proportion of reports indicating ≤ 1 assaults (c) Total number of reported assaults across all schools and all years (d) Within-school variability of reported numbers over time

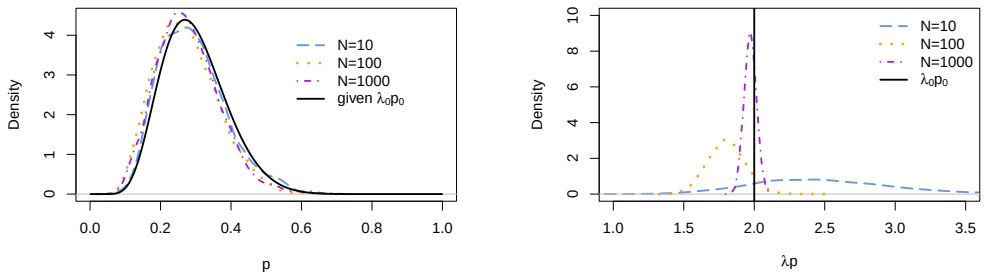
Other quantities of interest are the total number of reported assaults $\sum_{ij} x_{ij}^{new}$ and the amount of within-school variability. The predicted number of assaults reported by an individual school should not be too rigid or too flexible over time, relative to the trends seen in the actual data.

This summary statistic is quantified as $\sum_i \text{var}(x_{i\cdot}^{new})$. The total number of reported assaults and the within-school variability observed in the held out dataset are both plausible under their respective predictive distributions, shown in Figure 1.7(c)-(d), suggesting that the model adequately captures these characteristics of the data.

1.3.5 Capabilities and Limitations

Let us pause to examine what this estimation method can and cannot do in terms of resolving the central identifiability issue introduced at the beginning of this chapter and in Section 1.1. First, consider a toy example in which $z_1, \ldots, z_N \overset{iid}{\sim} \text{Poisson}(\lambda_0)$ are latent and $x_i \sim \text{Binom}(z_i, p_0)$, $i = 1, \ldots, N$ are observed. The observed dataset identifies the product $\lambda_0 p_0$, while only the prior distributions provide information about λ_0 and p_0 individually. The posterior distribution of the product $\lambda_0 p_0$ concentrates as the number of observations increases. The posterior distributions of λ_0 and p_0, however, do not become arbitrarily concentrated in the infinite data limit, but rather approach $p(\lambda_0|\lambda_0 p_0)$ and $p(p_0|\lambda_0 p_0)$ respectively, as demonstrated in Figure 1.8.

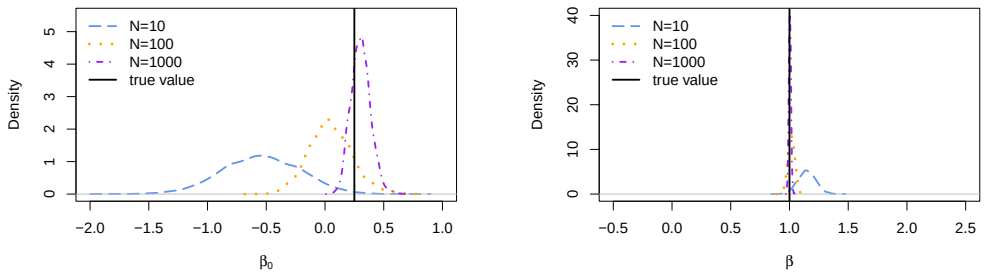
Furthermore, the posterior distribution $p(\lambda_0|x_1, \ldots, x_N)$ will not necessarily be centered at the true value λ_0 (and likewise for p_0); the data provide information about the product $\lambda_0 p_0$, but conditional on $\lambda_0 p_0$, posterior estimates of λ_0 and p_0 are determined by their respective prior distributions.



(a) (b)

Figure 1.8: The posterior distribution of $\lambda_0 p_0$ concentrates as the amount of observational data increases, but the posterior distribution of p_0 has some amount of uncertainty that cannot be eliminated by observational data; (a) Kernel density estimates of posterior samples of p_0, conditioned on observed datasets of size N, with a solid line representing $p(p_0|\lambda_0 p_0)$, (b) Kernel density estimates of posterior samples of $\lambda_0 p_0$, conditioned on observed datasets of size N

A similar relationship emerges when the toy model is extended to include covariates for λ and p. Let $\log(\lambda_i) = \beta_0 + \beta \mathrm{v}_i$ and $\log(\frac{p_i}{1-p_i}) = \alpha_0 + \alpha \mathrm{w}_i$ for $\mathrm{v}_i, \mathrm{w}_i \in \mathbb{R}$. For $i = 1, \ldots, N$ latent counts are generated as $z_i \sim \text{Poisson}(\lambda_i)$, and we observe the reported counts $x_i \sim \text{Binom}(z_i, p_i)$. Performing posterior inference on data simulated from this model reveals that increasing the amount of observational data is more effective at reducing posterior uncertainty about the slope parameters β and α than about the intercepts β_0 and α_0. Figure 1.9 illustrates the relative difficulty of learning one of the intercept parameters from data.



(a) (b)

Figure 1.9: Additional observational data is more effective at reducing posterior uncertainty about the slope coefficient β than about the intercept ; (a) Kernel density estimates of posterior samples of β_0, conditioned on observed datasets of size N (b) Kernel density estimates of posterior samples of β_0, conditioned on observed datasets of size N, displayed on x-axis interval of equal width as left-hand plot

In light of these dynamics, Appendix A.5 explores sensitivity to the prior information elicited from NCVS estimates. While some reasons for not reporting a sexual assault are less relevant in a survey setting (such as fear of retaliation or lack of faith that the perpetrator will be held accountable), it is nevertheless plausible that some NCVS respondents choose not to disclose their sexual assault victimization. In line with the discussion in this section, inferences about the role of covariates are relatively more stable if the extent of underreporting is larger than estimated by NCVS, while inferences about the true overall incidence and reporting rate exhibit more sensitivity.

1.4 Results

1.4.1 Role of Covariates

Sampling from the model in Section 1.3 yields the posterior means and quartiles given in Table 1.3 below.

Variable	Mean	25%	Median	75%	R hat†
β_1	0.82	0.81	0.82	0.84	1.00
β_2	-4.05	-4.53	-4.05	-3.56	1.00
$\beta_{0,1}$	-4.54	-4.67	-4.54	-4.42	1.00
$\beta_{0,2}$	-4.40	-4.53	-4.40	-4.28	1.00
$\beta_{0,3}$	-4.32	-4.45	-4.32	-4.21	1.00
α_0	-1.43	-1.50	-1.43	-1.36	1.00
α_1	-1.97	-2.06	-1.97	-1.88	1.00
α_2	-2.48	-2.73	-2.48	-2.23	1.00
α_3	0.49	0.26	0.49	0.73	1.00
α_4	-3.00	-3.18	-3.00	-2.83	1.00

†R hat, also known as the potential scale reduction factor, is a convergence diagnostic proposed in Gelman and Rubin (1992)

Table 1.3: Posterior means and quartiles of latent variables

The posterior distribution for β_1 is concentrated around 0.82, which corresponds to sub-linear growth in the expected number of assaults as a function of number of students. The proposed the power law relationship between student population and expected number of assaults (1.4) implies a similar relationship between student population and the expected per-capita rate of assaults, $\frac{z_{ij}}{\text{student population}_{ij}}$:

$$\mathbb{E}\left(\frac{z_{ij}}{\text{student population}_{ij}}|\boldsymbol{\beta}, \mathbf{v}\right) \quad \propto \quad (\text{student population}_{ij})^{\beta_1 - 1}.$$

Consequently, the posterior distribution suggests an inverse relationship between the size of the student population and the expected per-capita number of assaults in a given year. With all other covariates held equal, we expect a school to have 13% more assaults per capita than an equivalent school twice its size, and 28% more than a school 4 times its size. Similarly, because the posterior distribution suggests $\beta_1 < 1$, the variance of the per capita number of assaults, $\text{var}\left(\frac{z_{ij}}{\text{student population}_{ij}}|\boldsymbol{\beta}, \mathbf{v}\right)$, scales inversely with the number of students. With other covariates held constant, the expected number of assaults per capita is more variable at small schools than at large ones.

Posterior distributions for the intercepts $\beta_{0,1}$, $\beta_{0,2}$, and $\beta_{0,3}$ are similar to each other. Results are

mildly suggestive of higher sexual assault incidence in rural areas than in cities, but do not reveal a substantial difference in expected number of assaults based on degree of campus urbanization. The negative values for β_2 suggest that a gender imbalance in a school's student body is associated with lower expected incidence of sexual assault, while $\alpha_3 > 0$ suggests that a higher proportion of female students corresponds to a higher reporting probability. Conversely, a higher proportion of students receiving Pell grants (a proxy for lower socioeconomic status) appears associated with a lower probability of reporting. Negative estimates for α_1 and α_2 signal lower expected reporting probabilities for junior colleges and institutions of religious instruction.

1.4.2 Systemwide Results

Drawing posterior samples of the true number of assaults at each school induces a distribution over the total number of assaults across schools each year. The posterior median for total assaults ranges from 2.6 to 2.8 per 1000 persons over the 2014-2019 time period. Although the median estimated incidence is higher in later years compared to 2014, posterior uncertainty is large enough that the true trend in incidence could conceivably be flat or even decreasing, as can be seen in Figure 1.10 below.

Figure 1.10: Posterior distribution of total incidence per 1000 persons in the college population does not exhibit a clear pattern of year-over-year change.

Posterior estimates of the true reporting rate, however, show more movement over time. Figure 1.11 depicts posterior summaries for the actual number of reported assaults across schools as a

fraction of the estimated true number of assaults. The posterior median reporting rate was lowest in 2014, at 17.3%, and highest in 2018 at 24.2%, with reporting rates trending upward over those years.

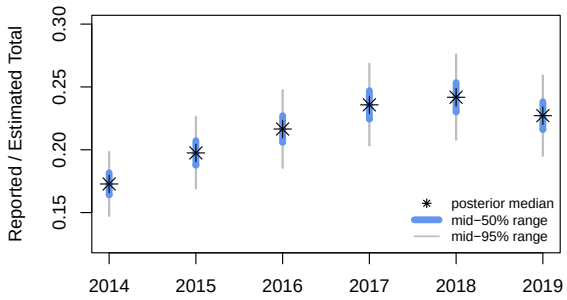
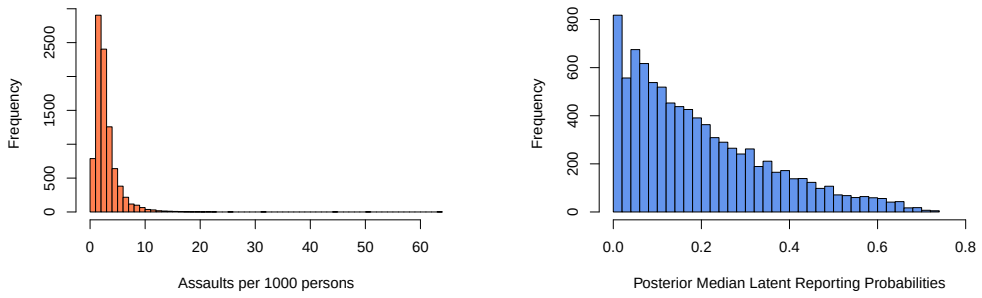


Figure 1.11: Posterior estimates of the overall reporting rate in the college population exhibit an increasing trend from 2014 to 2018.

Taken together, these two results suggest that the increase in total sexual assaults reported nationwide over the 2014-2019 period is more likely attributable to an increase in reporting rates than an increase in the true number of assaults that occurred.

1.4.3 School-Level Results

While the systemwide increase in reported assaults may be more easily explained by an increase in reporting rates, aggregate trends are not necessarily representative of the dynamics at an individual school. Examining the posterior medians of the reporting rate p_{ij} and true incidence per 1000 persons across all records in the dataset reveals considerable heterogeneity, as displayed in Figure 1.12.



(a) (b)

Figure 1.12: Per-record posterior medians for incidence and reporting rate exhibit heterogeneity; (a) Posterior medians for incidence are centered slightly below the overall population value at 2.3 per 1000, but 25% of records have posterior median incidence greater than 3.4, and another 25% below 1.6. (b) Posterior medians for reporting probabilities are predominantly below 30%, but for some records are as high as 74%.

While sexual assault is susceptible to underreporting at all schools, the estimated degree of underreporting varies widely. Figure 1.12(b) illustrates that some schools have extremely low reporting probabilities, while others are considerably higher than the population-level reporting rates published in the NCVS. This variation across schools has implications for how individual schools should interpret their reported crime statistics. With only the prior belief that sexual assault is significantly underreported, one might be equally inclined to attribute an increase in reported assaults at any school to an increase in the reporting probabilities. However, this may lead to flawed conclusions.

Suppose that for a particular school, the true number of assaults in a future year, not included in the dataset, remains the same as the true number of assaults in 2019. With an unchanged number of true assaults, how large of a change in the reported number of assaults is plausible due to variation in the reporting probability alone? For school i, the predictive distribution for a future year's number of reported assaults can be sampled as follows.

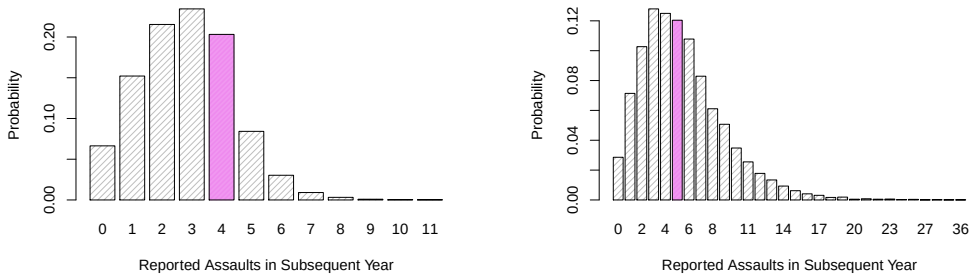
(1) Draw $(z_{i,2019}^{(s)}, \alpha_0^{(s)}, \boldsymbol{\alpha}^{(s)}, \gamma_i^{(s)})$ from model's posterior distribution

(2) Draw $\delta^{new} \sim N(0, 0.5)$

(3) Set $p^{new} = \text{invlogit}(\alpha_0^{(s)} + \langle \boldsymbol{\alpha}^{(s)}, \mathbf{w}_{i,2019} \rangle + \gamma_i^{(s)} + \delta^{new})$

(4) Draw $x^{new} \sim \text{Binom}(z_{i,2019}, p^{new})$

Randolph College and the University of South Carolina-Columbia (USCC) had a similar number of reported assaults in 2019 (4 and 5 respectively). However, Randolph College's estimated reporting rate was considerably higher, with a posterior median of 53% compared to only 14% at USCC. Assuming that the true number of assaults is held constant, the predictive distributions for the number of reported assaults are shown in Figure 1.13.



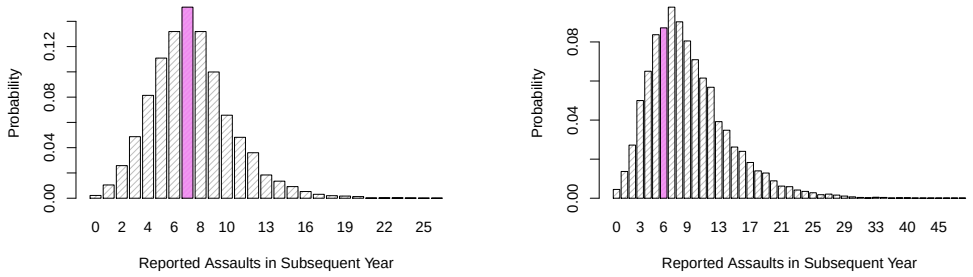
(a) (b)

Figure 1.13: Variation in reporting probability alone is less likely to produce an increased number of reported assaults in subsequent year at (a) Randolph College than at (b) University of South Carolina- Columbia. Solid-colored bar corresponds to number of reported assaults in 2019.

Under this scenario, there is a roughly 13% chance of an increased number of reported assaults at Randolph College in a subsequent year due to variation in the reporting rate, and less than a 1% chance of the reported number to double to 8 or more. At USCC, variation in reporting rate has a comparatively larger chance of producing an increased number of reported assaults (42%), and a 12% chance of doubling the number of reported assaults to 10 or more.

Ursinus College and Temple University provide another illustration of such school-level differences. Of the two, Ursinus had a much higher estimated reporting probability in 2019, and changes in reporting probability are comparatively less likely to drive an increase in reported as-

saults at Ursinus compared to Temple, as can be seen from their respective predictive distributions in Figure 1.14.



(a) (b)

Figure 1.14: Without a change in the true number of assaults, the number of reported assaults has a 44% chance of increasing in a subsequent year at (a) Ursinus College, and a 67% chance at (b) Temple University.

These disparities underscore the opaque nature of reported campus sexual assault data and the potential pitfalls of a one-size-fits-all approach to interpreting school-level reports over time.

1.5 Discussion

Colleges and universities prioritize reducing the incidence of sexual assault on their campuses, but underreporting diminishes the interpretability of their published sexual assault incidence data. This chapter constructs a generative model of underreported campus sexual assault data which allows for estimation of true incidence and reporting rates. Fitting this model reveals that lower socioeconomic status, as measured by the percentage of a school's undergraduates receiving Pell grants, appears associated with a lower probability of reporting. Status as a junior college or institution of religious instruction appear similarly associated with lower reporting probability, while gender imbalance in a school's student body is associated with lower true incidence of assault.

For the overall college population, estimated incidence of sexual assault remained fairly stable over 2014-2019, while reporting probabilities increased. Results at individual schools, however, varied widely. Estimates of school-level incidence and reporting probabilities may help university

officials, for instance Title IX coordinators, assess the effectiveness of initiatives to reduce incidence of sexual assault or to familiarize students with reporting resources, and decide where to make changes and improvements moving forward.

One avenue for future work concerns repeat victimizations of the same individual. High-frequency series victimizations, for instance in the case of patterns of intimate partner violence, can produce extreme values in the reported number of assaults. As an example, in 2017 the University of Nebraska-Lincoln received 104 reports of sexual assault corresponding to the same victim and perpetrator. Such extreme outliers are not easily accommodated, and in particular may violate the modeling assumption that the reporting decisions for each assault are independent (conditional on reporting probability). Violations of the independence assumption can degrade the quality of posterior inferences- see Appendix A.4 for further details. More broadly, the Bureau of Justice Statistics has acknowledged the difficulty of accurately incorporating series victimization in its national crime estimates (Lauritsen et al., 2012).

Chapter 2: Detecting Voting Irregularities in Ghana's 2020 Presidential Election

In Ghana's 2020 presidential election, incumbent Nana Akufo-Addo of the New Patriotic Party (NPP) ran for reelection against main challenger John Mahama of the National Democratic Congress (NDC), along with 10 additional candidates. Akufo-Addo was declared the winner, with 51.3% of the popular vote. The election sparked controversy, with consequences including a series of violent outbursts leading to several deaths (Kokutse, 2020). Mr. Mahama petitioned Ghana's supreme court to overturn the election results, alleging improper conduct by the election commission and ballot box stuffing on behalf of the NPP (Mahama v. Electoral Commission & Akufo-Addo, 2020). The petition was dismissed due to insufficient evidence, although witness testimony supported the allegation of ballot box stuffing (Mahama v. Electoral Commission & Akufo-Addo, 2021). In this chapter, we analyze voting results for statistical evidence of ballot box stuffing. Such analysis holds both historical value and future relevance, particularly in informing the strategic placement of election observers in future elections. The rest of the chapter is organized as follows. Section 2.1 reviews relevant background. Section 2.2 outlines a method for identifying constituencies with voting results suggestive of ballot box stuffing. Section 2.3 takes a closer look at results for the Ashanti region, inferring the occurrence of ballot box stuffing at individual polling places, and estimating the overall impact of that presumptive voter fraud.

2.1 Background

Administratively, Ghana is comprised of 16 regions, which are subdivided into 275 parliamentary constituencies[1]. Available data consists of the number of registered voters and total votes for

[1]There were 275 constituencies at the time of the 2020 national election. One additional constituency was subsequently added between the 2020 and 2024 elections.

each presidential candidate by polling place. The data set does not have 100% coverage. The available data comes from 274 constituencies (all except Techiman North in the Bono East region) and covers approximately 37,000 polling places, which account for 96% of the total votes cast in the 2020 presidential election. The nationwide aggregate number of votes for each candidate is also available. Different forms of fraud affect this type of voting data in different ways. For instance, falsification of the reported vote totals at individual polling places may alter the frequency of certain digits in the resulting vote totals. Beber and Scacco (2012) demonstrated the disproportionate number of polling places with vote totals ending in '0' in certain elections suspected of fraud, and Mebane (2008) highlighted a similar result for the second digit of vote totals. In a similar vein, Kobak et al. (2016) found an overabundance of whole integer percentages for the turnout (number of votes cast as a percentage of the number of registered voters) and share of votes for the winner across several Russian elections. Discarding or invalidating ballots cast for opposition candidates decreases turnout while increasing the share of votes for the party committing the fraud. Ballot box stuffing is a similar form of election fraud, in which additional fraudulent ballots are cast, replacing true abstentions with manufactured votes for a particular candidate. Ballot box stuffing increases both the turnout and the share of votes for the party committing fraud, inducing a positive correlation between these two quantities. This observation forms the basis of the analysis in Sections 2.2 and 2.3. **Note**: To simplify exposition, we will hereafter use "the incumbent" as shorthand for "the candidate to whose benefit the fraud is committed". The incumbent party was the target of fraud allegations in Ghana's 2020 presidential election, although this is not universally the case.

One of the central challenges in quantitative approaches to detecting election fraud is the absence of labeled data, particularly for fraudulent cases. Even for elections widely suspected of fraud, we generally do not know the mechanisms, the extent, or the specific locations where fraud occurred. Thus, these quantitative approaches necessarily rely on assumptions about the characteristics of fair and fraudulent elections. Zhang et al. (2019) approach this issue by using synthetic data to train a classifier to detect instances of fraud. Synthetic "fair" results for each polling place are generated by formulating the vote share of each candidate as a generalized linear model

based on local demographic factors. Vote totals at a subset of polling places are then manipulated, creating known "fraudulent" locations. How well the resulting classifier generalizes to real elections depends in part on how closely the synthetic "fair" vote totals resemble the results of actual elections free of fraud. Montgomery et al. (2015) also propose a supervised method for training a fraud detection model, but rather than utilizing synthetic data, the authors manually assign a "fraud score" to a set of several hundred elections around the world. The fraud scores are based on details such as denial of access to election monitors, protests or riots, and documented concerns of unfair election practices. The fraud scores themselves are then treated as the dependent variable to predict, using a collection of quantitative digit-based indicators and contextual factors such as regime type and socioeconomic inequality.

The digit-based approaches of Mebane (2008) and Beber and Scacco (2012) rely on assumptions about the distribution of digits in the vote totals of fair elections; the assumptions of Mebane (2008) in particular have faced criticism (Deckert et al., 2011). Klimek et al. (2012) assumes a parametric model for the marginal distributions of turnout and incumbent share in a fair election, and posits that deviations from the model are signals of potential fraud.

Another line of work, to which this present chapter belongs, supposes that we would not generally expect variation in voter turnout to be correlated with the level of support for a particular candidate in a fair election. This relationship was first noted by Sobyanin (1994) when studying the results of a 1993 constitutional referendum in Russia. Myagkov et al. (2009) further discusses this relationship, emphasizing that we expect those quantities to be uncorrelated in homogeneous voting populations. Klimek et al. (2012) studies turnout and incumbent share separately from one another, but also illustrates the presence of correlation in several elections suspected of fraud (Uganda 2011, Russia 2011, Russia 2012) as well as its absence in several elections generally considered fair (e.g. France 2007, Switzerland 2007, Canada 2011). A similar observation appears in Kobak et al. (2016). After identifying potential cases of fraud based on the frequency of exact integer percentages in local vote tallies in seven Russian elections, Kobak et al. (2016) remark that the districts suspected of fraud display moderate to strong correlation (ranging from approximately

0.6 to 0.8) between turnout and incumbent share, whereas correlation is materially lower among the remaining districts.

2.2 Methodology

Because ballot box stuffing was the primary form of fraud suspected in Ghana's 2020 presidential election, we rely on the intuition from Section 2.1 that ballot box stuffing tends to produce positive correlation between turnout and incumbent vote share across polling places. With this in mind, we aim to identify constituencies in which the per-polling-place turnout and incumbent vote share display a stronger positive correlation than could be explained by random chance, if indeed the turnout and incumbent vote share were independent.

For a particular constituency, define the following quantities.

n	number of polling places in constituency
R_i	number of registered voters at polling place i
V_i	number of votes cast at polling place i
W_i	number of votes cast for incumbent at polling place i
$t_i = \frac{V_i}{R_i}$	turnout proportion at polling place i
$w_i = \frac{W_i}{V_i}$	proportion of votes for incumbent at polling place i

Within the observed data for a particular constituency, suppose that the values of turnout and incumbent share were assigned independently of each other. Then every pairing of the t_i's in $t_{1:n}$ with the w_i's in $w_{1:n}$ would be equally likely. Each of those pairings would result in a different value of the correlation between t and w. Considering those values, a relevant question is whether the actual correlation in the observed data is implausibly large with respect to the distribution of correlations that would arise under independence. If so, we would be disinclined to believe that the turnout and incumbent share are independent within the constituency in question. Because the total number of possible pairings is prohibitively large, we approximate by randomly generating a smaller number of permutations.

As outlined in Algorithm 3, we modify the described randomization test with an additional layer of hierarchy. Polling place i gets assigned a random turnout π_i^v from the observed values $t_{1:n}$, and incumbent share π_i^w from $w_{1:n}$. Then, then new number of votes V_i^* at polling place i is

drawn from a binomial distribution parameterized by π_i^v and the number of registered voters R_i. Consequently, the new turnout value recorded for polling place i is $\frac{V_i^*}{R_i}$, rather than π_i^v itself. A new value for incumbent share is determined analogously. This allows for additional variability in the randomization distribution, in a fashion that incorporates the differing sizes of the polling places. We note that this construction shares some similarities to Kobak et al. (2016)'s approach to bootstrap resampling of election results.

Algorithm 3: Approximating the Randomization Distribution of Correlations

Input:

- Election data from polling places in an individual constituency $\{R_i, t_i, w_i\}_{i=1}^n$
- Desired number of samples B

Output: Samples of the correlation ρ between turnout and vote share for incumbent $\{\rho_b\}_{b=1}^B$;

for b *in* $1 : B$ **do**
  for i *in* $1 : n$ **do**
    Draw π_i^v uniformly from $\{t_i\}_{i=1}^n$ without replacement ;
    Draw π_i^w uniformly from $\{w_i\}_{i=1}^n$ without replacement ;
    Sample $V_i^* \sim \text{Binom}(N_i, \pi_i^v)$;
    Set $t_i^* \leftarrow \frac{V_i^*}{N_i}$;
    Sample $W_i^* \sim \text{Binom}(V_i, \pi_i^w)$;
    Set $w_i^* \leftarrow \frac{W_i^*}{V_i}$;
  end
  Set $\rho_b \leftarrow \text{SpearmanCorrelation}(\{t_i^*\}_{i=1}^n, \{w_i^*\}_{i=1}^n)$;
end
return $\{\rho_b\}_{b=1}^B$

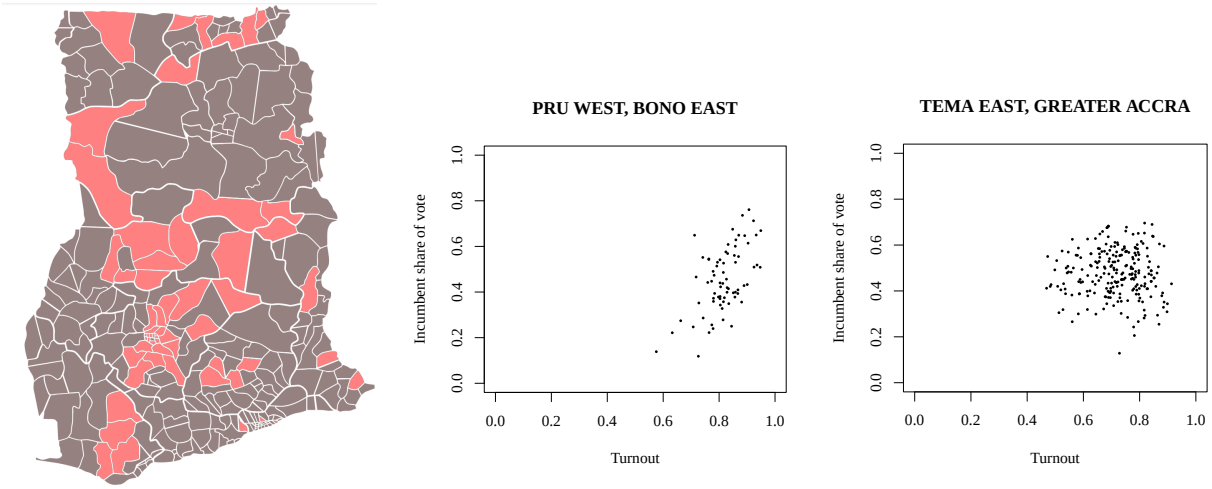
With B-many samples of correlation values, compare to the observed value ρ_{obs} by calculating

$$\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{\rho_b > \rho_{obs}\}. \tag{2.1}$$

If this fraction exceeds a desired threshold α, the constituency's observed correlation between

turnout and incumbent share is larger than could be reasonably explained by random chance under independence, and is consistent with what would be expected in the presence of ballot box stuffing. Note that the correlation measure employed here is the Spearman (rank) correlation, because of both its robustness to outliers, and its ability to capture monotone associations that are not necessarily linear, which are of interest in this context.

After performing the above exercise for each of the 274 constituencies in the data set, drawing 100,000 samples each time, a Benjamini Hochberg correction (Benjamini and Hochberg, 1995) to the values calculated in Equation 2.1 is appropriate to control the false discovery rate, as this is a multiple testing problem. At a significance level of 0.01, this procedure flags 61 constituencies as having significant positive correlation between turnout and vote share for incumbent. The flagged constituencies are highlighted on a map of Ghana in Figure 2.1(a) and enumerated in Appendix B. Figure 2.1(b) and (c) respectively show examples of voting results for a constituency that was flagged, and for one that was not.



(a) (b) (c)

Figure 2.1: (a) Map of Ghana. Constituencies highlighted in pink had voting results flagged for significant positive correlation, (b) Example of a constituency whose results were flagged, (c) Example of a constituency that was not flagged

Ballot box stuffing is one possible explanation for the significant correlation, but it is not the only possible explanation. In addition to the usual notions of Type I and Type II statistical er-

ror, there is additional possibility of errors due to the fact that correlation between turnout and incumbent share may exist in some regions for idiosyncratic reasons unrelated to fraud. A positive correlation between turnout and incumbent share does not definitively prove that ballot box stuffing occurred, nor does the absence of a positive correlation guarantee the absence of fraud. Regarding the latter, consider that a particular region may naturally exhibit a negative correlation between turnout and incumbent share. In such a region, ballot box stuffing could theoretically increase the correlation to zero, causing the above method to fail to detect fraud. Myagkov et al. (2009) offers a thorough discussion of the possibility and the pitfalls of natural associations between turnout and incumbent share for reasons other than fraud. We emphasize that these findings should not be treated as conclusive on their own, but rather, may be a valuable screening tool to use in conjunction with expert opinion on the local political and cultural landscape.

2.3 Voting Patterns in the Ashanti Region

A disproportionately large number of the flagged constituencies, 26 of 61, are located in the Ashanti region. The region is located in the central-southern portion of the map in Figure 2.1(a); refer to Figure 2.2 for a closer view of the Ashanti region by itself.

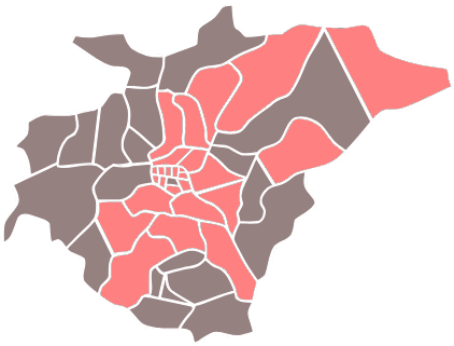


Figure 2.2: Constituencies in the Ashanti region. Constituencies flagged for significant positive correlation between turnout and incumbent share are highlighted in pink.

Ashanti has historically been a stronghold of the incumbent party. Note that unlike the United States, Ghana appoints a winner based on the national popular vote, so there is still incentive to manufacture additional votes for a candidate in a region where they already expect to receive

a majority of the vote. In addition to anecdotal reports of election interference in the region, independent election observers witnessed a voter attempt to fraudulently submit multiple ballots in Ashanti's Asawase constituency (Coalition of Domestic Election Observers, 2023). Given the results from Section 2.2 and the cultural context, the voting results from Ashanti warrant further consideration. Examining the election data from polling places in Ashanti, there is a pronounced difference between the data from "flagged" constituencies and from non-flagged ones, as illustrated in Figure 2.3.

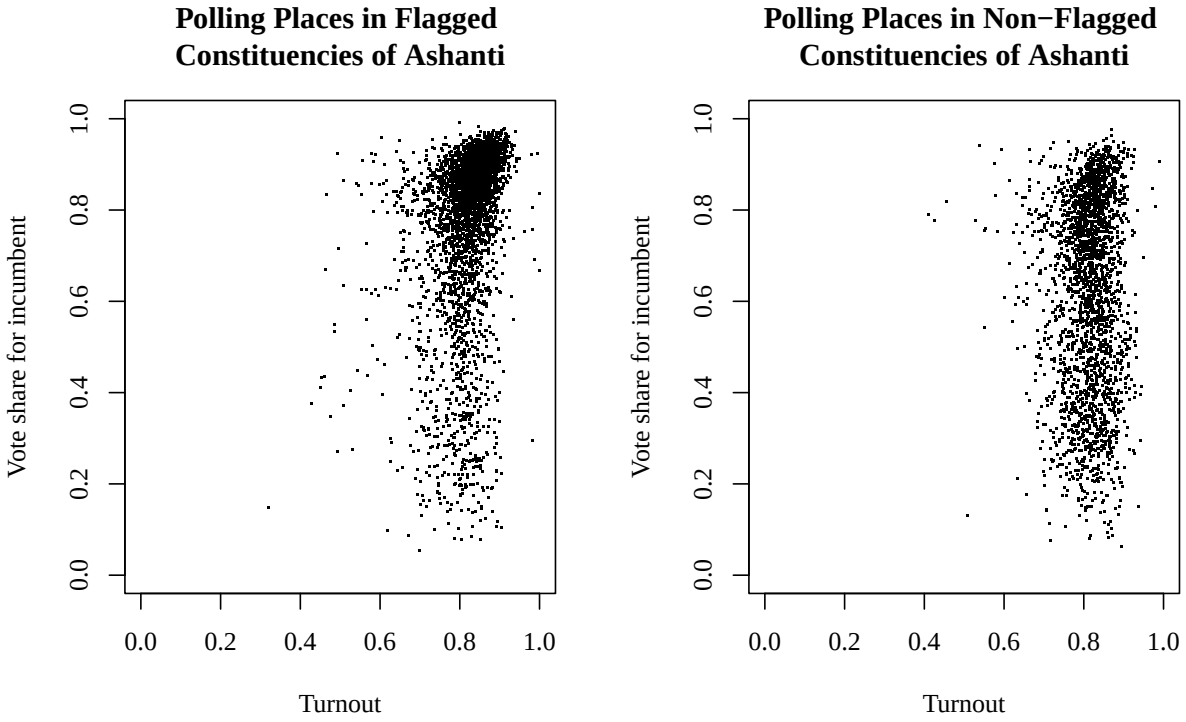


Figure 2.3: Voter turnout and vote share for incumbent candidate per polling station are distributed differently in the flagged constituencies.

2.3.1 Polling Place-Level Inferences

Within the flagged constituencies, there is a higher concentration of polling places with simultaneously high turnout and a high proportion of votes for the incumbent. The plot of polling places in flagged constituencies appears to consist of a tightly clustered group of polling places with high turnout and proportion of votes for the incumbent, and another less tightly-clustered

group of polling places with more diffuse values of vote share for incumbent, and a slightly lower average turnout. The cluster with simultaneously high turnout and high proportion of votes for the incumbent displays voting patterns we would expect in the presence of ballot box stuffing. Motivated by that observation, consider modeling the data from flagged constituencies as belonging to two clusters, estimating the properties of those clusters and the probability of each polling place belonging to each cluster. Assuming each cluster can be approximated as arising from a bivariate Gaussian distribution, a two-component Gaussian mixture model can provide estimates of the desired quantities.

Notationally, let "cluster 1" refer to the more spread-out cluster, and "cluster 2" refer to the denser cluster with high turnout and high incumbent share. We denote the total number of polling places in the flagged constituencies in Ashanti as N, and introduce an additional variable z_i for $i = 1, \ldots, N$, an indicator for the cluster to which polling place i belongs (i.e. $z_i = 1$ or $z_i = 2$). As outlined below, this model supposes that the cluster assignment is drawn from a Bernoulli distribution. Given the cluster assignment indicator z_i, the (turnout, incumbent share) pair for data point i is drawn from the Gaussian distribution for its respective cluster.

$$
\begin{aligned}
(t_i, w_i)|\mu, \Sigma, z_i &\sim \mathcal{N}(\mu_{[z_i]}, \Sigma_{[z_i]}) \qquad \text{for } i = 1, \ldots, N \\
(\mu_1, \Sigma_1) &\sim \text{Normal-inverseWishart}(\eta_1, \lambda = 1, \Psi, \nu = 4) \\
(\mu_2, \Sigma_2) &\sim \text{Normal-inverseWishart}(\eta_2, \lambda = 1, \Psi, \nu = 4) \\
\eta_1 &= [0.8,\ 0.65] \\
\eta_2 &= [0.8,\ 0.75] \\
\Psi &= \begin{bmatrix} 0.01 & 0 \\ 0 & 0.05 \end{bmatrix} \\
z_i|\theta &\sim \text{Bernoulli}(\theta) \\
\theta &\sim \text{Unif}(0, 1)
\end{aligned}
\tag{2.2}
$$

The posterior distribution of the cluster assignments $z_{1:N}$ and the per-cluster distribution parameters given the observed data $t_{1:N}$ and $w_{1:N}$ offer insight into the voting patterns and possibility of ballot box stuffing. The particular choice of prior distributions on θ, (μ_1, Σ_1) and (μ_2, Σ_2) in model 2.2 result in conditional conjugacy. Due to this property, Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990) provides a straightforward way to draw approximate samples from the posterior distribution. The posterior means for (μ_1, Σ_1) and (μ_2, Σ_2) indicate little correlation between turnout and incumbent share in cluster 1, compared to cluster 2 which exhibits relatively strong positive correlation (≈ 0.40), as well as typically higher turnout and incumbent share. Figure 2.4 displays quantiles of Gaussian distributions with parameters given by the posterior means of (μ_1, Σ_1) and (μ_2, Σ_2).

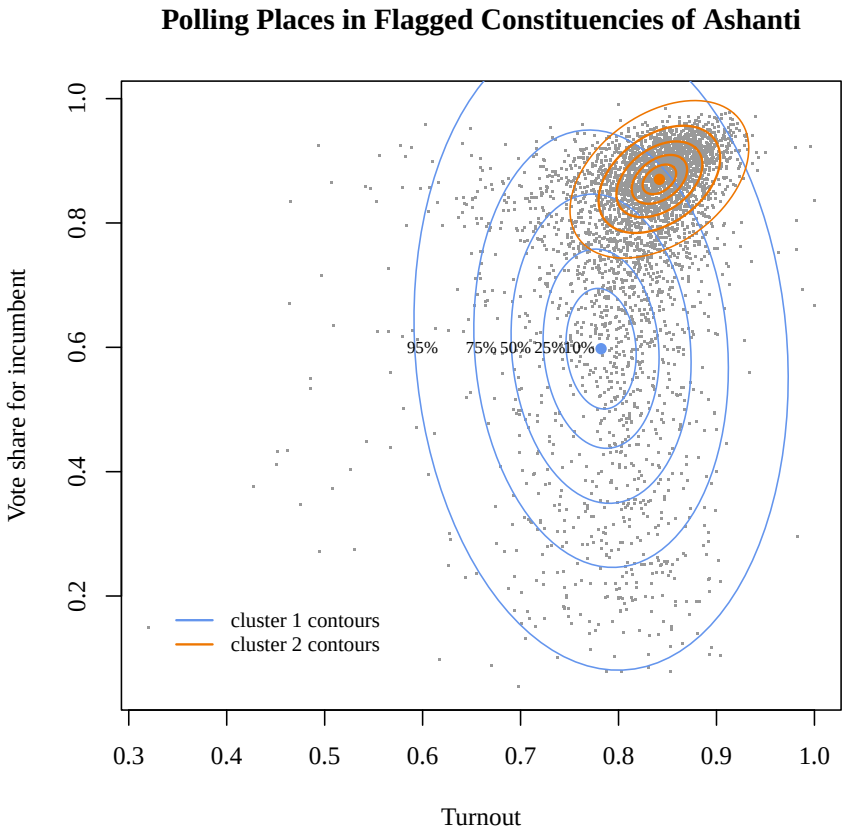


Figure 2.4: 10%, 25%, 50%, 75%, and 95% quantiles of the estimated distribution for each cluster

Because the voting pattern in cluster 2 is consistent with what we would expect in the event

of ballot box stuffing, we are interested in inferring which polling places belong to cluster 2. Figure 2.5 highlights each polling place's dot on the scatter plot according to its posterior probability of belonging to cluster 2, i.e. $p(z_i = 2|t_{1:N}, w_{1:N})$. The cluster membership probabilities also confirm that the two clusters do not merely partition the set of 26 constituencies; all constituencies have multiple polling places likely to belong to each of the two clusters.

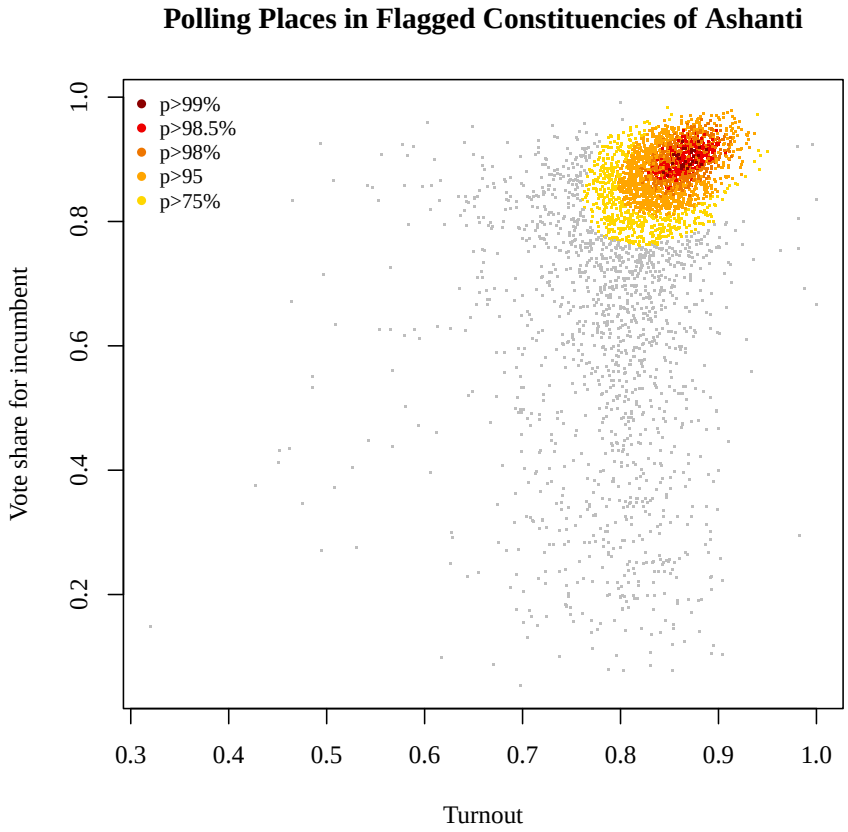


Figure 2.5: Colors correspond to the estimated posterior probability that a polling place belongs to cluster 2

As with the analysis in Section 2.2, this should not be construed as concrete proof of fraud at any particular polling station. These inferences may be useful, however, as one component of a broader strategy for detecting fraud and protecting the integrity of future elections.

2.3.2 Counterfactual Election Results

Having identified a pattern of suspicious voting results, we can conjecture what the election results would have looked like, had there been no interference in the Ashanti region (if, indeed, ballot box stuffing was responsible for the irregularities). In particular, it is interesting to consider minimally disruptive ways to modify the observed election results to be consistent with our notion of a fair election.

If both clusters are modeled as bivariate Gaussians, results from the field of optimal transport (Peyré and Cuturi, 2019) demonstrate that the most efficient mapping from $N(\mu_2, \Sigma_2)$ to $N(\mu_1, \Sigma_1)$ is given by mapping $x \sim N(\mu_2, \Sigma_2)$ as follows:

$$x \mapsto \mu_1 + A(x - \mu_2), \tag{2.3}$$

where

$$A = \Sigma_2^{-1/2}(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})\Sigma_2^{-1/2}.$$

This mapping is optimal in the sense that it minimizes the Euclidean distance over which each infinitesimal unit of probability mass is relocated. More formally, that quantity is the 2-Wasserstein distance, with respect to the Euclidean distance on $\mathbb{R}^2$. Figure 2.6 illustrates the results of this mapping, for one individual sample from the posterior distribution $p(z_{1:N}, \mu_1, \Sigma_1, \mu_2, \Sigma_2 | t_{1:N}, w_{1:N})$.

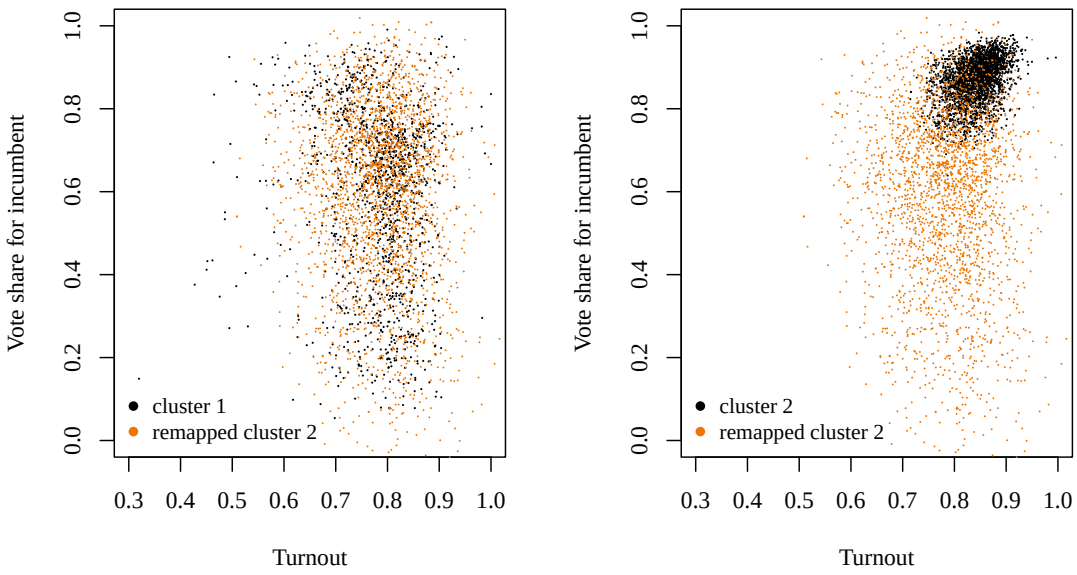


Figure 2.6: This mapping redistributes the data points from cluster 2 to match the approximate bivariate Gaussian distribution corresponding to cluster 1.

This mapping results in new values of turnout and incumbent share for each of the polling places belonging to the suspicious cluster (i.e. with $z_i = 2$ in this posterior sample). Because the number of registered voters for each polling place is fixed and known, this implies a new total number of votes for the incumbent and for the opposition at each polling place. Using those recalculated vote totals for polling places in cluster 2 allows for a tabulation of the aggregate vote totals across flagged constituencies in the Ashanti region, under the described modification. These vote totals are displayed in Table 2.1.

	observed vote total	modified vote total	change
incumbent	1,116,339	970,439	-145,900
other candidates	301,497	376,474	+74,977
difference	814,842	593,965	-220,877

Table 2.1: Actual and hypothetical aggregated vote totals for flagged constituencies in the Ashanti region

The optimal mapping of cluster 2 onto cluster 1 results in the incumbent candidate's margin over all other candidates shrinking by approximately 220,000 votes. In the actual election, the incumbent candidate received approximately 517,000 more votes than the second-place candidate, so the changes in the Ashanti region in this hypothetical scenario are not large enough to change the outcome of the national popular vote, although they are of a similar order of magnitude. Ghana's presidential election procedures stipulate that if no candidate wins a simple majority of votes cast, a runoff vote is held. In the hypothetical voting results above, the incumbent candidate only exceeds the 50% threshold by approximately 60,000 votes out of more than 13 million total votes cast. Accounting for the possibility of ballot box stuffing in some of the other flagged constituencies outside of the Ashanti region, this could plausibly lead to a runoff vote in the outlined hypothetical scenario.

However, the described method of constructing alternative "fair" voting results is mismatched with our notion of a "minimally disruptive" modification in two key ways. First, it does not preserve the total number of votes for the opposition. Experts believe that ballot box stuffing was the relevant form of interference in this particular election, and ballot box stuffing affects the vote totals by artificially inflating the number of votes for the incumbent candidate. It does not change the number of votes for opposing candidates. As shown in Table 2.1, the number of votes for opposition candidates differs materially between the observed vote totals and the hypothetical totals under the optimal mapping.

Second, it does not reflect the different number of voters at each polling place. The question we would most like to answer is "What is the minimum number of votes for the incumbent candidate we would need to delete in order to make these election results look plausibly fair?". Moving a polling place from (x_1, y_1) to (x_2, y_2) on the scatter plot of turnout vs. incumbent share is more "disruptive" in the latter sense if the polling place has a larger number of registered voters, but the mapping in Equation 2.3 does not take this into account.

Instead, we can imagine decreasing the number of votes for the incumbent at each polling place, while leaving the number of votes for opposition parties unchanged. If the number of votes

for the incumbent deleted from polling place i is δ_i, the new value of turnout is $\frac{V_i - \delta_i}{R_i}$ and the new value of incumbent share is $\frac{W_i - \delta_i}{V_i - \delta_i}$. This restricts the set of possible alternate outcomes for each polling place. On the scatter plot of turnout vs. incumbent share, the point corresponding to a given polling place can be moved along a particular curve, with locations along that curve corresponding to different numbers of deleted votes for the incumbent. Figure 2.7 illustrates three examples.

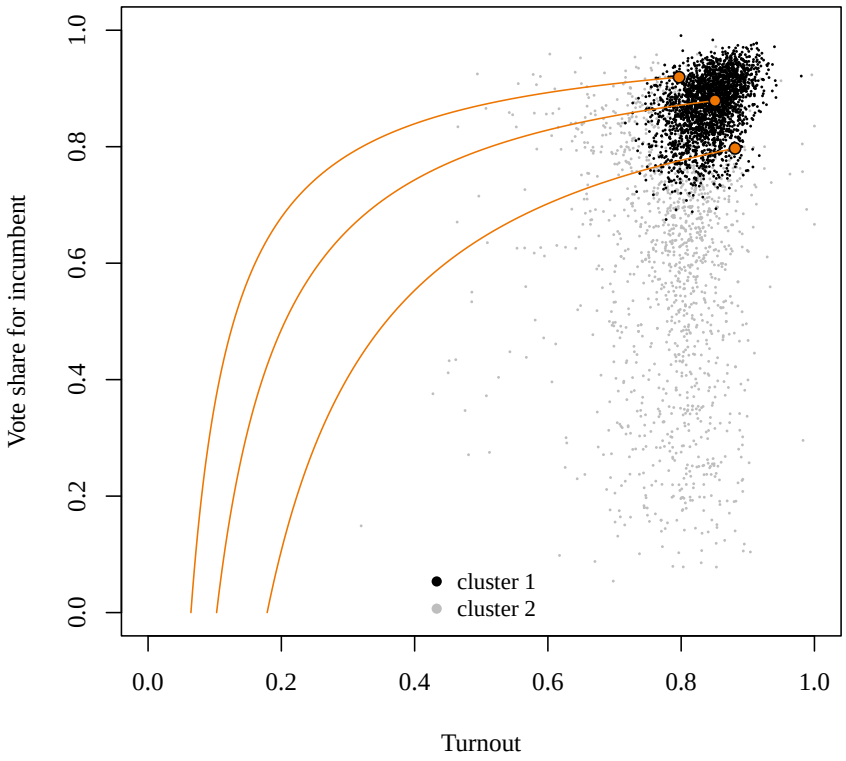


Figure 2.7: Example trajectories created by deleting votes for the incumbent at an individual polling place. As the number of deleted votes increases, the polling place's dot on the scatter plot moves from right to left along the orange curve.

Note that under this approach to modifying the vote numbers, it is no longer possible to re-arrange the points from cluster 2 to follow the same distribution as the points from cluster 1. Mapping cluster 2 onto cluster 1 in the previous approach, we implicitly assumed that in a counterfactual scenario absent of fraud, the voting characteristics of the potentially fraudulent polling

places (i.e. those with $z_i = 2$ when conditioning on the observed data) would be the same as the voting characteristics of the other polling places. It is not obvious a priori whether we should expect the occurrence of ballot box stuffing at a polling place to be independent of the natural voting characteristics of that polling place, so this new constraint is not necessarily problematic.

For this new approach to constructing a hypothetical fair election, our objective is formalized in Equation 2.4 below.

$$\begin{aligned} &\min \sum_{i:z_i=2} \delta_i \\ &\text{s.t.} \quad 0 \leq \delta_i \leq W_i \quad \forall i : z_i = 2 \\ &\qquad \text{corr}(\tilde{t}_{1:N}, \tilde{w}_{1:N}) \leq 0 \end{aligned} \tag{2.4}$$

where

$$\tilde{t}_i = \begin{cases} \frac{V_i}{R_i} & if \quad z_i = 1 \\ \frac{V_i - \delta_i}{R_i} & if \quad z_i = 2 \end{cases}$$

$$\tilde{w}_i = \begin{cases} \frac{W_i}{V_i} & if \quad z_i = 1 \\ \frac{W_i - \delta_i}{V_i - \delta_i} & if \quad z_i = 2 \end{cases}$$

Given cluster assignments $z_{1:N}$, the objective in Equation 2.4 seeks the smallest total number of votes for the incumbent that can be deleted at polling places belonging to cluster 2, such that among all polling places in the flagged constituencies in Ashanti, the correlation between turnout and incumbent share is not positive.[2] Requiring the correlation to be equal to zero would be most consistent with the conceptual ideal of a fair election described in Section 2.1, but empirically, optimization routines with an inequality constraint were more feasible than those with strict equality constraints. Because the stated constraint on correlation opposes the goal of minimizing the number of votes deleted, in practice the optimum occurs effectively at the boundary where correlation

[2]The correlation in Equation 2.4 is the Pearson (product-moment) correlation rather than the Spearman (rank) correlation used in Algorithm 3. Because the Pearson correlation is differentiable with respect to δ, this change greatly simplifies the optimization routine. Empirically, the Spearman correlation was also close to zero at the optimum (between 0 and 0.03, approximately, compared to 0.433 in the observed data).

equals zero.

This optimization applies to a specific set of cluster assignments $z_{1:N}$. Thus, the posterior distribution $p(z_{1:N}|t_{1:N}, w_{1:N})$ induces a corresponding distribution over the minimum number of incumbent votes to delete. Figure 2.8 shows an approximation of that distribution, generated by performing the minimization in Equation 2.4 at each of 2,000 samples from the (approximate) posterior $p(z_{1:N}|t_{1:N}, w_{1:N})$.

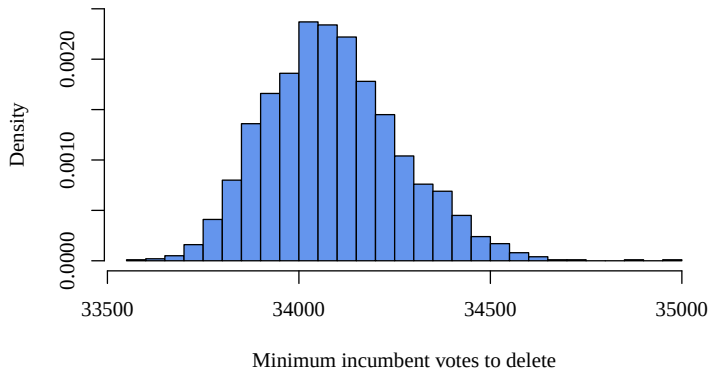


Figure 2.8: Discarding approximately 34,000 votes for the incumbent could be sufficient to produce voting results consistent with expectations for a "fair" election.

The minimum number of votes to deduct from the incumbent candidate is much smaller in this hypothetical scenario than in the previous exercise mapping cluster 2 onto cluster 1 via Equation 2.3. The central 95% credible interval is approximately 33,800 to 34,450 votes, compared to over 200,000. So, if all of the observed correlation between turnout and incumbent share in certain parts of Ashanti was due to ballot box stuffing, the extent of that fraud could potentially be as small as roughly 34,000 fraudulent ballots cast for the incumbent. Whether this level of fraud is tolerable in itself is a separate question, but in any case, it is inconsequential relative to the incumbent's 500,000 vote margin of victory.

2.4 Discussion

Ghana's 2020 presidential election drew allegations of fraud, namely in the form of ballot box stuffing. As discussed in the election forensics literature, ballot box stuffing affects local vote tallies by producing a positive correlation between turnout rates and proportion of the votes won by the party perpetrating the fraud. Several constituencies in Ghana had voting results that exhibited this pattern to a significant degree. The Ashanti region in particular had a high concentration of polling places with simultaneously high turnout and vote share for the incumbent. However, even if ballot box stuffing were the sole cause of this pattern, it could potentially be explained by an amount of fraud far too small to have materially impacted the outcome of the election. The findings herein are not intended to give conclusive proof of malfeasance, but rather, may be a valuable indicator to use in conjunction with expert opinion on the local political and cultural landscape.

Chapter 3: Leveraging Shared Information Across Multiple Synthetic Control Studies

Synthetic control methods, introduced by Abadie et al. (2010), provide a strategy for estimating causal effects from observational panel data. In a conventional use case, a collection of units would be observed over time, both before and after a treatment was applied to a particular "target" unit. Using data from the pre-treatment period, researchers learn to model the outcome values of the target unit as a weighted combination of the other, "donor", units' outcomes. In the post-treatment period, the donor units (who remain untreated) can be used to impute what the target unit's outcome would have been, had the target not been treated. Comparing the treatment unit's observed outcome under treatment to its estimated untreated outcome provides an estimate of the individual treatment effect on that unit.

In the following, we consider a slightly different scenario, in which a collection of units are observed over time, alternating between periods in which all units are untreated, and periods in which a subset of the units receive treatment, as illustrated in Figure 3.1. Such examples may arise when similar policy interventions are implemented in different regions at different times. For instance, Dube and Zipperer (2015) study the effect of US state minimum wage increases on teen employment. Each minimum wage increase constitutes its own study, and each study has a different set of donor units, depending on which other states do not also implement minimum wage increases during the respective time period. Our motivating example, detailed in Section 3.4, concerns the effect of lottery winnings on support for incumbent politicians. The lottery in question is a large national lottery held annually in Spain, and the units are the provinces of Spain. The "treatment" is a grand prize win, and each year grand prizes are won in different provinces (there may be multiple winners), changing the set of provinces that can serve as donor units each year.

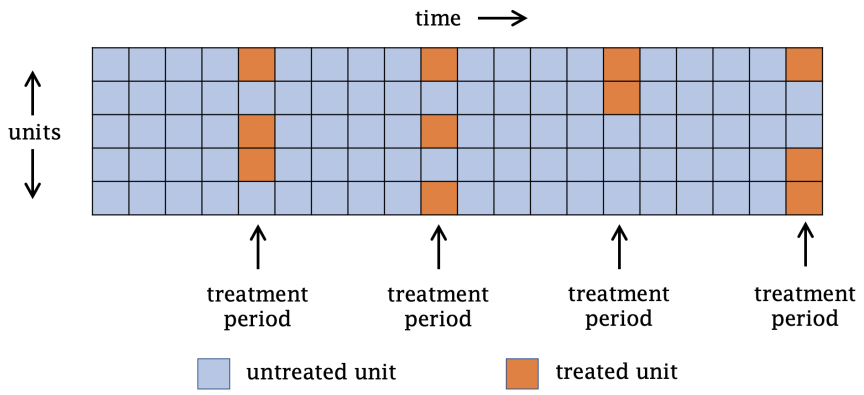


Figure 3.1: In the envisioned data setting, each treatment period can be considered to demarcate a separate study.

For our particular target unit of interest, we could view this as a series of sequential synthetic control studies. This scenario presents a challenge for the conventional synthetic control method, because the set of available donor units is not the same in each individual study. Other units may be ineligible to serve as donors in some of the studies, either because they themselves were also receiving the treatment at that time, or simply because their data from that time period is not available. One option is to fit a single synthetic control estimator, using only the donor units that are available in all time periods. However, if the number of such donors is very small, this may be quite limiting, making it difficult to construct a close approximation to the target unit. Another option is to fit separate synthetic control estimators to each study, independent of each other, using the pool of donors available in each individual study. If the number of time periods in each individual study is much smaller than the number of donors, this approach is vulnerable to overfitting, and can produce estimators that generalize poorly to new time periods. By fitting synthetic control estimators for each study individually, this approach also ignores any shared structure among the related estimation problems. In this chapter, we propose a hierarchical Bayesian approach to fitting synthetic control estimators, which takes advantage of shared information across multiple related synthetic control studies. The chapter is arranged as follows. Section 3.1 discusses relevant literature on synthetic control methods. Section 3.2 outlines the proposed estimator. Section 3.3 explores performance on synthetic data, and Section 3.4 applies the proposed methodology to real

data concerning the Spanish Christmas lottery. Section 3.5 concludes with a discussion of findings and directions for future work.

3.1 Related Work

There is a vast literature on synthetic control methods, pioneered by Abadie et al. (2010). Within that broader body of work, various Bayesian approaches have emerged in recent years. Both Amjad et al. (2018) and Kim et al. (2020) utilize Bayesian linear regression models to estimate synthetic control weights for a single study. These formulations remove the conventional constraint that the weights be non-negative and sum to one, whereas the method proposed in Section 3.2.3 maintains that constraint. In Kim et al. (2020) the motivation for taking a Bayesian approach is twofold: prior distributions on the synthetic control weights provide regularization, and and the Bayesian framework offers straightforward inference for model quantities. Amjad et al. (2018) also note the benefit of uncertainty quantification, although it is important to be aware that the posterior distribution on the weights quantifies uncertainty arising from the linear regression model, but does not address uncertainty arising from a preprocessing step in which singular value thresholding is performed on the matrix of untreated outcome data. In this chapter, the proposed hierarchical Bayesian estimation method is also motivated by the desire to partially pool information across multiple related studies. This aspect is not relevant to Amjad et al. (2018) and Kim et al. (2020), which focus on the case of a single panel study.

Brodersen et al. (2015) and Pang et al. (2022) also employ Bayesian methods for synthetic control estimation. The methods in these works are designed for data-rich applications with many pre-treamtent time periods. Both approaches explicitly incorporate time series structure, modeling the evolution of the units' outcomes over time. Such methods are well-suited to applications in topics like finance and online marketing, where high frequency data is available. However, they have limited applicability in settings with a smaller number of time periods, as is often the case when studying large-scale interventions in the social sciences. The method proposed in Section 3.2.3 is intended to address applications of the latter variety.

The specific case of multiple related synthetic control studies has received less attention. Dube and Zipperer (2015) explore the effect of US state minimum wage increases on teen employment, and each minimum wage increase constitutes its own synthetic control study. Each time a state increases its minimum wage, the treatment effect is estimated by fitting an individual synthetic control estimator using the states available to serve as donor units in a time window around the increase. For the purposes of fitting synthetic control estimators, no information is shared across the individual studies. In contrast, pooling information across estimation tasks is a central goal of the present chapter.

Athey et al. (2021) draws connections between matrix completion and several popular methods for estimating causal effects from (observational) panel data, including synthetic control. The authors reframe the goal of these methods as a matrix completion problem, in which we have observed the untreated outcomes for some units at some time points, and wish to impute the hypothetical untreated outcomes that we did not get to observe (because those units did receive the treatment). This approach is applicable to the case of multiple related synthetic control studies, if we imagine concatenating the outcome data matrices for each individual study into one large partially observed matrix of untreated outcomes. This approach shares a fundamental similarity with the method proposed in Section 3.2.3, in that all of the observed untreated outcomes from all of the studies are used to impute each missing untreated outcome.

3.2 Methodology

For purposes of exposition, consider the case of a single intervention applied to one target unit. Suppose we observe an outcome variable Y for units $\{1, \ldots, J\}$ at time periods $\{1, \ldots, T\}$. Without loss of generality, let unit 1 be the target unit. An intervention is applied to unit 1 beginning at time $t = T_0 + 1$. Let Y_{jt} denote the observed outcome for unit j at time t, and $Y_{jt}(0)$ and $Y_{jt}(1)$ denote the potential outcomes for unit j at time t under control and under treatment, respectively. We make the consistency assumption that at any given time, a unit's observed outcome matches its potential outcome for the condition to which it is assigned (i.e. either treatment or control). Note that

methods for estimating the target unit's untreated potential outcome during treated time periods are motivated by the assumption of a latent factor model for the untreated potential outcomes. In such a model, the untreated potential outcomes are expressed as

$$Y_{jt}(0) = \lambda_t^\top \mu_j + \varepsilon_{jt},$$

where $\lambda_t \in \mathbb{R}^r$ is a vector of unobserved common factors at time t, $\mu_j \in \mathbb{R}^r$ is a vector of unobserved factor loadings for unit j, and ε_{jt} are mean-zero independent noise terms.

3.2.1 Conventional Synthetic Control

Conventional synthetic control methods model the target unit's pre-intervention outcomes as a linear combination of the other donor units' outcomes. In a generic implementation of this method, the weights of the donor units in the linear combination may be chosen to minimize squared error, subject to the constraint that the weights are non-negative and sum to one. Let Y_{-jt} denote the vector of observed outcomes at time t for all units except unit j. That is, $Y_{-jt} = (Y_{1t}, \ldots, Y_{(j-1)t}, Y_{(j+1)t}, \ldots, Y_{Jt})$. Let Δ^k denote the k-simplex, consisting of vectors of length $k+1$ whose entries are non-negative real numbers that sum to one. The weights used to express unit 1 as a linear combination of the other $J-1$ units are given by

$$\boldsymbol{\beta}^{SC} \in \underset{\boldsymbol{\beta} \in \Delta^{J-2}}{\operatorname{argmin}} \sum_{t=1}^{T_0} (Y_{1t} - Y_{-1t}^\top \boldsymbol{\beta})^2. \tag{3.1}$$

The counterfactual control outcome for unit 1 at time $t > T_0$ is taken to be

$$\hat{Y}_{1t}^{SC}(0) = Y_{-1t}^\top \beta^{SC}.$$

This leads to an estimated individual treatment effect for unit 1 at time $t > T_0$ given by

$$\begin{aligned}\hat{\tau}_{1t} &= Y_{1t}(1) - \hat{Y}_{1t}^{SC}(0) \\ &= Y_{1t} - \hat{Y}_{1t}^{SC}(0)\end{aligned}$$

3.2.2 Matrix Completion

As discussed in Section 3.1, matrix completion methods are a relevant comparison. This line of work attempts to complete the partially observed matrix of control outcomes, $\mathbf{Y}(0)$. If unit j was receiving treatment at time t, or was not observed, its corresponding entry $Y_{jt}(0)$ is considered to be missing. The goal for causal inference is to impute the missing entries of the matrix. In Athey et al. (2021), this is achieved by finding a low-rank matrix $\hat{\mathbf{Y}}(0)$ that closely approximates $\mathbf{Y}(0)$ at the observed entries. Specifically, Athey et al. (2021)'s Matrix-Completion with Nuclear Norm Minimization minimizes the following objective:

$$\hat{\mathbf{Y}}^{MC}(0) = \underset{M \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{|O|} \sum_{(i,j) \in O} (Y_{jt} - M_{jt})^2 + \lambda \|M\|_* \tag{3.2}$$

In the above, O denotes the set of observed untreated outcomes, and $\|\cdot\|_*$ is the nuclear norm, i.e. the sum of the singular values of matrix M.

3.2.3 Hierarchical Bayesian Synthetic Control

Next, imagine that a subset of the $J - 1$ donor units in this particular study are unavailable, either due to receiving the treatment, or due to data missingness. Denote the set of available donor units in this study as $\mathcal{A}$, and $Y_{\mathcal{A}t}(0)$ the untreated potential outcomes, at time t, for the units in $\mathcal{A}$.

At each time point t, we model the target unit's untreated potential outcomes as

$$
\begin{aligned}
Y_{1t}(0)|\beta^{HB} &\sim N(Y_{\mathcal{A},t}(0)^\top \beta^{HB}, \sigma^2) \\
\beta^{HB}|\boldsymbol{\mu}, \boldsymbol{\varepsilon} &= \text{softmax}(\boldsymbol{\mu}_{\mathcal{A}} + \boldsymbol{\varepsilon}) \qquad (3.3) \\
\mu_j &\overset{iid}{\sim} N(0, \phi^2) \qquad j = 1, \ldots, J-1 \qquad (3.4) \\
\varepsilon_i &\overset{iid}{\sim} N(0, \gamma^2) \qquad i = 1, \ldots, |\mathcal{A}|.
\end{aligned}
$$

As in conventional synthetic control methods, the weight vector β^{HB} consists of non-negative values that sum to one. This weight vector is determined by an underlying vector $\boldsymbol{\mu}$ and random noise terms $\boldsymbol{\varepsilon}$. The vector $\boldsymbol{\mu}$ has length $J-1$, with one entry corresponding to each of the units other than the treatment unit. The variance ϕ^2 in Equation 3.4 controls the sparsity of the weight vector β^{HB}. Larger values of ϕ^2 correspond to weight vectors with many weights close to zero, and only a few larger ones. In contrast, the conventional synthetic control estimator from Equation 3.1 can be reframed as a MAP (maximum a posteriori) estimate of the weight vector β under a model in which the outcome data arises from a Gaussian distribution, and the prior distribution on β is uniform over the simplex.

In a scenario with one single study, this construction in Equation 3.3 is not meaningful. However, in a scenario with multiple studies, in which each study only a subset of the units serve as controls, the global $\boldsymbol{\mu}$ vector facilitates the sharing of information among the different studies. Suppose we have data from multiple studies, indexed by $p = 1, \ldots, P$, measuring the same J units over time. Within study p, the available donor units are $\mathcal{A}_p$. In each study, the goal is to estimate the counterfactual untreated outcome for the target unit. The target unit is the same unit in each

study. For each study p, we model the target unit's untreated potential outcome at time t as

$$
\begin{aligned}
Y_{1t,p}(0)|\beta_p^{HB} &\sim N(Y_{\mathcal{A}_p,t}(0)^\top \beta_p^{HB}, \sigma^2) \\
\beta_p^{HB}|\boldsymbol{\mu}, \boldsymbol{\varepsilon}_p &= \text{softmax}(\boldsymbol{\mu}_{\mathcal{A}_p} + \boldsymbol{\varepsilon}_p) \\
\mu_j &\overset{iid}{\sim} N(0, \phi^2) && j = 1, \ldots, J-1 \\
\varepsilon_{i,p} &\overset{iid}{\sim} N(0, \gamma^2) && i = 1, \ldots, |\mathcal{A}|.
\end{aligned}
$$

During the pre-treatment time periods, the untreated potential outcomes coincide with the observed outcomes for the target unit and for the donor units in $\mathcal{A}_p$. Conditioning on that data, we can approximate the posterior distribution of β_p^{HB} via Markov Chain Monte Carlo methods. The posterior distribution on β_p^{HB} induces a posterior predictive distribution on the target unit's untreated potential outcomes during treatment periods. For treated time period t in study p, the posterior predictive likelihood for the treated unit's counterfactual untreated outcome can be expressed as

$$
\begin{aligned}
p(Y_{1t,p}(0)|\mathcal{D}) &= \int p(Y_{1t,p}(0)|\beta_p^{HB}, \mathcal{D}) p(\beta_p^{HB}|\mathcal{D}) d\beta_p^{HB} \\
&= \int \mathcal{N}(Y_{1t,p}(0)|Y_{\mathcal{A}_p,t}(0)^\top \beta_p^{HB}, \sigma^2) p(\beta_p^{HB}|\mathcal{D}) d\beta_p^{HB}
\end{aligned} \tag{3.5}
$$

Where $\mathcal{D}$ denotes all data from the target and available donor units during all untreated time periods from all studies, and $\mathcal{N}(x|\mu, \sigma^2)$ refers to the normal density with mean μ and variance σ^2, evaluated at x. Comparing the predictive distribution for the untreated potential outcome with the observed outcome under treatment then yields a distribution on the target unit's treatment effect in a particular treatment period.

3.3 Simulation Study

To explore the performance of the estimators from Section 3.2 at modeling outcomes under control, we consider a synthetic data set in which no treatments occur. The data set consists of outcome measurements for 16 donor units over 30 time periods. Donor data $\mathbf{Y} \in \mathbb{R}^{16x30}$ was generated

by adding iid Gaussian noise to an underlying lower-rank matrix $\mathbf{Z}$. An artificial target unit was created by multiplying the underlying $\mathbf{Z}$ matrix by weight vector β^{sim}, then adding iid Gaussian noise. The simulated data was then partitioned into five individual "studies", each consisting of six consecutive time periods. Within each individual study, the final time period (of six) is defined to be the treatment period. In each study a subset of the donors will be masked off as "unavailable", and the set of missing donors will differ across the studies. Refer to Appendix C for further details of the simulation study.

We compare the performance of four methods for estimating the target unit's outcomes during the treatment periods: the hierarchical Bayesian estimator proposed in Section 3.2.3 (HB), the matrix completion method from Athey et al. (2021) (MC-NNM), separate conventional synthetic control estimators fit to each study independently (SCM-indiv), and a single conventional synthetic control estimator fit using only the donor units available across all studies (SCM-complete). The hierarchical Bayesian estimator can be evaluated based on the resulting log posterior predictive density at the actual values observed for the target unit in the treatment periods. Taking t to be the treated time period in each of the five studies, the posterior samples of β_p^{HB} facilitate a Monte Carlo estimate of the posterior predictive likelihood as follows:

$$\Lambda^{HB}(\{Y_{1t,p}\}_{p=1}^5) = \log\Big(\int p(\beta_1^{HB},\beta_2^{HB},\beta_3^{HB},\beta_4^{HB},\beta_5^{HB}|\mathcal{D})\Pi_{p=1}^5 p(Y_{1t,p}(0)|\beta_p^{HB},\mathcal{D})d\boldsymbol{\beta}^{HB}\Big) \tag{3.6}$$

$$= \log\Big(\int p(\beta_1^{HB},\beta_2^{HB},\beta_3^{HB},\beta_4^{HB},\beta_5^{HB}|\mathcal{D})\Pi_{p=1}^5 \mathcal{N}(Y_{1t,p}(0)|Y_{\mathcal{A}_p t}(0)^\top \beta_p^{HB},\sigma^2)$$

$$\approx \log\Big(\frac{1}{S}\sum_{s=1}^{S}\Pi_{p=1}^5 \mathcal{N}(Y_{1t,p}(0)|Y_{\mathcal{A}_p t}(0)^\top \beta_p^{HB},\sigma^2)\Big)$$

The subscript $s = 1,\ldots,S$ indexes the S-many posterior samples of the weight vectors.

The other three methods considered for this estimation task do not directly produce a posterior distribution on the weight vectors or a posterior predictive distribution as in Equation 3.5. However, for purposes of comparison, we imagine that the data likelihood follows a Gaussian distribution,

and the equivalent to the predictive distribution for the other methods is $N(\hat{Y}_{1t}, \sigma^2)$. In the style of Equation 3.6, we define

$$\Lambda^{SC}(\{Y_{1t,p}\}_{p=1}^5) \triangleq \log(\Pi_{p=1}^5 \mathcal{N}(Y_{1t,p}(0)|\hat{Y}_{1t,p}^{SC}(0), \sigma^2))$$
$$\Lambda^{MC}(\{Y_{1t,p}\}_{p=1}^5) \triangleq \log(\Pi_{p=1}^5 \mathcal{N}(Y_{1t,p}(0)|\hat{Y}_{1t,p}^{MC}(0), \sigma^2))$$

These predictive likelihoods can be used to compare the performance of the estimators considered below. Recall that a subset of the donor units will be masked off in each study. Because performance depends, perhaps obviously, on whether the unavailable donors are important for the task of modeling the target unit, we consider two cases: one in which the unavailable donors are less important, in the sense that their weights in the vector β^{sim} are small, and one in which donors with large weights in β^{sim} are unavailable in some studies.

3.3.1 Unimportant Missing Units

In this case, four of the 16 donor units are unavailable in each study. None of the units with missing data have weights larger than 0.1 in the vector β^{sim}, and the four units with the largest weights are available in all five studies. Overall, 8 of the 16 donor units have missing data in at least one study.

	Λ
SCM-indiv	-14.3
SCM-complete	-11.3
MC-NNM	-12.2
HB	-12.3

Table 3.1: When only unimportant units have missing data, SCM-complete assigns the highest predictive likelihood to the held-out data.

In this stylized example, the conventional synthetic control estimator that uses only the set of donors available in all five studies (SCM-complete) enjoys a unique advantage, because none of the missing units have a large contribution to the target unit in the generative model. In this particular

circumstance, ignoring the units with some missing data effectively removes nuisance information, allowing a more accurate estimate of the target unit. This scenario is more favorable than what we are likely to encounter in practice. Independently fitting synthetic control estimators for each study results in the worst performance among the four methods. Although the most "important" donors are available in each study, with 12 donors and only 5 pre-treatment time periods, constraining the weight vectors to the simplex appears to provide insufficient regularization to prevent overfitting.

In both this example and the following, the matrix completion method (MC-NNM) also has an advantage in the sense that its assumption of an underlying low-rank factorization is indeed true. Matrix completion does not perform quite as well as SCM-complete in this example, but does outperform SCM-indiv. The penalty parameter λ in Equation 3.2 was selected via cross-validation on subsets of the observed entries of the $\mathbf{Y}(0)$ matrix. The hierarchical Bayesian synthetic control estimator performed similarly to matrix completion. Following the example of Brodersen et al. (2015), the hyperparameter ϕ^2 in Equation 3.4 is chosen to reflect the "expected sparsity" of the corresponding weight vector β^{HB}.

3.3.2 Important Missing Units

In this case, three of the 16 donor units are unavailable in each study, and 8 of the 16 donor units have missing data in at least one study. The units with the two largest weights in β^{sim} are unavailable in some of the five studies.

	Λ
SCM-indiv	-16.5
SCM-complete	-14.2
MC-NNM	-13.6
HB	-13.1

Table 3.2: When important units have missing data, HB assigns the highest predictive likelihood to the held-out data.

Unsurprisingly, all four methods perform worse when the "important" donor units have missing data. Among the four methods, the hierarchical Bayesian synthetic control estimator exhibited the

smallest degradation in performance, and also had the best performance of the four estimators in this scenario. The SCM-complete estimator, which performed best in the previous scenario, experienced the largest degradation in performance, as it now discards all data from the most important donor units.

In real applications, we will not know whether the units with missing data are "important" for modeling the target unit. Adopting a hierarchical Bayesian approach serves well if the missing units are important, and comes at only a modest sacrifice in performance if not.

3.4 Application: Spanish Christmas Lottery

Next we turn our attention to research concerning the connection between economic conditions and voter sentiment toward incumbent politicians. This is a germane consideration following 2024, a year marked by a large number of national elections and a large number of incumbent losses in those elections. We follow the example of Bagues and Esteve-Volart (2016), who used results of a large national lottery in Spain to assess the effects of exogenous good economic conditions on voting behavior. While there is a ample evidence of a positive association between good economic conditions and votes received by incumbent political candidates (Lewis-Beck and Stegmaier, 2000), the mechanism driving that association is challenging to discern, with several competing theories proposed in the political science literature (Lewis-Beck and Paldam, 2000). Lotteries can offer a helpful perspective on this issue, as they create exogenous economic shocks that are essentially independent of the incumbent politician's actions. The Spanish Christmas lottery is a particularly promising case study due to its high participation rate and geographic concentration of winners. Each year, Spain hosts a national lottery with winners drawn in late December. Approximately 75% of the adult population participates, and ticket sales exceed EUR 3bn. Depending on the year, between 150 and 180 "grand prize" tickets are issued, and the structure of the ticket sale process is such that the grand prize-winning tickets are often sold in a single province or a small number of provinces, rather than spread out across the country. Group purchases of tickets among family and friends are common, and consequently the number of households sharing the

grand prize can be much larger than the number of winning tickets. Due to the scale of the lottery and the geographic concentration of winners, the influx of cash tends to be significant relative to a province's GDP: Bagues and Esteve-Volart reported that over 1986-2011, the amount of the grand prize was roughly 3.5% of GDP in the province(s) where the winners were located (Bagues and Esteve-Volart, 2016).

To investigate the political consequences of these lottery winnings, Bagues and Esteve-Volart incorporated data from political surveys and voting records. We focus on the political survey data in this work, specifically a survey question asking, "If a national election were held tomorrow, which party would you vote for?". Within each province, we are interested in the proportion of respondents who indicate they would vote for the incumbent party. Surveys are conducted each year in January, April, July, and October. If the Christmas lottery grand prize is won by residents of a particular province, how does that affect the rate of support for the incumbent political party among residents of that province? A synthetic control approach to estimating that effect consists of modeling support for incumbent in the winning province as a convex combination of support for incumbent in non-winning provinces in time periods leading up to the lottery draw, then in the January survey following the lottery drawing, taking the same convex combination of the outcomes for non-winning provinces to estimate the counterfactual support for incumbent in the province where the grand prize was won.

In the following, we focus on the provinces of Barcelona and Granada, estimating the effects of lottery grand prize wins between the years 1988 and 2008.

3.4.1 Barcelona

Lottery outlets in Barcelona sold grand prize-winning tickets in December of 1988, 1992, 1994, 1998, 2005, and 2007. What was the effect of those lottery wins on the political support for the incumbent government among residents of Barcelona? Complete data is also available for Barcelona in an additional seven years during which Barcelona had no grand prize wins. We consider this data set to contain 13 individual studies, each consisting of political survey data

from April, July, and October of a given year, plus January of the following year. In each study, January is considered to be the treatment period. The following model is fit to the data from the non-treatment periods, i.e. April, July, and October:

$$
\begin{aligned}
Y_{1t,p}(0)|\beta_p^{HB} &\sim N(Y_{\mathcal{A}_p,t}(0)^\top \beta_p^{HB}, 0.02^2) \\
\beta_p^{HB}|\boldsymbol{\mu}, \boldsymbol{\varepsilon}_p &= \text{softmax}(\boldsymbol{\mu}_{\mathcal{A}_p} + \boldsymbol{\varepsilon}_p) \\
\mu_j &\overset{iid}{\sim} N(0, 2.5^2) && j = 1, \ldots, 49 \\
\varepsilon_{i,p} &\overset{iid}{\sim} N(0, 0.2^2) && i = 1, \ldots, |\mathcal{A}|.
\end{aligned}
$$

The values of hyperparameters, i.e. $\phi^2 = 2.5^2$ and $\gamma^2 = 0.2^2$, were chosen to maximize the predictive likelihood of the January outcomes for Barcelona in the seven available years in which Barcelona did not receive the treatment.

Table 3.3 displays the weights placed on the control provinces to construct the synthetic control for Barcelona in each year. Granada, Gipuzkoa, Sevilla, and Almería all have large weights in the years for which their data is available. In 1998, none of those four provinces were available as control units. Instead, Lleida and Vizcaya have large weights in 1998, whereas those two provinces have weights less than 10% in all other years.

Province	Synthetic Control Weights					
	'88	'92	'94	'98	'05	'07
Granada	0.32	0.31	NA	NA	0.44	0.55
Gipuzkoa	0.21	0.33	0.27	NA	0.17	0.25
Sevilla	0.20	NA	0.22	NA	0.15	NA
Almería	0.14	0.11	0.21	NA	0.15	NA
Lleida	0.04	0.04	NA	0.22	NA	NA
Vizcaya	0.06	0.07	0.08	0.49	0.05	NA
Huesca	NA	0.07	0.12	NA	NA	0.13
Santa Cruz de Tenerife	0.01	0.01	0.01	0.04	0.01	NA
Palencia	NA	0.03	0.03	NA	NA	NA
Pontevedra	0.00	0.00	0.01	0.02	0.00	0.01
Cádiz	0.00	0.00	0.00	0.02	0.00	0.00

Table 3.3: Posterior means of synthetic control weights constituting synthetic Barcelona. Includes all provinces with a weight ≥ 0.02 in any of the six years.

During pre-treatment time periods, the rate of support for the (national) incumbent party in synthetic Barcelona follows a similar trajectory to the observed values for Barcelona, as depicted in Figure 3.2. The time periods displayed correspond to April, July, and October of each available year. Each year immediately following a lottery grand prize win in Barcelona is excluded, as we do not consider Barcelona to be under the control condition at those times.

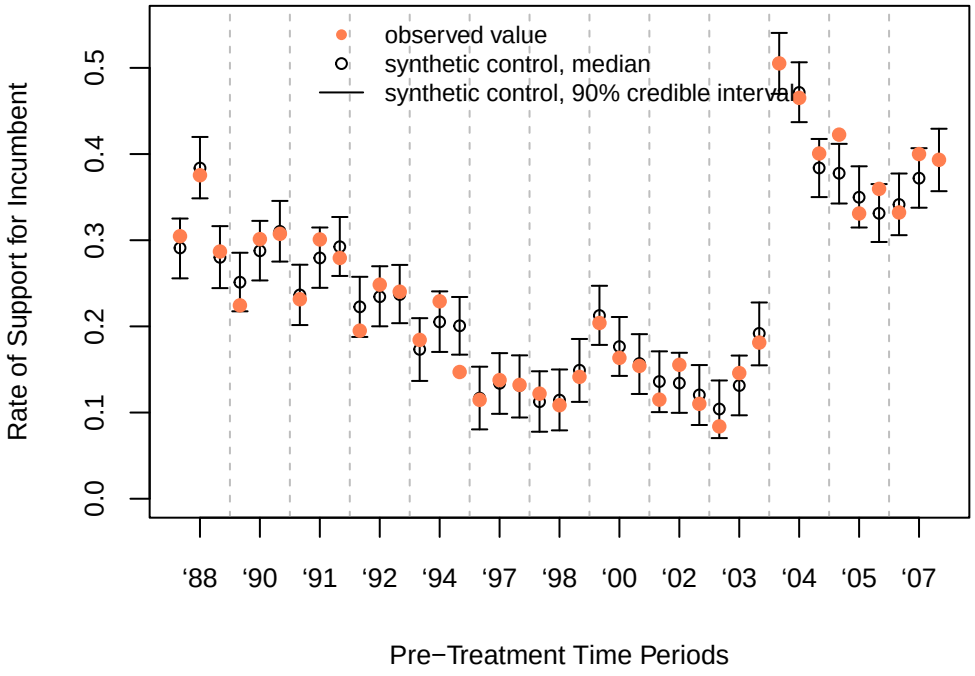


Figure 3.2: Political support for the incumbent party in synthetic Barcelona closely approximates the observed data for Barcelona during most pre-treatment time periods.

Posterior distributions for the treatment effects are summarized in Figure 3.3. For many of the years in which Barcelona won the grand prize, posterior uncertainty is too large to meaningfully distinguish the treatment effect from zero. Furthermore, the posterior median of the treatment effect is negative in two of the six years. Altogether, this analysis does not provide strong support for the notion that lottery winnings systematically influence Barcelona's support for the incumbent party.

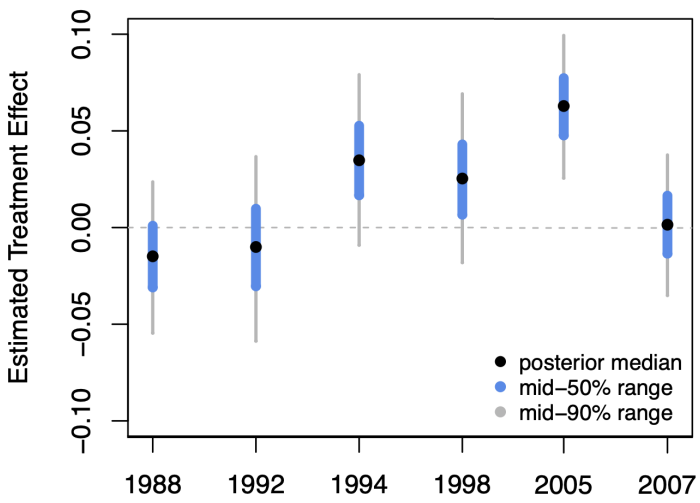


Figure 3.3: Estimated effect of lottery win on support for incumbent party in Barcelona

3.4.2 Granada

Lottery outlets in Granada sold grand prize-winning tickets in December of 1997 and 2002. Complete data is available for Granada in an additional 14 years during which Granada had no grand prize wins.

Granada and Barcelona differ considerably in population. Barcelona is the second most populous province in Spain, whereas Granada is a smaller province with a population roughly 16% of Barcelona's. Data values for the rate of political support for the incumbent are derived from sample surveys, which have fewer respondents from smaller provinces. Compared to Barcelona, Granada's estimates of political support for the incumbent are noisier, and this translates to greater uncertainty about the treatment effect.

Because the political surveys have a small (on the order of 50) and variable number of respondents from Granada, this sample variability is reflected in the likelihood model for the observations. In Equation 3.7 below, each time period has its own variance, $\sigma^2_{p,t}$, which is set to be inversely proportional to the number of survey respondents from Granada in that individual time period. The

full model fit to the data from the non-treatment periods, i.e. April, July, and October, is as follows:

$$
\begin{aligned}
Y_{1t,p}(0)|\beta_p^{HB} &\sim N(Y_{\mathcal{A}_p,t}(0)^\top \beta_p^{HB}, \sigma_{p,t}^2) && (3.7)\\
\beta_p^{HB}|\boldsymbol{\mu}, \boldsymbol{\varepsilon}_p &= \text{softmax}(\boldsymbol{\mu}_{\mathcal{A}_p} + \boldsymbol{\varepsilon}_p) \\
\mu_j &\overset{iid}{\sim} N(0, 2.5^2) && j = 1, \ldots, 49 \\
\varepsilon_{i,p} &\overset{iid}{\sim} N(0, 0.125^2) && i = 1, \ldots, |\mathcal{A}|.
\end{aligned}
$$

As in the Barcelona example above, the values of hyperparameters, i.e. $\phi^2 = 2.5^2$ and $\gamma^2 = 0.125^2$, were again chosen to maximize the predictive likelihood of the January outcomes for Granada in the available years in which Barcelona did not receive the treatment.

Table 3.4 displays the weights used to construct the synthetic control in each year. Málaga and Barcelona are large contributors to the synthetic Granada in both years. Madrid also had a large weight in 1997, but was unavailable in 2002; much of its weight shifted to Huelva instead.

Province	Synthetic Control Weights	
	'97	'02
Málaga	0.310	0.299
Barcelona	0.239	0.231
Madrid	0.227	NA
Huelva	0.169	0.315
Zamora	NA	0.086
Cádiz	0.043	0.065
Teruel	0.010	NA

Table 3.4: Posterior means of synthetic control weights constituting synthetic Granada. Includes all provinces with a weight ≥ 0.001 in either year.

Using the hierarchical Bayesian synthetic control method from Section 3.2.3, the posterior predictive distribution for Granada's untreated outcomes (obtained via Equation 3.5) is noticeably more spread out than was the case for Barcelona, largely due to the greater sampling variability in the data for Granada's level of political support for the incumbent. Figure 3.4 compares the estimated synthetic control to the observed values in Granada during pre-treatment time periods. Still, the synthetic control adequately reproduces the actual Granada in the sense that nearly all of

the 90% credible intervals cover the actual observed value.

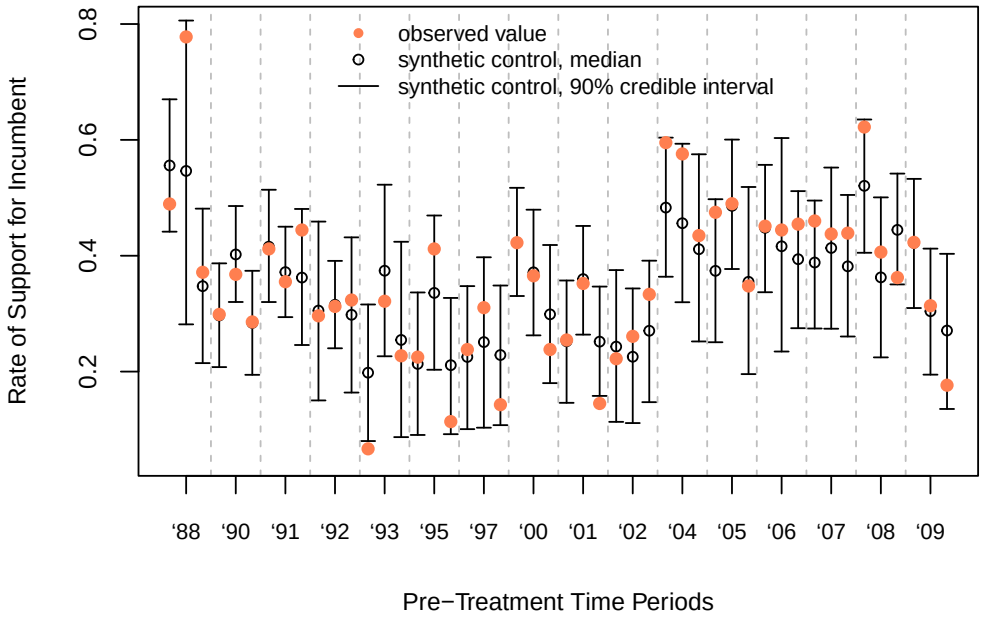


Figure 3.4: Estimates of political support for the incumbent party are less precise in Granada than in Barcelona, though most 90% posterior credible intervals for synthetic Granada cover the actual observed value during pre-treatment periods.

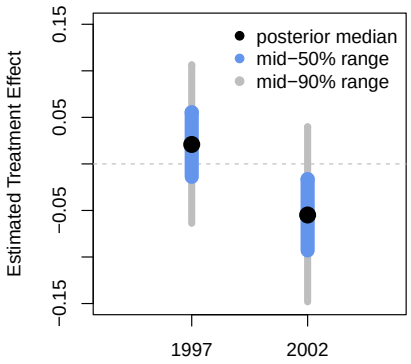


Figure 3.5: Estimated effect of lottery win on support for incumbent party in Granada

Figure 3.5 illustrates that the remaining posterior uncertainty regarding the treatment effects

is sufficiently large that we cannot confidently conclude that a grand prize lottery win affects Granada residents' rate of support for the incumbent political party. However, in both Granada and Barcelona, the median estimated effect was largest when the amount won was largest. This pattern is highlighted in Figure 3.6.

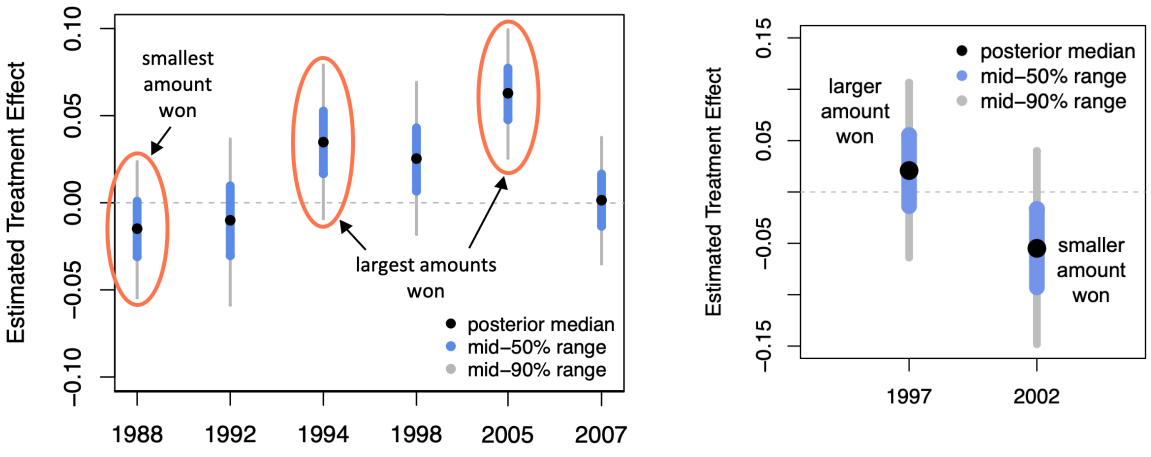


Figure 3.6: For both Barcelona and Granada, the estimated treatment effects are largest for the years in which the amount of money won was largest

Recall that multiple grand prize-winning tickets are sold each year. Often those tickets are all sold in a single province, but in some years they are distributed across several provinces. Consequently, the nominal amount won varies, as well as the amount relative to the size of the economy of the winning province(s).

In 2005 and 1994, Barcelona was the sole province where grand prize-winning tickets were sold. The amounts won in Barcelona were largest in those two years, with 2005 the larger of the two. Correspondingly, the estimated effect is largest for 2005, and second-largest for 1994. Among the years depicted in Figure 3.6, Barcelona won the smallest amount in 1988, totaling less than 0.02% of the province's GDP, compared to approximately 0.50% in 2005. The estimated effect of Barcelona's winnings in 1988 was also the smallest, with a negative median and distribution that implies the effect could plausibly be zero.

Likewise for Granada, the estimated effect has a positive median in 1997, when Granada was the sole winner. On the other hand, winning tickets were sold in nine different provinces in 2002, and the amount won in Granada was negligible. For 2002, the estimated effect of Granada's winnings had a negative median. Detailed exploration of a possible relationship between the specific amount won and the change in support for the incumbent party is deferred to future work.

3.5 Discussion

In this chapter, we develop a hierarchical Bayesian synthetic control estimator. The key contribution of this method is that it allows for the sharing of information across multiple related synthetic control studies, and allows the pool of available donor units to vary across the individual studies. The simulation studies of Section 3.3 demonstrate the potential benefits of this approach, which are most notable when donor units that are relevant for modeling the target unit are missing in some time periods.

Following the example of Doudchenko and Imbens (2016), this chapter focuses on using only the outcome values of the donor and target units to learn synthetic control weights. However, the hierarchical Bayesian method proposed herein could be extended to incorporate covariate data as well.

As discussed in Section 3.4.2, the "support for incumbent" response variable considered in the Spanish Christmas lottery application is derived from sample surveys. Due to varying numbers of respondents from each Spanish province, the measurement error differs across provinces and across time periods. When modeling Granada as the target unit, this sampling variability was encoded in the data likelihood, Equation 3.7. This approach could be extended to explicitly incorporate the differing amounts of measurement error among the other provinces used to construct the synthetic control. That information is already indirectly incorporated by the variability in the observed data from each province, but it may be interesting to explore whether explicitly incorporating it in the modeling approach would systematically shift the weights placed on each province comprising the synthetic control. Such a method could be broadly applicable to synthetic control studies in which

data quality differs among units.

Another possibly interesting direction for future work is to adapt this information sharing strategy to allow weight vectors that are not constrained to live on the simplex. A number of works in the synthetic control literature have considered relaxing that constraint on the weights, allowing negative weights and weights greater than one (Doudchenko and Imbens, 2016; Amjad et al., 2018; Kim et al., 2020). Indeed, the matrix completion approach of Athey et al. (2021) implicitly dispenses with this requirement as well.

Chapter 4: Differentially Private Inference via Noisy Optimization

As discussed in previous chapters, modern statistical tools can provide novel insights into social issues. However, when data sets contain the records of individuals, data analysis itself must be handled carefully to avoid harm from compromising privacy. This requires a delicate balance in order to make useful inferences about a population, without revealing too much about any one person. In recent decades, differential privacy has emerged as a rigorous framework for defining and quantifying the privacy loss incurred in data analysis. Informally, differential privacy is a guarantee that the conclusions drawn from analyzing a database only differ a negligible amount depending on whether a particular individual is included or excluded from that database. The privacy paradigm considered herein assumes that a trusted curator holds the data of individuals in a database, which a statistician may query without directly accessing the data. Privacy guarantees are achieved by injecting noise before returning a response to the statistician. This chapter develops an optimization-based approach to computing differentially private versions of M-estimators from the robust statistics literature. Inspired by the work of Song et al. (2013); Bassily et al. (2014); Lee and Kifer (2018); Feldman et al. (2020), the central idea of these methods is to add noise to every iterate of a gradient-based optimization routine such that each iterate satisfies a targeted differential privacy guarantee. The resulting private estimators converge with high probability to a small neighborhood around their non-private counterparts, and due to asymptotic normality, can also be extended to provide parametric confidence intervals.

The rest of this chapter is organized as follows. Section 4.1 reviews basic concepts of differential privacy and M-estimators, as well as related literature. Section 4.2 outlines a noisy gradient descent algorithm for producing differentially private M-estimators. Section 4.3 adapts the ideas of Section 4.2 to second-order methods, constructing a private version of Newton's method. Section 4.4 presents an approach for constructing confidence regions, including a small-sample bias

correction. Practical considerations for the deployment of these methods are discussed throughout, together with theoretical convergence results. Section 4.5 concludes the chapter with a discussion of results and future research directions. Proofs are deferred to Appendix D, together with real-data applications and empirical comparisons to existing methods.

4.1 Background

4.1.1 Notation

For a vector $v \in \mathbb{R}^m$, we write $\|v\|_2^2 = v^\top v$. For a positive definite matrix A, we denote $\|v\|_A^2 = v^\top A v$ and $\|A\|_2 = \max_{v:\|v\|_2=1} \|Av\|_2$. We denote the smallest and largest eigenvalues of a symmetric matrix A by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. We write $f = O(g)$ when $f(\cdot) \leq Cg(\cdot)$ for any admissible arguments of $f(\cdot)$ and $g(\cdot)$ and some positive constant C. Similarly, we write $f = \Omega(g)$ when $f(\cdot) \geq cg(\cdot)$ for any admissible arguments of $f(\cdot)$ and $g(\cdot)$ and some positive constant c. We write $f \asymp g$ when both $f = O(g)$ and $f = \Omega(g)$. For sequences of random variables $\{X_n\}$ and $\{Y_n\}$, we write $X_n = O_p(Y_n)$ to denote boundedness in probability, i.e., for every $\epsilon > 0$, there exist M and N such that $\mathbb{P}\left(\left|\frac{X_n}{Y_n}\right| < M\right) > 1 - \epsilon$ for all $n \geq N$. We write $x_{1:n} \in \mathbb{R}^{n\times m}$ to denote the data matrix with i^{th} row equal to $x_i \in \mathbb{R}^m$. For any two matrices $x_{1:n}, x'_{1:n} \in \mathbb{R}^{n\times m}$, we define their Hamming distance $\mathrm{d_H}(x_{1:n}, x'_{1:n}) := \left|\{i = 1, \ldots, n : x_i \neq x'_i\}\right|$ to be the number of coordinates which differ between $x_{1:n}$ and $x'_{1:n}$. For $\theta \in \mathbb{R}^p$ and $r > 0$, we write $\mathcal{B}_r(\theta)$ to denote the Euclidean ball of radius r around θ. For a set (event) E, let $\mathbb{1}_E$ or $\mathbb{1}\{E\}$ denote the indicator function.

4.1.2 Gaussian Differential Privacy

A *random function* $h : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ maps each $x_{1:n} \in \mathbb{R}^{n\times m}$ to a Borel-measurable random variable $h(x_{1:n})$. A statistic is a deterministic function of an observed data set $x_{1:n}$, e.g., the sample mean $\bar{x}$, whereas $h(x)$ is a randomized estimate, i.e., the output of a randomized algorithm obtained by perturbing the deterministic output $\bar{x}$.

Definition 1 ((ϵ, δ)-differential privacy)**.** *Let $\varepsilon, \delta > 0$. A random function h is (ε, δ)-differentially*

private (DP) if and only if for every pair of data sets $x_{1:n}, x'_{1:n} \in \mathbb{R}^{n\times m}$ *such that* $\mathrm{d_H}(x_{1:n}, x'_{1:n}) = 1$ *and all Borel sets* $B \subseteq \mathbb{R}^p$*, we have* $\mathbb{P}[h(x_{1:n}) \in B] \leq e^{\varepsilon}\mathbb{P}[h(x'_{1:n}) \in B] + \delta$.

The probabilities in Definition 1 are computed over the randomness of the function h, for any fixed neighboring data sets $x_{1:n}$ and $x'_{1:n}$. The most basic approach for constructing such a randomized function consists of adding random noise to a (deterministic) statistic, as we will discuss below. The definition of (ε, δ)-DP limits the ability of an adversary to identify the presence of an individual in a data set based on released outputs, as it is difficult to distinguish between the distributions of $h(x_{1:n})$ and $h(x'_{1:n})$. This leads to a hypothesis testing interpretation of differential privacy pointed out by Wasserman and Zhou (2010). Indeed, one can interpret differential privacy as a protection guarantee against an adversary that tests two simple hypotheses of the form $H_0 : x_i = s$ vs. $H_1 : x_i = t$. Privacy is ensured when this testing problem is difficult, and (ε, δ)-DP assesses the hardness of the problem via an approximate worst-case likelihood ratio of the distributions of $h(x_{1:n})$ and $h(x'_{1:n})$ over all neighboring data sets.

Building on the hypothesis testing interpretation, Dong et al. (2022) advocated for a new definition of Gaussian differential privacy that we will use for the remainder of this chapter. The definition involves a transparent interpretation of the privacy requirement: determining whether any individual is in a data set is at least as hard as distinguishing between two normal distributions $N(0, 1)$ and $N(\mu, 1)$ based on one random draw. The formal definition is a little more complex:

Definition 2 (Gaussian differential privacy)**.** *Let* $h : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ *be a random function.*

1. *We say that* h *is* f*-DP if any* α*-level test between simple hypotheses of the form* $H_0 : x_i = t$ *vs.* $H_1 : x_i = s$ *has power function* $\beta(\alpha) \leq 1 - f(\alpha)$*, where* f *is a convex, continuous, non-increasing function satisfying* $f(\alpha) \leq 1 - \alpha$ *for all* $\alpha \in [0, 1]$.

2. *We say that* h *is* μ*-Gaussian differentially private (GDP) if* h *is* f*-DP and* $f(\alpha) \geq \Phi(\Phi^{-1}(1-\alpha) - \mu)$ *for all* $\alpha \in [0, 1]$*, where* $\Phi(\cdot)$ *is the standard normal cdf.*

The following notion of sensitivity is central in the construction of differentially private procedures. In particular, it is used in the most basic algorithms that make some output $h(x)$ private by

simply releasing $h(x)+u$, where u is an independent noise term whose variance is scaled according to the sensitivity of h.

Definition 3 (Global sensitivity)**.** *Let $g : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ be deterministic. The global sensitivity of g is defined by* $\mathrm{GS}_g = \sup_{x_{1:n}, x'_{1:n} \in \mathbb{R}^{n\times m}} \{\|g(x'_{1:n}) - g(x_{1:n})\|_2 : \mathrm{d_H}(x_{1:n}, x'_{1:n}) = 1\}$.

The following theorem concerns a procedure that will be a primary building block for the algorithms proposed in Sections 4.2 and 4.3. This method is called the *Gaussian mechanism*, and can be easily tuned to achieve a desired (ε, δ)-DP (Dwork and Roth, 2014, Theorem A.1) or μ-GDP guarantee, as stated below:

Theorem 1 (Theorem 1 in Dong et al. (2022))**.** *Let $g : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ be a function with finite global sensitivity* GS_g*. Let Z be a standard normal p-dimensional random vector. For all $\mu > 0$ and $x \in \mathbb{R}^{n\times m}$, the random function $h(x) = g(x) + \frac{\mathrm{GS}_g}{\mu} Z$ is μ-GDP.*

The optimization methods proposed herein rely on compositions of a sequence of differentially private outputs computed using the same data set, where each step uses information from prior private computations. How to characterize the overall privacy guarantee of such an analysis is an important consideration. Intuitively, this guarantee should degrade as one composes more and more private outputs. Dong et al. (2022) very elegantly argued that in such scenarios, Gaussian differential privacy is very special: Its privacy guarantee under composition can be characterized as the K-fold composition of μ_k-GDP mechanisms, and is exactly μ-GDP, where $\mu = \sqrt{\mu_1^2 + \cdots + \mu_K^2}$. More fundamentally, GDP is a canonical privacy guarantee in an asymptotic sense, as a central limit theorem phenomenon shows that the composition of a large number of f-DP algorithms is approximately μ-GDP for some parameter μ (Dong et al., 2022).

4.1.3 M-estimators for Parametric Models

M-estimators are a simple class of estimators, stemming from robust statistics and constituting a general approach to parametric inference (Huber, 1964; Huber and Ronchetti, 2009). Suppose we observe an i.i.d. sample $x_1, \ldots, x_n$ with cdf F and wish to estimate a population parameter

$\theta_0 = T(F)$ lying in a parameter space Θ. The differentially private estimators proposed in this chapter rely on noisy optimization techniques that will lead to private counterparts of M-estimators $\hat{\theta} = T(F_n)$ of θ_0, defined as minimizers of the form

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \mathcal{L}_n(\theta) = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \rho(x_i, \theta) = \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_{F_n}[\rho(X, \theta)], \tag{4.1}$$

where F_n denotes the empirical distribution of $x_1, \ldots, x_n \in \mathbb{R}^m$. This class of estimators is a generalization of the class of maximum likelihood estimators.

When ρ is differentiable and convex, the estimator $\hat{\theta}$ can be viewed as a solution to

$$\frac{1}{n} \sum_{i=1}^{n} \Psi(x_i, \hat{\theta}) = 0,$$

where $\Psi(x, \theta) = \frac{\partial}{\partial \theta} \rho(x, \theta)$. M-estimators defined by functions $\Psi(x, \theta)$ that are bounded in $x \in \mathcal{X} \subseteq \mathbb{R}^m$ are particularly appealing in robust statistics, since the influence function of $T(F)$ is then bounded, ensuring infinitesimal robustness to outliers (Hampel et al., 1986).

Under mild general conditions, e.g., $\theta_0 = \arg\min_\theta \mathbb{E}_F[\rho(X, \theta)]$ and $\mathbb{E}_F[\Psi(X, \theta_0)] = 0$ (Huber, 1967), M-estimators are asymptotically normal: $\sqrt{n}(\hat{\theta} - \theta_0) \to_d \mathcal{N}(0, V(\Psi, F))$, where

$$\begin{aligned} V(\Psi, F) &:= M(\Psi, F)^{-1} Q(\Psi, F) M(\Psi, F)^{-1}, \\ M(\Psi, F) &:= -\mathbb{E}_F[\dot{\Psi}(Z, \theta_0)] \\ Q(\Psi, F) &:= \mathbb{E}_F[\Psi(Z, \theta_0) \Psi(Z, \theta_0)^\top], \end{aligned} \tag{4.2}$$

and $\dot{\Psi}(x, \theta) := \frac{\partial}{\partial \theta^\top} \Psi(x, \theta)$. Analogous results hold for their private counterparts described in Sections 4.2 and 4.3

4.1.4 Related Work

A sizable body of work is devoted to developing differentially private approaches for convex optimization in the context of empirical risk minimization. A first general optimization construc-

tion explored in the literature consists of perturbing the objective function so as to ensure that the resulting minimizer is differentially private. Representative work exploring this idea includes Chaudhuri and Monteleoni (2008), Chaudhuri et al. (2011), Kifer et al. (2012), Jain and Thakurta (2014) and more recently Slavkovic and Molinari (2021).

A related idea that is closer to the methods studied in this chapter considers iterative gradient-based algorithms, where the algorithm employs noisy gradients designed to ensure differential privacy at each iteration. Many such empirical risk minimization algorithms have been proposed (Song et al., 2013; Bassily et al., 2014; Wang et al., 2017a; Lee and Kifer, 2018; Iyengar et al., 2019; Bassily et al., 2019; Balle et al., 2020; Wang, 2018), motivated by online problems (Jain et al., 2012; Duchi et al., 2018), multiparty classification (Rajkumar and Agarwal, 2012), Bayesian learning (Wang et al., 2015), high-dimensional regression (Talwar et al., 2015; Cai et al., 2021, 2020), and deep learning (Abadi et al., 2016; Bu et al., 2020). Typical theoretical guarantees for such noisy gradient-based methods are stated in terms of the expected sub-optimality gap, i.e., the distance between the objective function at the private randomized solution and the non-private optimal solution. As will be discussed later in Section 4.2.2, such statements imply weaker estimation error guarantees than the main results of this chapter; the stronger guarantees are needed to construct valid confidence intervals.

While most of the privacy literature has focused on noisy stochastic gradient descent, in which gradients are estimated from a subset of the data, full gradient evaluations are used in this work, similar to standard implementations in non-private statistical software. Many existing results cannot be directly applied to our setting (Bassily et al., 2014; Feldman et al., 2020); moreover, even existing methods that use full gradients rely on different analyses (Cai et al., 2021, 2020), as we avoid truncation, allow unbounded input variables and parameter spaces, and explicitly track the impact of potentially bad starting values. All of the above can be achieved by considering locally strongly convex objective functions defining Fisher-consistent bounded-influence M-estimators.

On the optimization side, the present is related to inexact oracle methods (Sun et al., 2020; Devolder et al., 2014) and stochastic optimization (d'Aspremont, 2008; Ghadimi and Lan, 2012;

Wang et al., 2017b). Proofs for the convergence of noisy gradient descent and noisy Newton's method rely on showing that with high probability, the noise introduced to the gradients and Hessians has a negligible effect on the convergence of the iterates (up to the order of the statistical error of the non-noisy versions of the algorithms). The theory of inexact oracle methods similarly derives results for the output of iterative optimization algorithms when gradients and/or Hessians are computed up to a certain level of accuracy. However, whereas inexact oracles rely on approximate gradients, approximate gradients alone do not constitute inexact oracles unless the domain is bounded (Devolder et al., 2014), so the convergence results for the private estimators proposed in this chapter are not directly implied by existing literature in this area. The stochastic optimization literature is more directly applicable to the noisy gradient setting considered herein, and the differentially private gradients we construct can be viewed as a special instantiation of stochastic gradients, where noise is introduced not due to sampling error, but manually in order to preserve privacy. However, convergence results for the private estimators proposed in this chapter are also not direct consequences of existing literature on first-order (Ghadimi and Lan, 2012) or second-order (Wang et al., 2017b) methods.

The literature on private inference is relatively limited, and has only recently received some attention from computer scientists (Uhler et al., 2013; Gaboardi et al., 2016; Rogers and Kifer, 2017; Karwa and Vadhan, 2017; Awan and Slavković, 2018; Acharya et al., 2018; Wang et al., 2019; Covington et al., 2021; Chadha et al., 2021). This problem has also been studied in the statistics literature in a regression setting, but in many cases, suffers from the same drawbacks encountered in estimation, e.g. the need to assume bounded data (Yu et al., 2014; Wang et al., 2019), resorting to truncation (Barrientos et al., 2019), and requiring very large sample sizes for the methods to work well (Sheffet, 2017), especially in light of the expected $\sqrt{n}$-consistency of M-estimators (Barber and Duchi, 2014; Cai et al., 2021; Avella Medina, 2021). Similarly, while Avella Medina (2021) provides a simple method for differentially private inference, the method requires a minimum sample size that depends on problem parameters that are difficult to quantify in practice. Recent Bayesian approaches to private inference include Savitsky et al. (2019), Kulkarni

et al. (2021), and Peña and Barrientos (2021).

4.2 Randomized M-estimators via Noisy Gradient Descent

The theory of empirical risk minimization suggests various algorithms to compute the global optimum (4.1) over a convex set Θ (Boyd and Vandenberghe, 2004). One of the most elementary methods is gradient descent, which uses the iterates

$$\theta^{(k+1)} = \theta^{(k)} - \frac{\eta}{n} \sum_{i=1}^{n} \Psi(x_i, \theta^{(k)}). \tag{4.3}$$

In a classical statistical setting, one would continue the iterates (4.3) until a numerical condition is met, e.g., $\|\frac{1}{n} \sum_{i=1}^{n} \Psi(x_i, \theta^{(k)})\|_2 < \epsilon$, for a prespecified tolerance level ϵ. Since the optimization error can be made negligible by setting ϵ arbitrarily small, classical statistical theory usually ignores the effect of carrying out inference on the final iterate $\theta^{(K)}$ as opposed to the global optimum $\hat{\theta}$, which is typically analyzed theoretically. In principle, one can take K as large as needed in order to ensure that $\theta^{(K)}$ is numerically identical to $\hat{\theta}$. This is in stark contrast to the differentially private case, in which the number of iterations K must be set *before* running the algorithm in order to ensure that the final estimator respects a desired level of differential privacy. Intuitively, the larger the number of data (gradient) queries of the algorithm, the more prone it will be to privacy leakage. The gradient descent algorithm proposed in this section restricts to the class of M-estimators in order to satisfy the following uniform boundedness condition.

Condition 1. *The gradient of the loss satisfies* $\sup_{x \in \mathcal{X}, \theta \in \Theta} \|\Psi(x, \theta)\|_2 \leq B < \infty$.

The choice of the loss function ρ implies a known bound on the gradient B, which guarantees that $2B$ is an upper bound for the global sensitivity of $\Psi(x, \theta)$ by the mean value theorem. Thus, we can privately implement gradient descent by using the following noisy version of the iterates (4.3):

$$\theta^{(k+1)} = \theta^{(k)} - \frac{\eta}{n} \sum_{i=1}^{n} \Psi(x_i, \theta^{(k)}) + \frac{2\eta B \sqrt{K}}{\mu n} Z_k, \tag{4.4}$$

where the final estimate is again denoted by $\theta^{(K)}$. Here, η is the step size, and $\{Z_k\}_{k=0}^{K-1}$ is a sequence of i.i.d. standard p-dimensional Gaussian random vectors. The number of iterations K needs to be set beforehand and critically impacts the statistical performance of this estimator, as discussed below. For all examples in this chapter, the initial iterate $\theta^{(0)}$ is chosen in a non-data-dependent manner (e.g., $\theta^{(0)} = 0$).

4.2.1 Examples

The following numerical simulations illustrate the theory thus far, demonstrating the behavior of noisy gradient trajectories and the benefits of the proposed approach based on M-estimation, in contrast to gradient clipping. For more comparisons to clipping and other benchmark methods, see Appendix D.6.

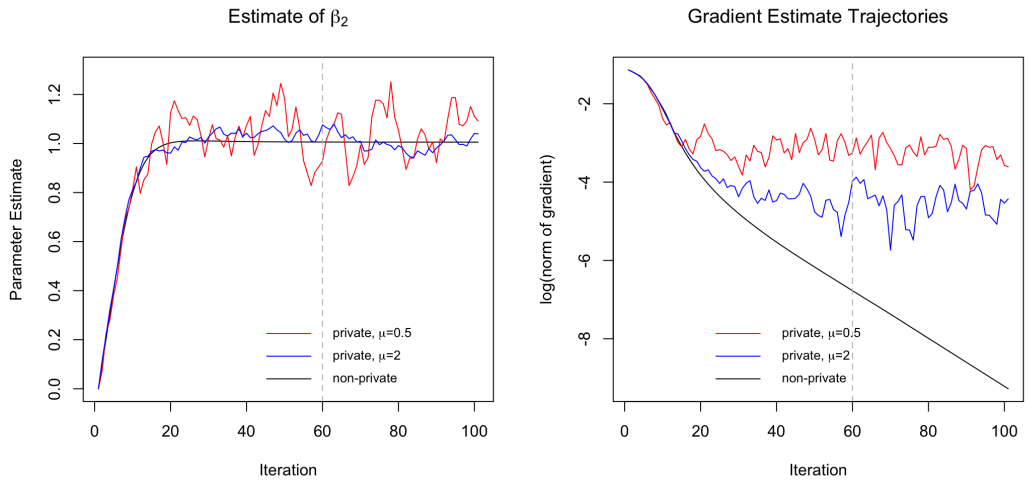
Linear Regression

We now explore the behavior of the noisy gradient descent algorithm when applied to simulated data from a linear regression model. The data $\{(x_i, y_i)\}_{i=1}^n$ are generated according to the model $y_i = x_i^\top \beta + \epsilon_i$, where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and the covariate vectors are given by $x_i = (1, z_i)^\top$, where $z_i \overset{i.i.d.}{\sim} N(0, \sigma_z^2 \mathbb{I}_3)$. The loss function is taken to be

$$\mathcal{L}_n(\beta, \sigma) = \frac{1}{n} \sum_{i=1}^n \left(\sigma \rho_c \left(\frac{y_i - x_i^\top \beta}{\sigma} \right) + \frac{1}{2} \kappa_c \sigma \right) w(x_i), \tag{4.5}$$

where ρ_c is the Huber loss with tuning parameter c, the constant κ_c is chosen to ensure consistency of $\hat{\sigma}$, and $w(x_i) = \min\left(1, \frac{2}{\|x_i\|_2^2}\right)$ downweights outlying covariates. By construction, the gradient of this loss function with respect to $\theta = (\beta, \sigma)$ has finite global sensitivity,so the noisy gradient descent iterates (4.4) can be used to estimate β and σ. The global sensitivity of $\nabla_\beta \mathcal{L}_n(\beta, \sigma)$ is $2\sqrt{2}c$ and the global sensitivity of $\nabla_\sigma \mathcal{L}_n(\beta, \sigma)$ is $\frac{1}{2}c^2$, resulting in a global sensitivity of $\sqrt{8c^2 + \frac{1}{4}c^4}$ for $\nabla_\theta \mathcal{L}_n(\theta)$ (this quantity takes the place of $2B$ in equation (4.4)). Setting $\beta = (1, 1, 1, 1)^\top$, $\sigma = 2 = \sigma_z$, $c = 1.345$, and $n = 1000$, Figure 4.1 shows example trajectories of the noisy gra-

dient descent iterates for the second coordinate of the parameter vector β. The optimization was initialized at $\beta^{(0)} = 0$ and $\sigma^{(0)} = 1$.



(a) (b)

Figure 4.1: Noisy gradient descent trajectories for linear regression. (a) Estimates of a single coordinate of the regression vector. (b) Gradient of the loss function evaluated at the current iterate, plotted on a log scale.

Figure 4.1(a) shows that in the early iterations, both private and non-private estimates move away from the initial value toward the true parameter value. While the non-private version converges to the true parameter value in later iterations, the μ-GDP version varies in a window around the true value as the iterates progress.

Since the random noise added to the gradient at each iteration (4.4) has the same fixed variance for a given sample size, the gradient of the loss function does not become arbitrarily small as the number of iterations increases, nor do the values at successive iterations become arbitrarily close to each other (see Figure 4.1(b)). From a practical standpoint, we can assess convergence of our algorithm by considering whether the gradient of the loss function is still large relative to the standard deviation of the random noise term. As noted above, the maximum number of iterations K must be set beforehand. However, if the loss function gradient is already small relative to the standard deviation of the noise term at some iteration $k < K$, empirical evidence suggests no

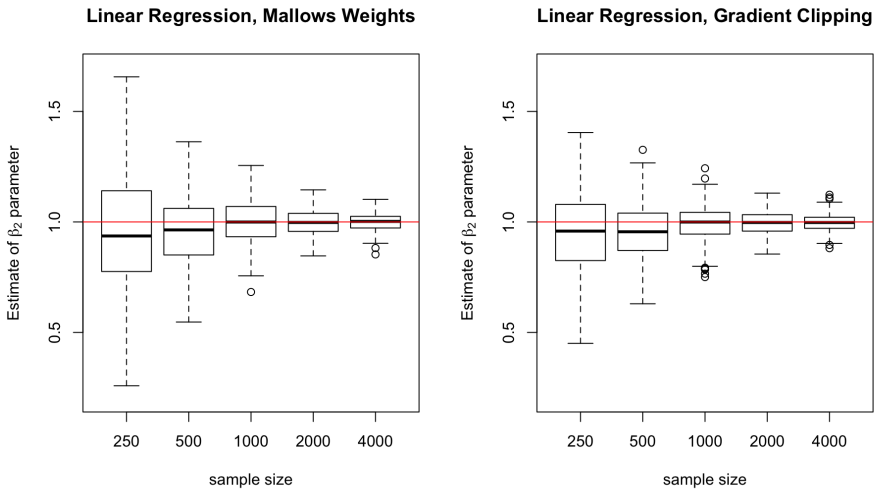
practical advantage to continuing through all K budgeted iterations to obtain $\left(\beta^{(K)}, \sigma^{(K)}\right)$.

Clipping and logistic regression

In this context, "clipping" refers to the practice of rescaling loss function gradients to obey a specified maximum magnitude. This directly implies a known, finite sensitivity for the gradient. Applying noisy gradient descent to optimize an objective function with bounded gradients is an alternative to explicitly clipping gradients to achieve finite global sensitivity. One motivation for avoiding clipping is that the resulting estimators may fail to be consistent: Suppose $x_1, \ldots, x_n \in \mathcal{X} \subseteq \mathbb{R}^m$ are i.i.d. according to F_{θ_0}. If one seeks a differentially private maximum likelihood estimator via clipped gradients, one will in fact be computing a differentially private counterpart of the clipped maximum likelihood estimator $\tilde{\theta} = T(F_n)$ defined as the solution to the equation $\frac{1}{n}\sum_{i=1}^{n} h_c\left(\nabla \log f(x_i; \tilde{\theta})\right) = 0$, where $f(\cdot; \theta_0)$ is the density function and $h_c(z) = z \min\left\{1, \frac{c}{\|z\|_2}\right\}$ is the multivariate Huber function (Hampel et al., 1986, p.239). As Song et al. (2021) demonstrate in the case of generalized linear models, this clipped estimator can be characterized as the minimizer of a Huberized version of the original loss function. However, while clipping guarantees a bounded sensitivity of the estimating equations, the clipped maximum likelihood estimator is in general not consistent, since the estimating equations are in general not unbiased, i.e., $\mathbb{E}_{F_{\theta_0}}\left[h_c\left(\nabla \log f(x_i; T(F_{\theta_0}))\right)\right] \neq 0$. Hence, even though gradient clipping is a common suggestion in the differential privacy literature, it is not the most appealing from a statistical viewpoint.

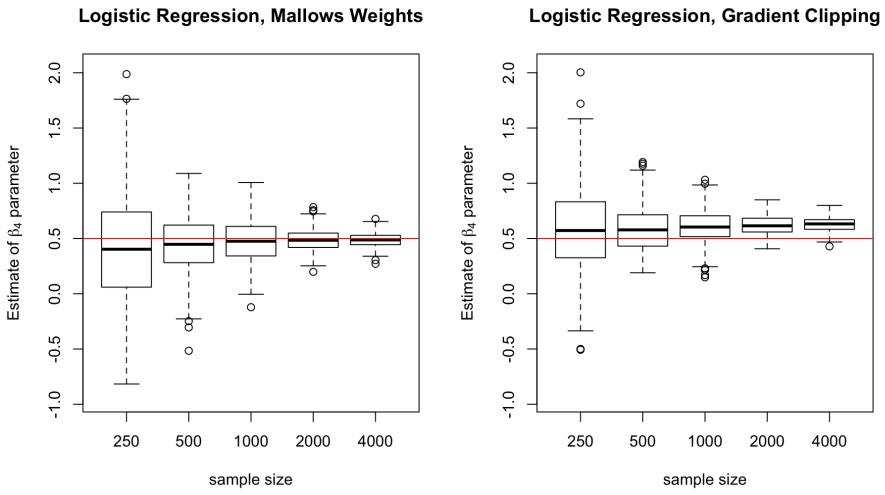
In a classical linear regression setting with squared loss and symmetric errors about zero, clipping does not lead to inconsistent estimators. However, this issue does arise when estimating the parameters of a logistic regression model (with the cross-entropy loss). Figures 4.2 and 4.3 demonstrate the behavior of the proposed noisy gradient descent algorithm (4.4) on simulated data from a linear regression model and a logistic regression model, using either Mallows weights (as in equation (4.5)) or gradient clipping. In the case of linear regression, both methods produce consistent estimators. For logistic regression, Figure 4.3(b) shows that the parameter estimates from gradient

clipping exhibit bias which does not shrink toward zero with the sample size.



(a) (b)

Figure 4.2: Gradient clipping and consistency. In the linear regression setting, both Mallows weights (plot (a)) and gradient clipping (plot (b)) approaches are consistent.



(a) (b)

Figure 4.3: Gradient clipping and consistency. Clipping results in a positive bias for logistic regression (plot (b)), which does not arise using Mallows weights (plot (a)).

Data for the linear regression simulation were generated according to the model $y_i = x_i^\top \beta + \epsilon_i$, where $\beta = (1.5, 1, -1, 0.5)^\top$, $\epsilon_i \overset{i.i.d.}{\sim} N(0, 2^2)$, and the covariate vectors are $x_i = (1, z_i)^\top$, where

$z_i \overset{i.i.d.}{\sim} N(0, \mathbb{I}_3)$. At each sample size, 400 repetitions were performed. The gradient of the loss function was clipped such that its ℓ_2-norm was no larger than 1. The optimization was initialized at $\beta^{(0)} = 0$, with σ^2 assumed to be known. For the logistic regression simulation, data were generated from the model $y_i \sim \text{Bernoulli}\left(\left\{1 + \exp(-x_i^\top \beta)\right\}^{-1}\right)$, with the same value of β and the same scheme for generating the covariates x_i as in the linear regression simulation. The instantiation with Mallows weights used a weighted version of the usual cross-entropy loss, i.e.,

$$\mathcal{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(-y_i \log\left(\frac{1}{1+\exp(x_i^\top\beta)}\right) + (1-y_i)\log\left(\frac{\exp(x_i^\top\beta)}{1+\exp(x_i^\top\beta)}\right)\right) w(x_i), \tag{4.6}$$

where $w(x_i) = \min\left(1, \frac{2}{\|x_i\|_2^2}\right)$. Since $\nabla\mathcal{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^n((1+e^{-x_i^\top\beta})^{-1} - y_i)x_i w(x_i)$, the update (4.4) uses $B = \sup_{x,\beta\in\mathbb{R}^4, y\in\{0,1\}} \|((1+e^{-x^\top\beta})^{-1} - y)xw(x)\|_2 = \sqrt{2}$.

The gradient clipping instantiation used the usual cross-entropy loss, with loss function gradients clipped at a threshold of 1. Again, 400 repetitions were performed at each sample size, and the optimization was initialized at $\beta^{(0)} = 0$.

4.2.2 Convergence Analysis

Taking $B = 0$ in the iterates (4.4) recovers standard gradient descent. Classical optimization theory tells us that for the choice $K = O(\log(1/\Delta))$, gradient descent incurs an optimization error of $\|\theta^{(K)} - \hat{\theta}\|_2 = O(\Delta)$. Therefore, $\log(n)$ steps suffice to ensure that the optimization error matches the parametric convergence rate $O\left(\sqrt{\frac{p}{n}}\right)$. We will require local strong convexity (LSC) of the loss $\mathcal{L}_n$ in a ball of radius r around the true parameter θ_0, as well as global smoothness. Strong convexity and smoothness are both standard conditions for the convergence analysis of gradient-based optimization methods (Boyd and Vandenberghe, 2004). Useful properties of strongly convex and smooth functions are reviewed in Appendix D.2.

Condition 2. *The loss function $\mathcal{L}_n$ is locally τ_1-strongly convex and τ_2-smooth, i.e.,*

$$\mathcal{L}_n(\theta_1) - \mathcal{L}_n(\theta_2) \geq \langle \nabla \mathcal{L}_n(\theta_2), \theta_1 - \theta_2 \rangle + \tau_1 \|\theta_1 - \theta_2\|_2^2, \quad \forall \theta_1, \theta_2 \in \mathcal{B}_r(\theta_0),$$
$$\mathcal{L}_n(\theta_1) - \mathcal{L}_n(\theta_2) \leq \langle \nabla \mathcal{L}_n(\theta_2), \theta_1 - \theta_2 \rangle + \tau_2 \|\theta_1 - \theta_2\|_2^2, \quad \forall \theta_1, \theta_2 \in \Theta \subseteq \mathbb{R}^p.$$

The following theorem shows that the noisy gradient descent iterates (4.4) are μ-GDP and converge to the non-private M-estimator as sample size n increases. Specifically, it shows that the private iterates lie in a neighborhood of the target non-private M-estimator whose radius is comparable to the privacy-inducing noise added in each noisy gradient descent step. The proof is in Appendix D.3.1, and hinges on the fact that the assumptions on the loss function and $\theta^{(0)}$ ensure that all iterates lie in the LSC ball $\mathcal{B}_r(\theta_0)$, with high probability (see Lemmas 6 and 7.)

Theorem 2. *Assume $\mathcal{L}_n$ satisfies Conditions 1 and 2, and suppose $\mathcal{L}_n(\theta^{(0)}) - \mathcal{L}_n(\hat{\theta}) \leq \frac{r^2}{4}\tau_1$ and $\hat{\theta} \in \mathcal{B}_{r/2}(\theta_0)$. Suppose $\mathcal{L}_n$ is twice-differentiable almost everywhere in $\mathcal{B}_r(\theta_0)$. Further let $\eta \leq \frac{1}{2}\min\left\{\frac{1}{\tau_2}, 1\right\}$, $n = \Omega\left(\frac{\sqrt{K\log(K/\xi)}}{\mu}\right)$, and $K = \Omega(\log n)$. Then (i) $\theta^{(K)}$ is μ-GDP, and (ii) $\|\theta^{(K)} - \hat{\theta}\|_2 \leq C\frac{\sqrt{K}(\sqrt{p}+\sqrt{2\log(K/\xi)})}{\mu n}$, with probability at least $1-\xi$, where C is a constant depending on B, τ_1, τ_2, and η.*

It is standard in differential privacy to present convergence results in terms of the expected sub-optimality gap (Bassily et al., 2014, 2019; Feldman et al., 2020; Song et al., 2021), building on stochastic gradient descent analysis. Assuming strong convexity, the best results among them give, in our notation,

$$\mathbb{E}[\mathcal{L}_n(\theta^{(K)})] - \mathcal{L}_n(\hat{\theta}) \leq C\frac{p}{\mu n}. \tag{4.7}$$

Instead, Theorem 2 proves a sub-Gaussian deviation error of the last iterate by adapting gradient descent analysis. This distinction is important for the following reasons:

1. A tight characterization of the convergence rate of $\theta^{(K)}$ itself is necessary for accurate inference. The analysis in this chapter bypasses the need to state convergence results in expectation.

2. Results regarding the convergence of the expected sub-optimality gap can be misleading in the sense that they do not point out the correct statistical cost of privacy induced by larger noise due to a large number of iterations. As a result, it is common to see theorems where the number of iterations is taken to be $K = Cn$ or $K = Cn^2$ (Bassily et al., 2014, 2019; Feldman et al., 2020; Song et al., 2021). It is easy to see from our proof of Lemma 8 that taking expectations in the bound (D.18) leads to

$$\mathbb{E}[\mathcal{L}_n(\theta^{(K)})] - \mathcal{L}_n(\hat{\theta}) \leq \kappa^K [\mathcal{L}_n(\theta^{(0)}) - \mathcal{L}_n(\hat{\theta})] + \frac{4\eta^2 B^2 K p}{\mu^2 n^2},$$

where $\kappa \in (0, 1)$. This bound is misleading because it suggests one could take $K = n$ and have a negligible error compared to the statistical error $O\left(\sqrt{\frac{p}{n}}\right)$. This is because by taking the expectation, one ignores a term linear in $\frac{\sqrt{K}}{n} Z_k$, which does not affect the bias but is the dominating variance term. In contrast, the high-probability analysis supporting Theorem 2 keeps track of the increasing effect of the total number of iterations on the variance, explicitly highlighting the importance of choosing a small number of iterations.

Theorem 2 immediately allows us to derive statistical properties of the noisy gradient descent estimator. The following result is a consequence of the triangle inequality and standard M-estimation theory, e.g. Huber and Ronchetti (2009, Corollary 6.7), which guarantees that $\|\hat{\theta} - \theta_0\|_2 = O_p\left(\sqrt{\frac{p}{n}}\right)$.

Corollary 1. *Assume the conditions of Theorem 2, and suppose θ_0 is such that $\mathbb{E}[\Psi(Z, \theta_0)] = 0$ and $\mathbb{E}[\dot{\Psi}(x, \theta)]$ is continuous and invertible in a neighborhood of θ_0. Then*

$$(i)\ \theta^{(K)} - \theta_0 = \hat{\theta} - \theta_0 + O_p\left(\frac{\sqrt{K \log K}}{\mu n}\right), \quad \text{and} \quad (ii)\ \sqrt{n}(\theta^{(K)} - \theta_0) \to_d N(0, V(\Psi, F_{\theta_0})).$$

Theorem 2 requires the sub-optimality gap of the objective function between the initial value $\theta^{(0)}$ and the global optimum $\hat{\theta}$ to be bounded by $\frac{r^2}{4}\tau_1$. The following proposition, proved in Appendix D.3.2, shows that an initial fixed number of noisy gradient descent iterations can be used to

ensure that the starting value condition is met.

Proposition 1. *Assume Conditions 1 and 2 hold and* $\eta \leq \frac{1}{2\tau_2}$. *Let* $R = \|\theta^{(0)} - \hat{\theta}\|_2$.

Then there exists a constant $c_0 > 0$ *such that with probability at least* $1 - \xi_0$, *after* $K_0 = \frac{R^2}{\eta\Delta}$ *noisy gradient descent iterations* (4.4) *and for* $n \geq C_0 \frac{(R+K_0+1)\{4\sqrt{p}+2\sqrt{2\log(K_0/\xi_0)}\}\sqrt{K_0}}{\Delta\mu}$, *we have* $\mathcal{L}_n(\theta^{(K_0)}) - \mathcal{L}_n(\hat{\theta}) \leq \Delta$, *where* C_0 *is a constant which depends on* η *and* B.

Taking $\Delta = \frac{r^2}{4}\tau_1$ in Proposition 1, we see that $K_0 = \frac{4}{\tau_1 r^2 \eta}\|\theta^{(0)} - \hat{\theta}\|_2^2$ iterations of noisy gradient descent are sufficient to ensure that $\theta^{(K_0)}$ meets the initialization condition of Theorem 2. Both Theorem 2 and Proposition 1 require the minimum sample size scaling $n = \Omega\left(\frac{\sqrt{K\log K}}{\mu}\right)$. Therefore, the additional $K_0 \asymp \|\theta^{(0)} - \theta_0\|_2^2$ iterations needed to ensure the starting value assumption of Theorem 2 do not substantively affect the conclusion of the theorem, since that result already assumes that $K = \Omega(\log n)$.

4.3 Randomized M-estimators via Noisy Newton's Method

Using similar ideas, one can construct a noisy version of Newton's method, as a second-order alternative to the noisy gradient descent algorithm described in Section 4.2. Recall that the classical Newton-Raphson algorithm finds the global optimum $\hat{\theta}$ of the objective (4.1) via the iterations

$$\theta^{(k+1)} = \theta^{(k)} - \left(\sum_{i=1}^{n} \dot{\Psi}(x_i, \theta^{(k)})\right)^{-1} \sum_{i=1}^{n} \Psi(x_i, \theta^{(k)}). \tag{4.8}$$

The key difference between the Newton-Raphson optimization procedure and gradient descent is the use of the Hessian term $\nabla^2\mathcal{L}_n$. The differentially private version of this algorithm will therefore add noise to both the gradient and the Hessian of the empirical loss function. Note that we will assume throughout this section that $\nabla^2\mathcal{L}_n$ exists everywhere.

4.3.1 Matrix-Valued Noise for Private Hessians

An intuitive idea for outputting a differentially private matrix is to add i.i.d. noise to each individual component. In many M-estimation problems, the matrix $M_n(\theta)$ can be viewed as an

empirical covariance matrix of the form $\frac{1}{n}\sum_{i=1}^{n} a_i a_i^\top$. Conveniently, the following result shows that for matrices with that form, we can indeed achieve privacy by adding a symmetric matrix with appropriately scaled element-wise i.i.d. Gaussian noise (Dwork et al., 2014, Algorithm 1):

Lemma 1 (Matrix Gaussian mechanism)**.** *Consider a data matrix $A \in \mathbb{R}^{n\times m}$ such that each row vector a_i satisfies $\|a_i\|_2 \leq 1$. Further define the function $h(A) = \frac{1}{n}A^\top A$, and let W be a symmetric random matrix whose upper-triangular elements, including the diagonal, are i.i.d. $\frac{1}{\mu n}N(0,1)$. Then the random function $\tilde{h}(A) = h(A) + W$ is μ-GDP.*

It is important for convergence of the iterations (4.8) that the matrix $M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \dot{\Psi}(x_i, \theta)$ be positive definite. Since we will use noisy versions of this matrix, we also want to guarantee that the randomized quantities remain positive definite. One obvious drawback of the matrix Gaussian mechanism described in Lemma 1 is that the noise is not positive definite, so the resulting privatized Hessian may also fail to be positive definite, especially in the case of small sample sizes. Note, however, that $\tilde{h}(A)$ can be projected onto a cone of positive definite matrices $\{H : H \succeq \varepsilon I\}$ defined by $\varepsilon > 0$ via the convex optimization problem $\tilde{h}(A)_+ = \arg\min_{H\succeq\varepsilon I} \|H - \tilde{h}(A)\|_2$. By definition, $\|\tilde{h}(A)_+ - \tilde{h}(A)\|_2 \leq \|h(A) - \tilde{h}(A)\|_2$, so the triangle inequality yields

$$\|\tilde{h}(A)_+ - h(A)\|_2 \leq \|\tilde{h}(A)_+ - \tilde{h}(A)\|_2 + \|\tilde{h}(A) - h(A)\|_2 \leq 2\|\tilde{h}(A) - h(A)\|_2.$$

Hence, the price to pay is no more than a factor of two, which does not affect the order of the convergence rate. In practice, the projection amounts to truncating the eigenvalues as $\max\{\lambda_j, \varepsilon\}$ (Boyd and Vandenberghe, 2004, p.399). The projected matrix is clearly also differentially private, as it results from a deterministic post-processing step applied to a differentially private output.

4.3.2 Differentially Private Newton's Method

We will require some regularity conditions on the Hessian matrix $\nabla^2\mathcal{L}_n(\theta)$. In particular, the Hessian will need to be factorizable in a way that allows the application of Lemma 1, which is typical for problems with linear predictors such as linear regression, robust regression, and

generalized linear models. We will also assume that the spectral norm of the Hessian is uniformly bounded.

Condition 3. *The Hessian is positive definite for* $\theta \in \mathcal{B}_r(\theta^*)$ *and of the form* $\nabla^2 \mathcal{L}_n(\theta) = \frac{1}{n} A^\top A = \frac{1}{n} \sum_{i=1}^n a(x_i, \theta) a(x_i, \theta)^\top$, *where* $\sup_{x,\theta} \|a(x, \theta)\|_2^2 \leq \bar{B} < \infty$.

Note that the constant $\bar{B}$ introduced in Condition 3 implies τ_2-smoothness with $\tau_2 = \frac{\bar{B}}{2}$.

Applying the matrix Gaussian mechanism of Lemma 1 to the Hessian and the Gaussian mechanism of Theorem 1 to the gradient, we can implement a noisy damped quasi-Newton method by updating the parameter estimate according to:

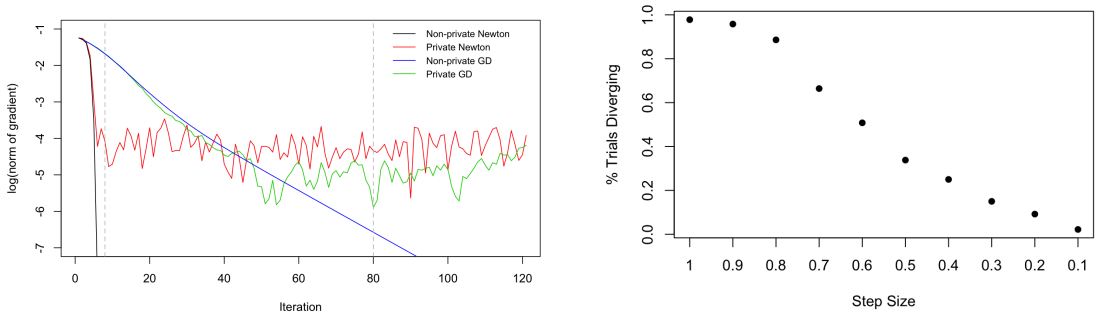
$$\theta^{(k+1)} = \theta^{(k)} - \eta H_k^{-1} \left(\frac{1}{n} \sum_{i=1}^n \Psi(x_i, \theta^{(k)}) + \frac{2B\sqrt{2K}}{\mu n} Z_k \right), \tag{4.9}$$

where $\eta > 0$ is the step size, $B \geq \sup_{x \in \mathcal{X}, \theta \in \Theta} \|\Psi(x, \theta)\|_2$ upper-bounds the gradients as before, $\bar{B}$ is as in Condition 3, $\{Z_k\}_{k=1}^K$ is a sequence of i.i.d. standard p-dimensional Gaussian random vectors, and $H_k = \frac{1}{n} \sum_{i=1}^n \dot{\Psi}(x_i, \theta^{(k)}) + \frac{2\bar{B}\sqrt{2K}}{\mu n} W_k$ is a noisy Hessian, where $\{W_k\}_{k=1}^K$ is a sequence of i.i.d. symmetric random matrices whose upper-triangular elements, including the diagonals, are i.i.d. standard normal. In practice, one can set the smallest eigenvalues of H_k to some small positive value ε, as discussed in Section 4.3.1. If $\varepsilon \to 0$, the effect will be asymptotically negligible and will not affect the theoretical conclusions stated in this work. When the step size $\eta = 1$, this algorithm corresponds to the usual, "pure" Newton's method. As in the case of noisy gradient descent, we choose the initial iterate $\theta^{(0)}$ in a non-data-dependent manner (e.g., $\theta^{(0)} = 0$), with specific choices of $\theta^{(0)}$ to be highlighted in the simulations to follow. We note that this noisy Newton's method algorithm can be interpreted as iterative applications of the SSP procedure (introduced by Vu and Slavkovic (2009) and applied to linear regression by Wang (2018)), as each step minimizes a certain least squares objective, although this interpretation does not directly imply the convergence results of Theorem 3 below.

The level of noise introduced to both the gradient and Hessian terms to ensure privacy is $O\left(\frac{1}{n}\right)$, which is appreciably smaller than the $\Theta\left(\frac{1}{\sqrt{n}}\right)$ fluctuations involved if we were to treat the gradient

and Hessian as empirical versions of $\nabla \mathcal{L}_n(\theta^{(k)})$ and $\nabla^2 \mathcal{L}_n(\theta^{(k)})$ (or the level of noise introduced by subsampling-type stochastic optimization methods) (Roosta-Khorasani and Mahoney, 2019). To our knowledge, the bounds derived for the iterative algorithms described in Sections 4.2 and 4.3 do not immediately follow from any known results in stochastic optimization.

The noisy Newton algorithm with $\eta = 1$ converges quadratically to a nearly-optimal neighborhood of the target parameter $\hat{\theta}$ when the starting value lies in a suitable neighborhood of the solution (specifically, the gradient of the loss function at the initial point must be small), as will be formalized below in Theorem 3. Figure 4.4(a) illustrates this improved performance of (noisy) Newton's method relative to (noisy) gradient descent, applying both methods to data simulated from a linear regression model.



(a) (b)

Figure 4.4: (a) Noisy Newton's method vs. gradient descent. The vertical axis records the norm of the gradient trajectory on a log scale. (b) Divergence of noisy damped Newton. When $\eta = 1$ (pure Newton), the algorithm may not converge if the initial point is far from the global optimum $\hat{\theta}$. Using smaller values of η remedies this problem.

In this example, the noisy Newton algorithm is calibrated to achieve 2-GDP in 8 iterations, while noisy gradient descent is calibrated to achieve 2-GDP in 80 iterations. This is meant to reflect the fact that Newton's method tends to converge faster, so in practice, we would schedule fewer iterations of Newton's method compared to gradient descent. To retain the 2-GDP privacy guarantees, we would terminate the algorithms at the gray lines on the plot (iteration 8 for Newton and iteration 80 for gradient descent), but for purpose of illustration, we have forced both algo-

rithms to continue, with the same amount of noise added in the extra steps. All optimizations were initialized at $\beta^{(0)} = 0$, with σ^2 assumed to be known. Taking the loss function (4.5), the Hessian is $\nabla^2 \mathcal{L}_n(\beta) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\left\{\left|\frac{y_i - x_i^\top \beta}{\sigma}\right| < c\right\} x_i x_i^\top w(x_i)$. Thus, to implement the update (4.9), we choose $\bar{B} = \sup_{x,\beta \in \mathbb{R}^4, y \in \mathbb{R}} \left\| \mathbb{1}\left\{\left|\frac{y - x^\top \beta}{\sigma}\right| < c\right\} x\sqrt{w(x)} \right\|_2^2 = 2$.

4.3.3 Initial Phase

The starting value condition required for the quadratic convergence of Newton's method cannot typically be guaranteed a priori. Indeed, even in general non-private settings, Newton's method with step size $\eta = 1$ may or may not converge depending on the initial point chosen. To avoid divergences due to poor initialization, we must consider an initial phase in which we approach the region of the solution via alternate means to Newton's method. Figure 4.4(b) displays the results of simulations using a damped version of Newton's method, in which the update steps are scaled by a step size $\eta < 1$. The data in this simulation, $\{(x_i, y_i)\}_{i=1}^{1000}$, are generated according to the model $y_i = x_i^\top \beta + \epsilon_i$, where $\beta = (1, 1, 1, 1)^\top$, $\epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} N(0, 2^2)$, and the covariate vectors are given by $x_i = (1, z_i)^\top$, where $z_i \overset{i.i.d.}{\sim} N(0, 2^2 \cdot \mathbb{I}_3)$. At each step size, 500 repetitions were performed, each initialized at $\beta^{(0)} = 0$ and $\sigma^{(0)} = 1$, and tuned for a 2-GDP privacy guarantee. For the same sample size, initial point, and true parameter values, we see that using damped Newton with a fixed step size η tends to help with the divergence problem: With all else equal, the proportion of trials leading to divergent iterates tends to decrease as the step size shrinks toward zero, although at very small step sizes, the algorithm may fail to converge within the budgeted number of iterations.

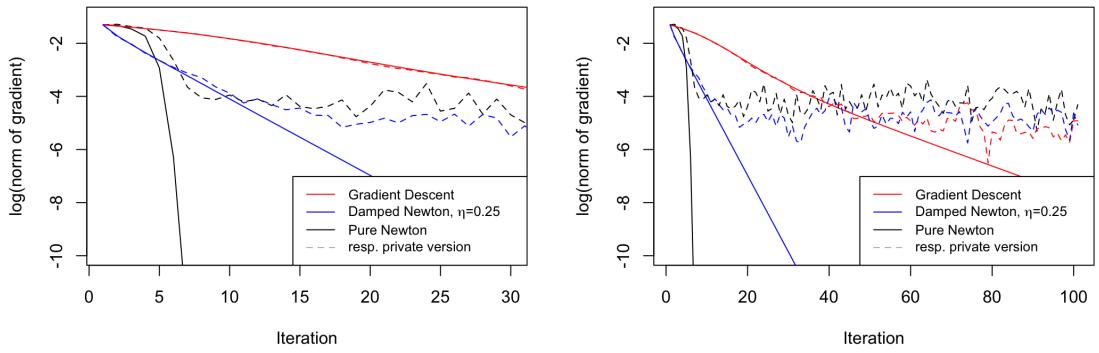


Figure 4.5: Trajectories of noisy gradient descent, pure Newton, and damped Newton algorithms. All optimizations were initialized at $\beta^{(0)} = 0$, with σ^2 assumed to be known.

Although taking smaller Newton step sizes guarantees global convergence of the algorithm, it only guarantees linear convergence, as illustrated in Figure 4.5. A popular strategy employed to in order to obtain a quadratically convergent Newton algorithm is to rely on a damped version of Newton's method, in which the update is scaled by an adaptively-chosen step size via backtracking line search (Boyd and Vandenberghe, 2004, Ch. 9.5.2). Below, we outline a differentially private implementation of backtracking line search for selecting these step sizes.

Backtracking Procedure: To avoid divergence associated with poor initializations in Newton's method, backtracking line search may be used to adaptively select a step size smaller than 1. The template for backtracking line search is as follows: At iteration k of Newton's method, set step size $\eta = 1$, and choose constants $\gamma \in (0, 1)$ and $\alpha \in (\frac{1}{2}, 1)$. Then,

$$\textbf{while } \mathcal{L}_n(\theta^{(k)} - \eta \nabla^2 \mathcal{L}_n(\theta^{(k)})^{-1} \nabla \mathcal{L}_n(\theta^{(k)})) - \mathcal{L}_n(\theta^{(k)}) > \\ - \alpha\eta \nabla \mathcal{L}_n(\theta^{(k)})^T \nabla^2 \mathcal{L}_n(\theta^{(k)})^{-1} \nabla \mathcal{L}_n(\theta^{(k)}) :$$

$$\textbf{do } \eta \leftarrow \gamma\eta.$$

That is, multiplicatively reduce the step size until the condition above (the Armijo-Goldstein condition) no longer holds.

At iteration k of the differentially private Newton's method from equation (4.9), differentially

private estimates of $\nabla \mathcal{L}_n(\theta^{(k)})$ and $\nabla^2 \mathcal{L}_n(\theta^{(k)})$ are already available. We can substitute these private quantities into the backtracking condition above without affecting the privacy guarantee, due to closure under postprocessing. For the moment, let H_k denote the private Hessian of the loss function at iteration k:

$$H_k = \frac{1}{n}\sum_{i=1}^{n} \dot{\Psi}(x_i, \theta^{(k)}) + \frac{2\bar{B}\sqrt{2K}}{\mu n} W_k,$$

and let G_k denote the corresponding gradient:

$$G_k = \frac{1}{n}\sum_{i=1}^{n} \Psi(x_i, \theta^{(k)}) + \frac{2B\sqrt{2K}}{\mu n} Z_k,$$

where K is the number of iterations budgeted for the Newton's method algorithm, and B, $\bar{B}$, W_k, and Z_k are as defined in Section 4.3.2.

The backtracking condition we wish to evaluate can then be privately approximated as

$$\mathcal{L}_n(\theta^{(k)} - \eta H_k^{-1} G_k) - \mathcal{L}_n(\theta^{(k)}) > -\alpha\eta G_k^T H_k^{-1} G_k. \tag{4.10}$$

The right-hand side of inequality (4.10) can be computed at no additional privacy cost, while the left-hand side requires a modification to achieve differential privacy. Recall that the loss function is of the form $\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \rho(x_i, \theta)$. Furthermore, we assume throughout that $\rho(x, \theta)$ is differentiable with respect to θ; and denoting $\Psi(x, \theta) = \frac{\partial}{\partial\theta}\rho(x, \theta)$, we assume an upper bound $B \geq \sup_{x\in\mathcal{X}, \theta\in\Theta} \|\Psi(x, \theta)\|_2$. It follows that

$$|\rho(x, \theta - \eta H_k^{-1} G_k) - \rho(x, \theta)| \leq \|\eta H_k^{-1} G_k\|_2 \|\Psi(x, \tilde{\theta})\|_2$$

for some $\tilde{\theta} \in (\theta, \theta - \eta H_k^{-1} G_k)$, implying that

$$|\rho(x, \theta - \eta H_k^{-1} G_k) - \rho(x, \theta)| \leq \eta \|H_k^{-1} G_k\|_2 B.$$

Thus, the global sensitivity of the function $\mathcal{L}_n(\theta^{(k)} - \eta H_k^{-1} G_k) - \mathcal{L}_n(\theta^{(k)})$ is $\frac{2\eta}{n}\|H_k^{-1} G_k\|_2 B$.

The backtracking condition may therefore be privately evaluated by checking the condition

$$\mathcal{L}_n(\theta^{(k)} - \eta H_k^{-1} G_k) - \mathcal{L}_n(\theta^{(k)}) + \frac{2B\eta \|H_k^{-1} G_k\|_2 \sqrt{T}}{n\mu} \xi_k > -\alpha\eta G_k^T H_k^{-1} G_k, \tag{4.11}$$

where ξ_k is a standard Gaussian random variable. Each evaluation of this condition satisfies $\frac{\mu}{\sqrt{T}}$-GDP. As with the number of algorithm iterations, the total number of backtracking steps must be budgeted ahead of time to ensure a specific privacy guarantee. Budgeting for T total backtracking steps, the entirety of the noisy Newton's method algorithm satisfies $\sqrt{2}\mu$-GDP. The one theoretical drawback of this proposal is that we do not have an a priori upper bound on the number of backtracking iterations needed before exiting the damped Newton phase, though this number is often not large in practice (see the example below).

As discussed in (Boyd and Vandenberghe, 2004, Ch. 9.5.3), this scheme for selecting step sizes causes Newton's method to exhibit an initial "damped" phase, followed by a "pure" phase in which all step sizes are equal to 1. Convergence in the "damped" phase is slower, while the "pure" phase is characterized by faster, quadratic convergence. Empirically, the differentially private counterpart exhibits similar behavior, as illustrated in Figure 4.6 below. Note that for implementation, we configured the private algorithm to discontinue backtracking in subsequent iterations after the first time a step size of 1 was chosen, as this performed favorably in numerical simulations.

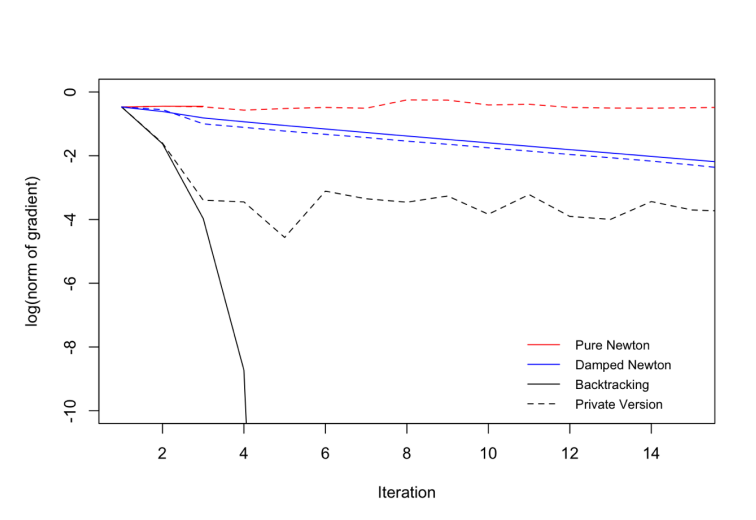


Figure 4.6: Behavior of Newton's method with backtracking line search, compared to pure and damped Newton's method

In this example, Newton's method is used to estimate the slope parameter in a linear regression model. The data set $\{(x_i, y_i)\}_{i=1}^{2000}$ was generated according to the model $y_i = x_i^T\beta + \epsilon_i$, where $\epsilon_i \overset{i.i.d.}{\sim} N(0,1)$, the covariate vectors are given by $x_i = (1, z_i)^T$, where $z_i \overset{i.i.d.}{\sim} N(0, \mathbb{I}_3)$, and the true slope is $\beta = (1,1,1,1)^T$. We take our loss function to be

$$\mathcal{L}_n(\beta, \sigma) = \frac{1}{n}\sum_{i=1}^{n} \rho_c(y_i - x_i^T\beta)w(x_i),$$

where ρ_c is the Huber loss function with tuning parameter c, and $w(x_i) = \min\left(1, \frac{2}{\|x_i\|_2^2}\right)$.

With an initial point $\beta_0 = (4,-2,0,4)^T$, both the private and non-private versions of pure Newton diverge. With the same initial point, backtracking line search performs 4 backtracking steps in the first iteration, then chooses a step size of 1 in subsequent iterations. (In this example, both the private and non-private algorithms performed the same number of backtracking steps, though in general those numbers often differ.) This switch to a step size of 1 allows for faster convergence compared to damped Newton. In this example, a fixed step size of approximately $\eta = 0.1$ was sufficiently small to avoid divergence. However, using this small step size at every iteration caused damped Newton to converge more slowly and output a less accurate estimate of β

within the budgeted number of iterations.

Practical Consideration: In privately evaluating the backtracking condition in Equation (4.11), it is possible for the added noise term to reverse the inequality, resulting in an incorrect choice of step size. This introduces two potential pitfalls: first, the algorithm may prematurely stop backtracking, and second, the algorithm may continue backtracking longer than necessary. A premature exit from backtracking can lead the algorithm to diverge; taking extra, unnecessary backtracking steps can cause the algorithm to exhaust its backtracking budget without reaching the pure Newton phase, which can prevent the algorithm from reaching a region around the true solution within the budgeted number of iterations (supposing that after the backtracking budget has been consumed, any remaining iterations of Newton's method proceed with a fixed step size equal to the most recent step size). The tuning parameter α influences the tradeoff between these two vulnerabilities. Small values of α make it easier to satisfy the backtracking exit condition, which tends to reduce the risk of unnecessary backtracking steps, but increase the risk of stopping backtracking too early. This tradeoff is illustrated in Figure 4.7 below.

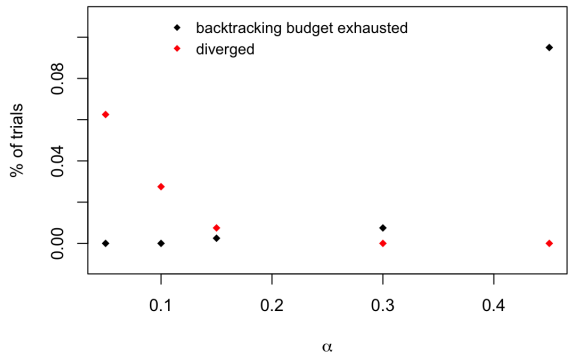


Figure 4.7: Behavior of 400 repetitions of noisy Newton's method with backtracking line search, applied to data generated from a logistic regression model with true $\beta = (1, 1, 1, 1)^T$ and initialization at $\beta^{(0)} = (-4, 6, -5, 6)^T$

Alternatives to this backtracking line search procedure are detailed in Avella-Medina et al. (2023). One suggestion is to begin with noisy gradient descent or a damped Newton procedure,

and privately check at each iteration whether the requisite starting condition for pure noisy Newton is satisfied. However, that strategy requires estimation of the local strong convexity constant of the loss function, which presents its own challenges.

4.3.4 Convergence Analysis

Assuming local strong convexity, it can be shown that the noisy Newton algorithm leads to the same statistical error bounds as noisy gradient descent, but requires fewer iterations to achieve convergence. In order to establish this result, we will also require the Hessian to be Lipschitz continuous, which is commonly assumed to establish the quadratic convergence of Newton's method under strong convexity (Boyd and Vandenberghe, 2004, Ch 9.5):

Condition 4 (Lipschitz continuity of Hessian)**.** *The Hessian $\nabla^2 \mathcal{L}_n(\theta)$ is L-Lipschitz continuous, i.e., $\|\nabla^2 \mathcal{L}_n(\theta_1) - \nabla^2 \mathcal{L}_n(\theta_2)\|_2 \leq L\|\theta_1 - \theta_2\|_2$ for all $\theta_1, \theta_2 \in \mathbb{R}^p$.*

The following theorem, proved in Appendix D.4, shows that under standard regularity conditions for robust M-estimators, $\Omega(\log\log n)$ iterations of the noisy Newton step (4.9) suffice to obtain a μ-GDP estimator lying in a neighborhood of $\hat{\theta}$ whose radius is proportional to the privacy-inducing noise of the algorithm.

Theorem 3. *Assume Conditions 1, 2, 3, and 4, and suppose $\hat{\theta} \in \mathcal{B}_{r/2}(\theta_0)$ and $\|\nabla \mathcal{L}_n(\theta^{(0)})\|_2 \leq \min\left\{\tau_1 r, \frac{\tau_1^2}{L}\right\}$.*
Let $n = \Omega\left(\frac{\sqrt{Kp\log(Kp/\xi)}}{\mu}\right)$ and $K = \Omega(\log\log n)$. The K^{th} noisy Newton iterate (4.9) with $\eta = 1$ satisfies (i) $\theta^{(K)}$ is μ-GDP, and (ii) $\|\theta^{(K)} - \hat{\theta}\|_2 \leq C\frac{\sqrt{Kp\log(K/\xi)}}{\mu n}$, with probability at least $1 - \xi$, where C depends on $(L, \tau_1, B, \bar{B})$.

As in the case of Corollary 1, it follows directly from Theorem 3 and standard M-estimation theory that the noisy Newton algorithm leads to μ-GDP estimators that are $\sqrt{n}$-consistent and asymptotically normally distributed.

Corollary 2. *Assume the conditions of Theorem 3 and let θ_0 be such that $\mathbb{E}[\Psi(Z, \theta_0)] = 0$ and $\mathbb{E}_F[\dot{\Psi}(x, \theta)]$ is continuous and invertible in a neighborhood of $\theta = \theta_0$. Then*

$$(i)\ \theta^{(K)} - \theta_0 = \hat{\theta} - \theta_0 + O_p\left(\frac{\sqrt{K \log K}}{\mu n}\right), \quad \text{and} \quad (ii)\ \sqrt{n}(\theta^{(K)} - \theta_0) \to_d N(0, V(\Psi, F_{\theta_0})).$$

4.4 Inference via Noisy Variance Estimators

Asymptotic normality of the private estimators proposed in the preceding sections facilitates the construction of private confidence regions for θ_0.

4.4.1 Private Sandwich Formula

The most common approach for constructing confidence intervals using an M-estimator $\hat{\theta}$ is to use an asymptotic pivot. In particular, under regularity conditions, we have $\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, V(\theta_0))$, where

$$V(\theta_0) := M(\theta_0)^{-1} Q(\theta_0) M(\theta_0)^{-1},$$
$$M(\theta_0) := -\mathbb{E}_F[\dot{\Psi}(X, \theta_0)], \text{ and}$$
$$Q(\theta_0) := \mathbb{E}_F[\Psi(X, \theta_0)\Psi(X, \theta_0)^\top]$$

are the analogs of equation (4.2). The asymptotic variance can accordingly be estimated via the plug-in sandwich estimator $V_n(\hat{\theta}) = M_n(\hat{\theta})^{-1} Q_n(\hat{\theta}) M_n(\hat{\theta})^{-1}$, where $M_n(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^n \dot{\Psi}(X_i, \hat{\theta})$ and $Q_n(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^n \Psi(X_i, \hat{\theta})\Psi(Z_i, \hat{\theta})^\top$. Consistency of $\hat{\theta}$ and the continuous mapping theorem show that $V_n(\hat{\theta}) \to_p V(\theta_0)$, so Slutsky's theorem justifies the use of the $(1-\alpha)$-confidence intervals $\mathrm{CI}_{1-\alpha}(\theta_{0j}) = \hat{\theta}_j \pm \frac{\sqrt{V_n(\hat{\theta})_{jj}}}{\sqrt{n}} z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard normal. However, in the differential privacy setting, this plug-in construction cannot be applied directly, as neither $M_n(\theta^{(K)})$ nor $Q_n(\theta^{(K)})$ are differentially private. Instead, we use the approach discussed in Section 4.3.1 to construct differentially private analogues of these two matrices: Assuming

Condition 3 holds, Lemma 1 suggests using the estimates

$$\tilde{M}_n(\theta^{(K)}) = M_n(\theta^{(K)}) + \frac{2\bar{B}}{\mu n} G_1 \quad \text{and}$$
$$\tilde{Q}_n(\theta^{(K)}) = Q_n(\theta^{(K)}) + \frac{2B^2}{\mu n} G_2,$$

where G_1 and G_2 are i.i.d. symmetric random matrices whose upper-triangular elements, including the diagonals, are i.i.d. standard normal. The differentially private matrices $\tilde{M}_n(\theta^{(K)})$ and $\tilde{Q}_n(\theta^{(K)})$ can be easily projected onto a cone of positive definite matrices $\{H : H \succeq \varepsilon I\}$ without paying a large statistical cost (see Section 4.3.1). The projected matrices $\tilde{M}_n(\theta^{(K)})_+ = \arg\min_{H \succeq \varepsilon I} \|H - \tilde{M}_n(\theta^{(K)})\|_2$ and $\tilde{Q}_n(\theta^{(K)})_+ = \arg\min_{H \succeq \varepsilon I} \|H - \tilde{Q}_n(\theta^{(K)})\|_2$, can be combined to construct the differentially private sandwich estimator:

$$\tilde{V}_n(\theta^{(K)}) = \tilde{M}_n(\theta^{(K)})_+^{-1} \tilde{Q}_n(\theta^{(K)})_+ \tilde{M}_n(\theta^{(K)})_+^{-1}. \tag{4.12}$$

Essentially the same idea was suggested by Wang et al. (2019) in the context of differentially private estimators computed via objective perturbation and output perturbation, although those techniques have some of the drawbacks mentioned in the introduction (e.g. assuming bounded data).

The composition property of μ-GDP estimators and consistency imply the following:

Proposition 2. *Assume the conditions of Theorems 2 or 3 hold. Then $\tilde{V}_n(\theta^{(K)})$ is $\sqrt{3}\mu$-GDP and $\tilde{V}_n(\theta^{(K)}) \to_p V(\theta_0)$.*

We can therefore release the $\sqrt{3}\mu$-GDP intervals $\mathrm{CI}^{priv}_{1-\alpha}(\theta_{0j}) = \theta_j^{(K)} \pm \frac{\sqrt{\tilde{V}_n(\theta^{(K)})_{jj}}}{\sqrt{n}} z_{1-\alpha/2}$. Although this construction is asymptotically valid, it tends to be too liberal in small samples. The correction proposed in the next subsection partially addresses this issue.

4.4.2 Finite-Sample Correction

We now discuss a correction to the noisy gradient descent and noisy Newton's method algorithms which leads to better performance in practice.

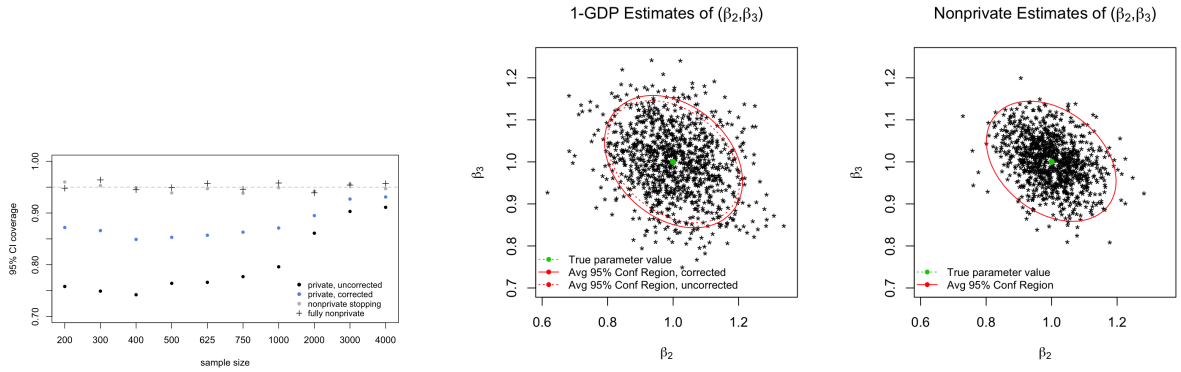
Noisy Gradient Descent

The formula (4.12), by construction, underestimates the variance of $\theta^{(K)}$ in finite samples, as it fails to account for the additional variability introduced by the privacy-preserving random noise mechanism. In the case of noisy gradient descent, this can be mitigated by making the following correction to equation (4.12):

$$\hat{V}_n(\theta^{(K)}) = \tilde{V}_n(\theta^{(K)}) + \frac{8\eta^2 B^2 K}{n\mu^2} I. \tag{4.13}$$

The correction (4.13) is motivated by the behavior of the noisy iterates of our algorithms at convergence. From the proof of Theorem 2, with high probability, every iteration of the algorithm grows closer to the non-private solution $\hat{\theta}$, up to a privacy-preserving noisy neighborhood of radius $\frac{4\eta^2 B^2 K}{n\mu^2}$. Once this neighborhood is reached, the iterates approximately behave as a random walk around the boundary. Since we inject Gaussian noise with variance $\frac{4\eta^2 B^2 K}{n\mu^2}$, we see that $\theta^{(K)}$ will be approximately within a $\frac{8\eta^2 B^2 K}{n\mu^2}$-neighborhood of $\hat{\theta}$, which is accounted for in equation (4.13). This corrected variance formula yields intervals $\mathrm{CI}^{cor}_{1-\alpha}(\theta_{0j}) = \theta^{(K)}_j \pm \sqrt{\frac{\tilde{V}_n(\theta^{(K)})_{jj}}{n} + 2\left(\frac{2\eta B\sqrt{K}}{n\mu}\right)^2} z_{1-\alpha/2}$.

Figure 4.8 demonstrates the practical impact of this correction, using noisy gradient descent on simulated data to estimate the parameters in multivariate linear regression. Data were drawn from the same model as in Section 4.2.1, and optimization began at the same initial point ($\beta^{(0)} = 0$ and $\sigma^{(0)} = 1$). In this example, the empirical coverage of the 95% confidence intervals (for one particular regression coefficient) is indeed closer to the nominal 95% level with the correction; the effect is more pronounced for smaller sample sizes. Naturally, one can also construct corrected confidence regions given ellipsoids centered at $\theta^{(K)}$ and shaped according to equation (4.13). This construction is illustrated in Figure 4.8, where data were generated from a variation of the linear model from Section 4.2.1, in which the covariates are correlated rather than independent.



(a) (b)

Figure 4.8: (a) Empirical confidence interval coverage with noisy gradient descent estimators. The blue dots show the empirical coverage probabilities of the last iterate utilizing the noisy sandwich formula and the corrected formula. The gray points illustrate the coverage of the corrected formula computed using the first iterate whose gradient is smaller (in Euclidean norm) than the standard error of the noise added at each iteration. This gradient uses a non-private stopping rule. (b) 2D coverage region for a pair of parameters, showing the tighter confidence region resulting from our finite-sample correction to the sandwich formula. Nonprivate counterpart included on right.

Noisy Newton's Method

In Appendix D.4, the noisy Newton's method update is expressed as the sum of an ordinary Newton's method update plus a noise term. Specifically, the noise term is $\eta\tilde{N}_k$, where $\tilde{N}_k$ is an infinite sum given by equation (D.28). This noise introduces additional variance not reflected in the sandwich variance estimator in Equation 4.12. We consider only the first term in $\tilde{N}_k$, as the other terms are higher-order. Thus, truncating the error term to $\eta\left\{\nabla^2\mathcal{L}_n(\theta^{(k)})\right\}^{-1}\tilde{Z}_k$, the variance attributable to this term is $\eta^2\left\{\nabla^2\mathcal{L}_n(\theta^{(k)})\right\}^{-1}\mathrm{var}(\tilde{Z}_k)\left\{\nabla^2\mathcal{L}_n(\theta^{(k)})\right\}^{-1}$. Finally, to maintain the desired privacy guarantee, we substitute the noisy estimate of the Hessian at the current iterate for $\nabla^2\mathcal{L}_n(\theta^{(k)})$. This gives an approximate correction to the variance of the form $C_{Newton} := \eta^2\left\{\nabla^2\mathcal{L}_n(\theta^{(k)})+\tilde{W}_k\right\}^{-1}\left(\frac{2B\sqrt{2K}}{\mu n}\right)^2\left\{\nabla^2\mathcal{L}_n(\theta^{(k)})+\tilde{W}_k\right\}^{-1}$. In the case of noisy Newton's method, the variance of $\theta^{(K)}$ can be approximated by making the following correction to equation (4.12): $\hat{V}_n(\theta^{(K)}) = \tilde{V}_n(\theta^{(K)}) + nC_{Newton}$. Analogously to the gradient descent case above, this corrected variance formula yields $(1-\alpha)$-confidence intervals of the form $\mathrm{CI}^{cor}_{1-\alpha}(\theta_{0j}) = \theta^{(K)}_j \pm \sqrt{\frac{\tilde{V}_n(\theta^{(K)})_{jj}}{n}+(C_{Newton})_{jj}}z_{1-\alpha/2}$.

Figure 4.9 demonstrates the practical impact of this correction. Simulated data for this exercise was generated from the same model as in Figure 4.8. All optimizations were initialized at $\beta^{(0)} = 0$, with σ^2 assumed to be known.

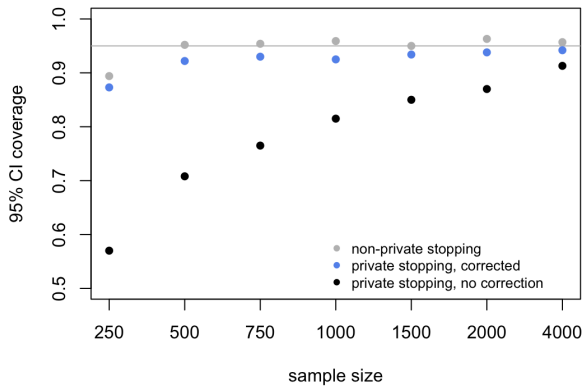


Figure 4.9: Empirical confidence interval coverage for confidence intervals constructed using noisy Newton's method estimators. The blue dots show the empirical coverage probabilities of the last iterate utilizing the noisy sandwich formula and the corrected formula, for an algorithm that terminates before the budgeted number of iterations if the noisy gradient is sufficiently small in magnitude. The gray dots illustrate the coverage of the corrected formula with a stopping rule based on the true gradient.

4.5 Discussion

This chapter introduces several tools for privacy-preserving data analysis: private versions of gradient descent and Newton's method for M-estimators, and a means to construct private confidence regions for the parameter estimates obtained from these optimization methods. Examples from this chapter and in Appendix D explore the behavior of the optimization algorithms in practice. One notable feature is that by construction, the iterates cannot approach the non-private M-estimators beyond a privacy-preserving neighborhood proportional to the noise injected at each step of the algorithms. They instead bounce around this neighborhood once it is attained.

The asymptotic variance (and asymptotic normality) of non-private M-estimators provides a basis for computing differentially private confidence regions. Since any noisy iterate has additional built-in noise that is not captured by the usual asymptotic variance, simple corrections that account for an extra noise term significantly mitigate the systematic underestimation of the variance from

the usual sandwich formula, as demonstrated in Section 4.4.2.

Supporting theory shows that the convergence rates of these private M-estimators are nearly optimal, preserving the same rates of convergence as their standard non-noisy counterparts, since they converge in the same order of iterations to a neighborhood of the solution to the non-private objective function minimization problem. This result highlights the statistical importance of (local) strong convexity, since it justifies scheduling only $O(\log n)$ noisy iterates, which in turn leads to the nearly-optimal statistical cost of privacy of the order $O\left(\frac{p \log n}{n\mu^2}\right)$. Without strong convexity, standard gradient-based algorithms are known to require $O(\sqrt{n})$ iterations to achieve an optimization error comparable to the statistical parametric rate $O\left(\sqrt{\frac{p}{n}}\right)$. This would be quite damaging for differentially private M-estimators, as they would lead to a statistical cost of $O\left(\sqrt{\frac{p}{n}}\frac{1}{\mu^2}\right)$, which would in fact be the dominating term driving the asymptotic efficiency of the resulting estimators.

An interesting extension of this work would to more deeply explore the utility of these methods on real data sets for which privacy concerns are relevant.

Bibliography

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

J. Acharya, Z. Sun, and H. Zhang. Differentially private testing of identity and closeness of discrete distributions. *Advances in Neural Information Processing Systems*, 31:6878–6891, 2018.

Muhammad Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. *Journal of Machine Learning Research*, 19(22):1–51, 2018.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

M. Avella Medina. Privacy-preserving parametric inference: A case for robust statistics. *Journal of the American Statistical Association*, 116(534):969–983, 2021.

Marco Avella-Medina, Casey Bradshaw, and Po-Ling Loh. Differentially private inference via noisy optimization. *The Annals of Statistics*, 51(5):2067–2092, 2023.

J. Awan and A. Slavković. Differentially private uniformly most powerful tests for binomial data. *Advances in Neural Information Processing Systems*, 2018:4208–4218, 2018.

Ben Axelson. These are the 45 upstate NY college campuses that reported the most rapes, ranked. *NewYorkUpstate.com*, 2019. 14 May 2019, `https://www.newyorkupstate.com/news/g66l-2019/05/b1452bae03654/these-are-the-45-ny-college-campuses-that-reported-the-most-rapes-ranked.html`. Retrieved 3 Aug 2023.

Manuel Bagues and Berta Esteve-Volart. Politicians' luck of the draw: Evidence from the Spanish Christmas lottery. *Journal of Political Economy*, 124(5):1269–1294, 2016.

T C Bailey, M S Carvalho, T M Lapa, W V Souza, and M J Brewer. Modeling of under-detection of cases in disease surveillance. *Annals of Epidemiology*, 15(5):335–343, 2005.

B. Balle, P. Kairouz, B. McMahan, O. D. Thakkar, and A. Thakurta. Privacy amplification via random check-ins. *Advances in Neural Information Processing Systems*, 33, 2020.

R. F. Barber and J. Duchi. Privacy: A few definitional aspects and consequences for minimax mean-squared error. In *53rd IEEE Conference on Decision and Control*, pages 1365–1369. IEEE, 2014.

A. F. Barrientos, J. P. Reiter, A. Machanavajjhala, and Y. Chen. Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics*, 28(2):440–453, 2019.

R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

R. Bassily, V. Feldman, K. Talwar, and A. Thakurta. Private stochastic convex optimization with optimal rates. *arXiv preprint arXiv:1908.09970*, 2019.

Bernd Beber and Alexandra Scacco. What the numbers say: A digit-based test for election fraud. *Political Analysis*, 20(2):211–234, 2012.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Luís M A Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301–7306, 2007.

George EP Box. Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 143(4):383–404, 1980.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Johannes Bracher and Leonhard Held. A marginal moment matching approach for fitting endemic-epidemic models to underreported disease surveillance counts. *Biometrics*, 77(4):1202–1214, 2021.

Kay H Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L Scott. Inferring causal impact using Bayesian structural time-series models. *THE ANNALS of APPLIED STATISTICS*, pages 247–274, 2015.

Z. Bu, J. Dong, Q. Long, and W. J. Su. Deep learning with Gaussian differential privacy. *Harvard Data Science Review*, 2020(23), 2020.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

T. T. Cai, Y. Wang, and L. Zhang. The cost of privacy in generalized linear models: Algorithms and minimax lower bounds. *arXiv preprint arXiv:2011.03900*, 2020.

T. T. Cai, Y. Wang, and L. Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *Annals of Statistics*, 49(5):2825–2850, 2021.

David Cantor, Darby M. Steiger, Reanne Townsend, John Y. Hartge, Robert E. Fay, Antonia Warren, Leanne L. Heaton, Suzanne Kaasa, Aaron Maitland, Hanyu Sun, Greg Norman, Michael Jones, Shannan Catalano, and Allen J. Beck. Methodological research to support the National Crime Victimization Survey: Self-report data on rape and sexual assault–pilot test. *Bureau of Justice Statistics*, NCJ 256011, 2021.

K. Chadha, J. Duchi, and R. Kuditipudi. Private confidence sets. In *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021.

YuSang Chang, SungSup Brian Choi, Jinsoo Lee, and Won Chang Jin. Population size vs. number of crimes: Is the relationship superlinear? *International Journal of Information Systems and Social Change (IJISSC)*, 9(1):26–39, 2018.

K. Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems*, volume 22, pages 289–296. Citeseer, 2008.

K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Raj Chetty, John N Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan. Income segregation and intergenerational mobility across colleges in the United States. *The Quarterly Journal of Economics*, 135(3):1567–1633, 2020.

Coalition of Domestic Election Observers. Final report on Ghana's 2020 presidential and parliamentary elections, 2023. `https://codeoghana.org/wp-content/uploads/2023/07/CODEOs-2020-Election-Final-Report_web.pdf`.

C. Covington, X. He, J. Honaker, and G. Kamath. Unbiased statistical estimation and valid confidence intervals under differential privacy. *arXiv preprint arXiv:2110.14465*, 2021.

Maria Cuellar. Trends in self-reporting of marijuana consumption in the United States. *Statistics and Public Policy*, 5(1):1–10, 2018.

A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.

Guilherme L de Oliveira, Juliane F Oliveira, Júlia M Pescarini, Roberto F S Andrade, Joilda S Nery, Maria Y Ichihara, Liam Smeeth, Elizabeth B Brickley, Maurício L Barreto, Gerson O Penna, et al. Estimating underreporting of leprosy in Brazil using a Bayesian approach. *PLoS Neglected Tropical Diseases*, 15(8), 2021.

Guilherme Lopes de Oliveira, Rosangela Helena Loschi, and Renato Martins Assunção. A random-censoring Poisson model for underreported data. *Statistics in medicine*, 36(30):4873–4892, 2017.

Joseph Deckert, Mikhail Myagkov, and Peter C Ordeshook. Benford's law and the detection of election fraud. *Political Analysis*, 19(3):245–268, 2011.

O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, 2014.

J. Dong, A. Roth, and W. J. Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B*, 84(1):3–37, 2022.

Nikolay Doudchenko and Guido W Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Working Paper 22791, National Bureau of Economic Research, October 2016.

Arindrajit Dube and Ben Zipperer. *Pooling multiple case studies using synthetic controls: An application to minimum wage policies*. Number 8944 in Discussion Paper Series. SSRN, 2015.

J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

Michaela Dvorzak and Helga Wagner. Sparse Bayesian modelling of underreported count data. *Statistical Modelling*, 16(1):24–46, 2016.

C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

C. Dwork, K. Talwar, A. Thakurta, and L. Zhang. Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, pages 11–20, 2014.

Peter S Fader and Bruce GS Hardie. A note on modelling underreported Poisson counts. *Journal of Applied Statistics*, 27(8):953–964, 2000.

V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: Optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

Amanda Fernández-Fontelo, Alejandra Cabaña, Harry Joe, Pedro Puig, and David Moriña. Untangling serially dependent underreported count data for gender-based violence. *Statistics in Medicine*, 38(22):4404–4422, 2019.

Bonnie S Fisher, Leah E Daigle, Francis T Cullen, and Michael G Turner. Reporting sexual victimization to the police and others: Results from a national-level study of college women. *Criminal Justice and Behavior*, 30(1):6–38, 2003.

M. Gaboardi, H. Lim, R. Rogers, and S. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *International Conference on Machine Learning*, pages 2111–2120. PMLR, 2016.

Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472, 1992.

Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6:721–741, 1984.

S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Chris Glynn, Thomas H. Byrne, and Dennis P. Culhane. Inflection points in community-level homeless rates. *The Annals of Applied Statistics*, 15(2):1037 – 1053, 2021. doi: 10.1214/20-AOAS1414.

James Goodman. State steps up efforts to make colleges more responsive to sexual assault allegations. *Democrat and Chronicle*, 2017. 10 July 2017, `https://www.democratandchronicle.com/story/news/2017/07/10/training-session-sexual-assault-puts-priority-understanding-trauma/454590001/`. Retrieved 3 Aug 2023.

Irwin Guttman. The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 29(1):83–100, 1967.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*, volume 196. John Wiley & Sons, 1986.

P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233. University of California Press, 1967.

P. J. Huber and E. Ronchetti. *Robust Statistics*. Wiley, New York, second edition, 2009.

R. Iyengar, J. P. Near, D. Song, O. Thakkar, A. Thakurta, and L. Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 299–316. IEEE, 2019.

P. Jain and A. Guha. Thakurta. (Near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, pages 476–484. PMLR, 2014.

P. Jain, P. Kothari, and A. Thakurta. Differentially private online learning. In *Conference on Learning Theory*, pages 24–1. JMLR Workshop and Conference Proceedings, 2012.

Edwin T Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4 (3):227–241, 1968.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

V. Karwa and S. Vadhan. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*, 2017.

D. Kifer, A. Smith, and A. Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. JMLR Workshop and Conference Proceedings, 2012.

Sungjin Kim, Clarence Lee, and Sachin Gupta. Bayesian synthetic control methods. *Journal of Marketing Research*, 57(5):831–852, 2020.

Peter Klimek, Yuri Yegorov, Rudolf Hanel, and Stefan Thurner. Statistical detection of systematic election irregularities. *Proceedings of the National Academy of Sciences*, 109(41):16469–16473, 2012.

Dmitry Kobak, Sergey Shpilkin, and Maxim S Pshenichnikov. Integer percentages as electoral falsification fingerprints. *The Annals of Applied Statistics*, 10(1):54–73, 2016.

Francis Kokutse. Ghana's main opposition party will contest election results. *The Associated Press*, 2020. 10 Dec 2020, `https://apnews.com/general-news-12cb47611276c43b5efc5b7a20218a1d`. Retrieved 8 Jan 2025.

Candace Kruttschnitt, William D. Kalsbeek, and Carol C. House, editors. *Estimating the incidence of rape and sexual assault*. Washington, DC: The National Academies Press, 2014.

T. Kulkarni, J. Jälkö, A. Koskela, S. Kaski, and A. Honkela. Differentially private Bayesian inference for generalized linear models. In *International Conference on Machine Learning*, pages 5838–5849. PMLR, 2021.

SSP Kumara and Hoong Chor Chin. Application of Poisson underreporting model to examine crash frequencies at signalized three-legged intersections. *Transportation Research Record*, 1908(1): 46–50, 2005.

Janet L. Lauritsen, Jennifer Gatewood Owens, Michael Planty, Michael R. Rand, and Jennifer L. Truman. Methods for counting high-frequency repeat victimizations in the National Crime Victimization Survey. *Bureau of Justice Statistics*, NCJ 237308, 2012. URL `http://www.bjs.gov/index.cfm?ty=pbdetail&iid=2240`.

J. Lee and D. Kifer. Concentrated differentially private gradient descent with adaptive per-iteration privacy budget. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1656–1665, 2018.

J. Lei. Differentially private M-estimators. *Advances in Neural Information Processing Systems*, 2011:361–369, 2011.

Fabien Letouzey, François Denis, and Rémi Gilleron. Learning from positive and unlabeled examples. In *International Conference on Algorithmic Learning Theory*, pages 71–85. Springer, 2000.

Michael S Lewis-Beck and Martin Paldam. Economic voting: An introduction. *Electoral studies*, 19(2-3):113–121, 2000.

Michael S Lewis-Beck and Mary Stegmaier. Economic determinants of electoral outcomes. *Annual Review of Political Science*, 3(1):183–219, 2000.

Jiawei Li and Jonathan H Huggins. Calibrated model criticism using split predictive checks. *arXiv preprint arXiv:2203.15897*, 2022.

Jianming Ma and Zheng Li. Bayesian modeling of frequency-severity indeterminacy with an application to traffic crashes on two-lane highways. In *ICCTP 2010: Integrated Transportation Systems: Green, Intelligent, Reliable*, pages 1022–1033. International Chinese Transportation Professionals Association, 2010.

Mahama v. Electoral Commission & Akufo-Addo. Petition. Supreme Court of Ghana, 2020.

Mahama v. Electoral Commission & Akufo-Addo. Writ No J1/05/2021. Supreme Court of Ghana, 2021.

Walter R. Mebane. Election forensics: The second-digit benford's law test and recent American presidential elections. In R. Michael Alvarez, Thad E. Hall, and Susan D. Hyde, editors, *Election Fraud: Detecting and Deterring Electoral Manipulation*, chapter 10, pages 162–181. Brookings Institution Press, 2008.

Jacob M Montgomery, Santiago Olivella, Joshua D Potter, and Brian F Crisp. An informed forensics approach to detecting vote irregularities. *Political Analysis*, 23(4):488–505, 2015.

Gemma E Moran, David M Blei, and Rajesh Ranganath. Holdout predictive checks for Bayesian model criticism. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86 (1):194–214, 2024.

Elias Moreno and Javier Giron. Estimating with incomplete count data: A Bayesian approach. *Journal of Statistical Planning and Inference*, 66(1):147–159, 1998.

Rachel E. Morgan and Grace Kena. Criminal victimization, 2016: Revised. *Bureau of Justice Statistics*, NCJ 252121, 2018.

Rachel E. Morgan and Alexandra Thompson. Criminal victimization, 2019. *Bureau of Justice Statistics*, NCJ 255113, 2020.

Rachel E. Morgan and Jennifer L. Truman. Criminal victimization, 2017. *Bureau of Justice Statistics*, NCJ 252472, 2018.

P. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Mikhail Myagkov, Peter C. Ordeshook, and Dimitri Shakin. *The Forensics of Election Fraud*, chapter 2, pages 30–70. Cambridge University Press, 2009.

Barack Obama. Memorandum – Establishing a White House task force to protect students from sexual assault. *The White House Office of the Press Secretary*, 2014. 22 Jan 2014, `https://obamawhitehouse.archives.gov/the-press-office/2014/01/22/memorandum-establishing-white-house-task-force-protect-students-sexual-a`. Retrieved 17 Aug 2023.

Xun Pang, Licheng Liu, and Yiqing Xu. A Bayesian alternative to synthetic control for comparative case studies. *Political Analysis*, 30(2):269–288, 2022.

V. Peña and A. F. Barrientos. Differentially private methods for managing model uncertainty in linear regression models. *arXiv preprint arXiv:2109.03949*, 2021.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

Stephanie Powers, Richard Gerlach, and James Stamey. Bayesian variable selection for Poisson regression with underreported responses. *Computational statistics & data analysis*, 54(12): 3289–3299, 2010.

A. Rajkumar and S. Agarwal. A differentially private stochastic gradient descent algorithm for multiparty classification. In *Artificial Intelligence and Statistics*, pages 933–941. PMLR, 2012.

P. Rigollet and J.-C. Hütter. High-dimensional statistics. *Lecture Notes for Course 18S997*, 2017.

R. Rogers and D. Kifer. A new class of private chi-square hypothesis tests. In *Artificial Intelligence and Statistics*, pages 991–1000. PMLR, 2017.

F. Roosta-Khorasani and M. W. Mahoney. Sub-sampled Newton methods. *Mathematical Programming*, 174(1):293–326, 2019.

Donald B Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, pages 1151–1172, 1984.

Chiara Sabina and Lavina Y Ho. Campus and college victim responses to sexual assault and dating violence: Disclosure, service utilization, and service provision. *Trauma, Violence, & Abuse*, 15 (3):201–226, 2014.

T. D. Savitsky, M. R. Williams, and J. Hu. Bayesian pseudo posterior mechanism under differential privacy. *arXiv preprint arXiv:1909.11796*, 2019.

Carl P Schmertmann and Marcos R Gonzaga. Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. *Demography*, 55(4):1363–1388, 2018.

Divya Shanmugam and Emma Pierson. Quantifying inequality in underreported medical conditions. *arXiv preprint arXiv:2110.04133*, 2021.

O. Sheffet. Differentially private ordinary least squares. In *International Conference on Machine Learning*, pages 3105–3114. PMLR, 2017.

Sofi Sinozich and Lynn Langton. Rape and sexual assault victimization among college-age females, 1995-2013. *Bureau of Justice Statistics*, NCJ 248471, 2014. URL `http://www.bjs.gov/index.cfm?ty=pbdetail&iid=5176`.

A. Slavkovic and R. Molinari. Perturbed M-estimation: A further investigation of robust statistics for differential privacy. *arXiv preprint arXiv:2108.08266*, 2021.

Aleksandr Sobyanin. Results of the 12 December 1993 elections and the future federal election law. *Constitutional Law: Eastern European Review*, 1:2–10, 1994.

Tanya Somanader. President Obama launches the "It's On Us" campaign to end sexual assault on campus. *The White House Blog*, 2014. 19 Sept 2014, `https://obamawhitehouse.archives.gov/blog/2014/09/19/president-obama-launches-its-us-campaign-end-sexual-assault-campus`. Retrieved 17 Aug 2023.

S. Song, K. Chaudhuri, and A. D. Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248. IEEE, 2013.

S. Song, T. Steinke, O. Thakkar, and A. Thakurta. Evading the curse of dimensionality in unconstrained private GLMs. In *International Conference on Artificial Intelligence and Statistics*, pages 2638–2646. PMLR, 2021.

Stan Development Team. RStan: the R interface to Stan, 2023. URL `https://mc-stan.org/`. R package version 2.21.8.

Oliver Stoner, Theo Economou, and Gabriela Drummond Marques da Silva. A hierarchical framework for correcting under-reporting in count data. *Journal of the American Statistical Association*, 114(528):1481–1492, 2019.

T. Sun, I. Necoara, and Q. Tran-Dinh. Composite convex optimization with global and local inexact oracles. *Computational Optimization and Applications*, pages 1–56, 2020.

K. Talwar, A. Thakurta, and L. Zhang. Nearly-optimal private LASSO. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 3025–3033, 2015.

J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Jennifer L. Truman and Heather Brotsos. Update on the NCVS instrument redesign: Operational pilot test and split sample. *Bureau of Justice Statistics*, NCJ 306051, 2021.

Jennifer L. Truman and Lynn Langton. Criminal victimization, 2013 (revised). *Bureau of Justice Statistics*, NCJ 247648, 2014.

Jennifer L. Truman and Lynn Langton. Criminal victimization, 2014. *Bureau of Justice Statistics*, NCJ 248973, 2015.

Jennifer L. Truman and Rachel E. Morgan. Criminal victimization, 2015. *Bureau of Justice Statistics*, NCJ 250180, 2016.

C. Uhler, A. Slavković, and S. E. Fienberg. Privacy-preserving data sharing for genome-wide association studies. *The Journal of Privacy and Confidentiality*, 5(1):137, 2013.

D. Vu and A. Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, pages 138–143. IEEE, 2009.

D. Wang, M. Ye, and J. Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2719–2728, 2017a.

X. Wang, S. Ma, D. Goldfarb, and W. Liu. Stochastic quasi-Newton methods for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 27(2):927–956, 2017b.

Y. Wang, S. Fienberg, and A. Smola. Privacy for free: Posterior sampling and stochastic gradient Monte Carlo. In *International Conference on Machine Learning*, pages 2493–2502. PMLR, 2015.

Y. Wang, D. Kifer, and J. Lee. Differentially private confidence intervals for empirical risk minimization. *Journal of Privacy and Confidentiality*, 9(1), 2019.

Y.-X. Wang. Revisiting differentially private linear regression: Optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1803.02596v2*, 2018.

L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

Nathan B. Wikle and Corwin M. Zigler. Causal health impacts of power plant emission controls under modeled and uncertain physical process interference. *The Annals of Applied Statistics*, 18 (4):2753 – 2774, 2024. doi: 10.1214/24-AOAS1904.

Rainer Winkelmann. Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics*, 21(4):575–587, 1996.

Kevin Wu, Dominik Dahlem, Christopher Hane, Eran Halperin, and James Zou. Collecting data when missingness is unknown: A method for improving model performance given underreporting in patient populations. In *Conference on Health, Inference, and Learning*, pages 229–242. PMLR, 2023.

F. Yu, M. Rybar, C. Uhler, and S. E. Fienberg. Differentially-private logistic regression for detecting multiple-SNP association in GWAS databases. In *International Conference on Privacy in Statistical Databases*, pages 170–184. Springer, 2014.

Mali Zhang, R Michael Alvarez, and Ines Levin. Election forensics: Using machine learning and synthetic data for possible election anomaly detection. *PloS one*, 14(10):e0223950, 2019.

Appendix A: Supplementary Material for Chapter 1

A.1 Dataset Details

The analysis in Chapter 1 pertains to postsecondary schools in the United States which grant academic degrees, reported a positive number of in-person students, and which submitted campus security reports under the Clery Act during the 2014-2019 period. A limited number of US colleges and universities are not subject to these reporting requirements.

Data for number of reported assaults was furnished by the US Department of Education Office of Postsecondary Education, Campus Safety and Security database. All other data is sourced from the Integrated Postsecondary Education Data System (IPEDS), maintained by the National Center for Education Statistics (NCES), a division of the US Department of Education.

Number of reported assaults- For each school in each year, this is the number of sexual assaults disclosed in the campus security report corresponding to that calendar year.

Student population- For each school in each year, this consists of the total number of students enrolled in the Fall, minus the number of students enrolled in 100% distance learning programs.

Percent female- For each school in each year, this is the percent of the student body enrolled in the fall who were classified as female ('male' and 'female' are currently the only gender identity options in the reporting system).

Degree of urbanization- Degree of urbanization is categorized on a 12-point scale defined by the NCES (`https://nces.ed.gov/programs/edge/docs/LOCALE_CLASSIFICATIONS.pdf`) For the purpose of the foregoing analysis, this scale was collapsed into 3 categories corresponding to 'urban', 'suburban', and 'rural'.

Pell grant recipients- For each school in each year, this is the percent of undergraduate students who received Pell grants.

Associate degree only- For each school, this is an indicator for whether the school offers only associate degrees. This equals zero for schools offering bachelor, master, or doctoral degrees.

Religious instruction- For each school, this is an indicator for whether the school is substantially engaged in religious instruction. Theological seminaries are one such example. This is a manually curated subset of schools designated as “Private not-for-profit (religious affiliation)" in IPEDS.

Modifications and exclusions

- Michigan State University is excluded from the data set due to notable unreliability of its reporting systems, for which the school was fined over $4 million by the US Department of Education in 2019.

- Northern Oklahoma College (NOC) is a community college. One of its campuses is co-located with the main Stillwater, OK campus of Oklahoma State University (OSU), a large flagship public university. Beginning in 2019, campus crime statistics for NOC include crimes on the shared OSU Stillwater campus. Consequently, these assaults are double counted in the dataset. For 2019, we attribute these assaults to OSU only and do not duplicate them for NOC.

- For Ohio State University (main campus), the reported number of assaults in 2018 and 2019 included assaults perpetrated by former Ohio State physician Richard Strauss. These assaults occurred during his 20-year employment at Ohio State, from 1978 to 1998. We exclude these assaults from our analysis (30 in 2018 and 97 in 2019), as they occurred long before the relevant reporting period (which we can be certain of, as Strauss died in 2005.)

- In 2017 University of Nebraska-Lincoln had a total of 119 reported assaults. Of those 119 reported assaults, 104 corresponded to the same victim-perpetrator pair. This type of high-frequency serial victimization is beyond the scope of what the current model can handle;

we defer this consideration to future work. Instead, we record one assault for the repeat victim/perpetrator pair.

- In 2017 Wells College had 488 students and 29 reported assaults. Of those 29 reported assaults, 28 corresponded to the same victim and perpetrator (Axelson, 2019). This type of high-frequency serial victimization is beyond the scope of what the current model can handle; we defer this consideration to future work. Instead, we record one assault for the repeat victim/perpetrator pair.

- In 2015 Genesee Community College had 20 reported assaults, all 20 of which corresponded to the same victim and perpetrator (Goodman, 2017). This type of high-frequency serial victimization is beyond the scope of what the current model can handle; we defer this consideration to future work. Instead, we record one assault for the repeat victim/perpetrator pair.

A.2 Derivations

Claim 1. $z|\lambda, p \sim Poisson(\lambda p)$

Given parameters λ and p, the true number of events z is independent of covariates $\mathbf{w}$ and $\mathbf{v}$, rendering the distribution of $z|\lambda, p$ equal to the distribution of $z|\lambda, p, \mathbf{w}, \mathbf{v}$. For simplicity of notation, the covariates are suppressed in the following.

For an arbitrary nonnegative integer k, we have:

$$\begin{aligned}
\mathbb{P}(x = k|\lambda, p) &= \sum_{j=0}^{\infty} \mathbb{P}(x = k|\lambda, p, z = j)\mathbb{P}(z = j|\lambda, p) \\
&= \sum_{j \geq k} \mathbb{P}(x = k|\lambda, p, z = j)\mathbb{P}(z = j|\lambda, p) && \text{(A.1)} \\
&= \sum_{j \geq k} \binom{j}{k} p^k (1-p)^{j-k} \frac{e^{-\lambda}\lambda^j}{j!} \\
&= \sum_{j \geq k} \frac{j!}{k!(j-k)!} p^k (1-p)^{j-k} e^{-\lambda} \frac{\lambda^j}{j!} \\
&= \sum_{j \geq k} \frac{1}{k!(j-k)!} (p\lambda)^k ((1-p)\lambda)^{j-k} e^{-\lambda} \\
&= \frac{(p\lambda)^k e^{-\lambda}}{k!} \sum_{j \geq k} \frac{((1-p)\lambda)^{j-k}}{(j-k)!} \\
&= \frac{(p\lambda)^k e^{-\lambda}}{k!} \sum_{m=0}^{\infty} \frac{((1-p)\lambda)^m}{(m)!} && \text{(A.2)} \\
&= \frac{(p\lambda)^k e^{-\lambda}}{k!} e^{(1-p)\lambda} \\
&= \frac{(p\lambda)^k}{k!} e^{-\lambda p}.
\end{aligned}$$

In the above, (A.1) holds because given $z = j$, x is supported on $\{0, 1, \ldots, j\}$. Step (A.2) utilizes a change of variables to $m \stackrel{\text{def}}{=} j - k$.

Claim 2. $u|x, \lambda, p \sim Poisson((1-p)\lambda)$

As in Section 1.3, consider decomposing the total number of assaults z into the number of reported assaults x and the number of unreported assaults u. When x is known, the randomness in z comes only from u, and the distribution of $z|x, \lambda, p$ is the distribution of $u|x, \lambda, p$ shifted to the right by the observed quantity x.

First consider the joint distribution of u and x.

$$\begin{aligned}
\mathbb{P}(u = l, x = k|\lambda, p) &= \sum_{j=0}^{\infty} \mathbb{P}(u = l, x = k|z = j, \lambda, p)\mathbb{P}(z = j|\lambda, p) \\
&= \mathbb{P}(u = l, x = k|z = l + k, \lambda, p)\mathbb{P}(z = l + k|\lambda, p) \qquad \text{(A.3)} \\
&= \binom{l+k}{k} p^k (1-p)^l \frac{e^{-\lambda}\lambda^{l+k}}{(l+k)!} \\
&= \frac{(p\lambda)^k((1-p)\lambda)^l}{k!l!} e^{-\lambda} \qquad \text{(A.4)}
\end{aligned}$$

Step (A.3) holds because $\mathbb{P}(u = l, x = k|z, \lambda, p)$ equals zero for any value of z such that $z \neq l + k$.

Then,

$$\mathbb{P}(u = l|x = k, \lambda, p) = \frac{\mathbb{P}(u = l, x = k|\lambda, p)}{\sum_{m=0}^{\infty} \mathbb{P}(u = m, x = k|\lambda, p)}.$$

Substituting in the expression from step (A.4), we have:

$$\begin{aligned}
\mathbb{P}(u = l|x = k, \lambda, p) &= \frac{(p\lambda)^k((1-p)\lambda)^l e^{-\lambda}/(k!l!)}{\sum_{m=0}^{\infty}(p\lambda)^k((1-p)\lambda)^m e^{-\lambda}/(m!k!)} \\
&= \frac{((1-p)\lambda)^l/l!}{\sum_{m=0}^{\infty}((1-p)\lambda)^m/m!} \\
&= \frac{((1-p)\lambda)^l/l!}{e^{(1-p)\lambda}} \\
&= \frac{((1-p)\lambda)^l e^{-(1-p)\lambda}}{l!}.
\end{aligned}$$

A.3 Pooling in Hierarchical Models

In the model in Section 1.3, coefficients α and β are global, encouraging the sharing of information across schools, while school-level terms γ and ε capture local properties of individual schools. This model structure is a compromise between complete pooling, in which no school-level differences are permitted (γ and ε omitted), and no pooling, in which each school is modeled separately from all others, with its own coefficient vectors α_i and β_i.

In the case of complete pooling or no pooling, the observed counts still arise from a binomial

distribution,

$$x_{ij}|z_{ij}, p_{ij} \sim \text{Binom}(z_{ij}, p_{ij}),$$

but the true number of assaults and probability of reporting are constructed differently.

A.3.1 Complete Pooling

The complete pooling scenario assumes that sexual assault reporting patterns do not systematically differ across schools, except to the extent accounted for by their covariates.

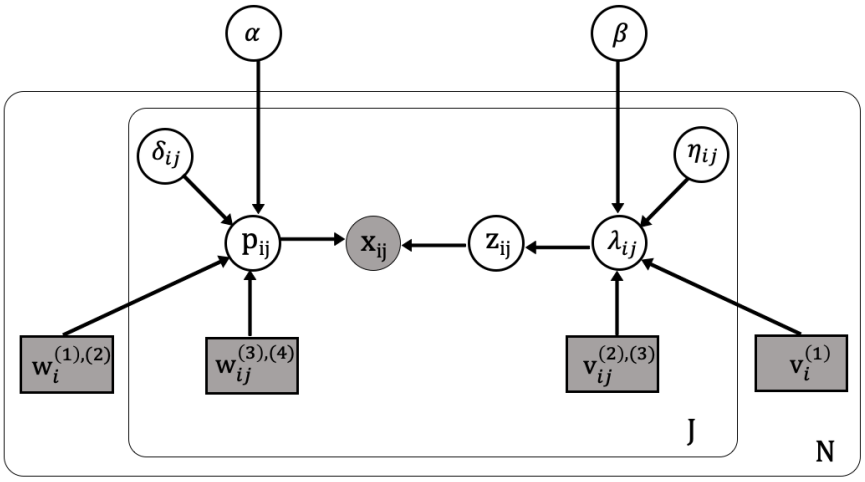


Figure A.1: Complete Pooling Model Diagram

The true number of assaults, z, follows a model similar to that of Section 1.3, but with the school-level offset ε_i omitted from Equation 1.3, that is:

$$\log(\lambda_{ij}) = \beta_{0,\mathrm{v}_i^{(1)}} + \beta_1 \mathrm{v}_{ij}^{(2)} + \beta_2 \mathrm{v}_{ij}^{(3)} + \eta_{ij}.$$

Likewise, the probability of reporting is modeled as in Section 1.3, but with the school-level offset γ_i omitted from Equation 1.5, that is:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_0 + \alpha_1 \mathrm{w}_i^{(1)} + \alpha_2 \mathrm{w}_i^{(2)} + \alpha_3 \mathrm{w}_{ij}^{(3)} + \alpha_4 \mathrm{w}_{ij}^{(4)} + \delta_{ij}.$$

A.3.2 No Pooling

In the no-pooling scenario, each school is modeled separately.

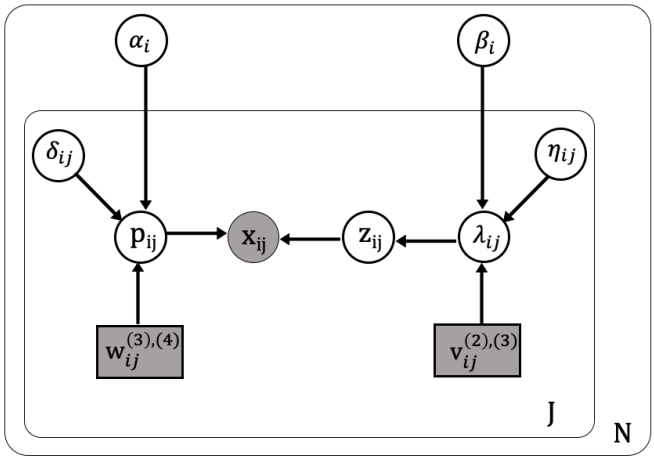


Figure A.2: No-Pooling Model Diagram

This model has a separate set of coefficients $\beta_{0,i}, \beta_{1,i}, \beta_{2,i}$ for each school i. Each model has a single intercept rather than separate intercepts for urban, suburban, and rural locations, as each school has a single location and does not allow for this distinction. The school-level offset ε_i is omitted; when modeling a single school, the dataset does not provide any information that distinguishes ε_i from the intercept $\beta_{0,i}$. When modeling z, we thus replace Equation 1.3 from the original partial pooling model with:

$$\log(\lambda_{ij}) = \beta_{0,i} + \beta_{1,i} v_{ij}^{(2)} + \beta_{2,i} v_{ij}^{(3)} + \eta_{ij}.$$

Each school i has a separate set of coefficients $\alpha_{0,i}, \alpha_{3,i}, \alpha_{4,i}$. As above, the school-level offset γ_i is omitted, as the dataset does not distinguish it from the intercept $\alpha_{0,i}$. Coefficients α_1 and α_2 are omitted. These coefficients correspond whether a school is a junior college, and whether a school is primarily engaged in religious instruction, respectively. As these are fixed characteristics of a school, the data for a single school do not provide information to help distinguish these coefficients from the intercept $\alpha_{0,i}$. When modeling z, we thus replace Equation 1.5 from the original partial

pooling model with:

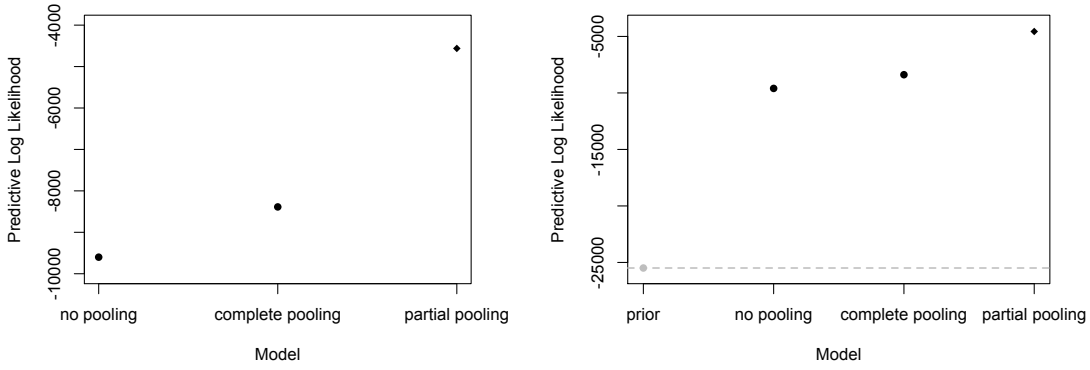
$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_{0,i} + \alpha_{3,i}\mathrm{w}_{ij}^{(3)} + \alpha_{4,i}\mathrm{w}_{ij}^{(4)} + \delta_{ij}.$$

A.3.3 Comparison

Roughly 20% of records in the campus sexual assault dataset were held out, and posterior inference for the partial pooling, complete pooling, and no pooling models was carried out on the remaining 80% of records. To compare the three models, we approximate the predictive likelihood on the held-out data. Denoting a model's collection of latent variables as $\boldsymbol{\theta}$ and suppressing covariates, this quantity is

$$p(x^{held-out}|x^{train}) = \int p(x^{held-out}|\boldsymbol{\theta})p(\boldsymbol{\theta}|x^{train})d\boldsymbol{\theta}.$$

Figure A.3 shows that the partial pooling model assigns the highest likelihood to the held-out data. This result that favors the partial pooling model over the two alternatives, as it suggests the partial pooling model does a better job approximating the true distribution of the data.



(a) (b)

Figure A.3: The partial pooling model assigns higher likelihood to held-out observational data; (a) Estimated predictive log likelihood of the held-out dataset under each model (b) Left-hand side plot rescaled to include log likelihood of the held-out dataset under the prior from the partial pooling model

A.4 Sensitivity to Correlated Reporting

The model in Section 1.3 assumes that, for a given school in a given year, each assault is reported or not reported independently of all others. In practice, this assumption could plausibly be violated. To explore sensitivity to violations of this assumption, we perform posterior inference on synthetic data with non-independent reporting decisions.

For each school i in year j, fix the true number of assaults z_{ij} and the reporting probability p_{ij}. Given z_{ij} and p_{ij}, generate a reported number of assaults x_{ij} under different assumptions:

- **Independent:** For each assault at school i in year j, assume the reporting decision is independent with probability p_{ij} of reporting.
- **ρ=0.05:** Assume that each assault at school i in year j has marginal probability p_{ij} of being reported, but that the reporting decisions are correlated with correlation coefficient $\rho = 0.05$, rather than being independent. Simulate this by sampling z_{ij}-many correlated Bernoulli variables, each with probability p_{ij}, and taking their sum to get x_{ij}.
- **ρ=0.10:** Same as above, but with correlation coefficient $\rho = 0.10$.
- **Pairwise:** Split the total number of assaults at school i in year j into pairs, and assume that each **pair** of assaults is reported independently with probability p_{ij}. The marginal reporting probability is still p_{ij} for each assault. This creates correlation in the reporting decisions, but unlike the previous scenario, the amount of correlation varies (schools with a large number of assaults will be less affected by this coarsening of the reported numbers, compared to schools where the true number of assaults is low).

To assess the impact of correlated reporting decisions, we obtain posterior estimates of the true number of assaults for each record in the dataset under each data-generating scenario. For each data-generating scenario, we regress the true number of assaults z on the posterior mean estimate $\hat{z} \triangleq \mathbb{E}[z|x]$. For the dataset simulated with independent reporting decisions, the slope of

this regression line is 1.00, indicating that the posterior means of the number of assaults tend to be proportional to the true counts z_{ij}. Figure A.4 depicts that the three scenarios with correlated reporting decisions produce slopes less than 1, indicating that the model tends to overestimate the true number of assaults when reporting decisions are positively correlated. This form of model misspecification can significantly degrade the model's ability to recover the true number of assaults.

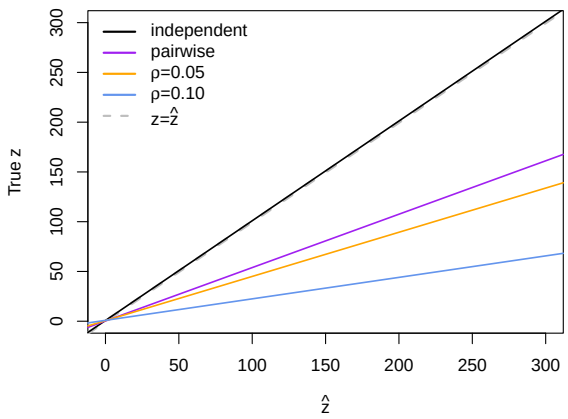


Figure A.4: When reporting decisions within each school are positively correlated, the model tends to overestimate the true underlying number of assaults.

A.5 Sensitivity to Prior Specification

While the National Crime Victimization Survey reflects a significant effort to measure the true extent of sexual assault victimization in the United States, it is nevertheless true that survey respondents may choose not to disclose their sexual assault victimization to interviewers. Consequently, it is worthwhile to examine how a change in the priors informed by NCVS estimates affects the modeling results presented in Section 1.4. Below we consider a range of alternative scenarios.

(a) Original prior specification presented in Sections 1.3.1 and 1.3.2 (Figure 1.4 Figure 1.6).

(b) True reporting rates are approximately 10% lower than NCVS estimates. Prior distributions are adjusted such that the prior median reporting rate is approximately 19.7%, rather than the

22% presented in Section 1.3.2 (Figure 1.6). Prior distribution on total incidence of assaults is increased correspondingly.

(c) True reporting rates are approximately 25% lower than estimated by NCVS, with corresponding increase in incidence of assaults.

(d) True reporting rates are approximately 50% lower than estimated by NCVS, with corresponding increase in incidence of assaults.

(e) True reporting rates are approximately 75% lower than estimated by NCVS, with corresponding increase in incidence of assaults.

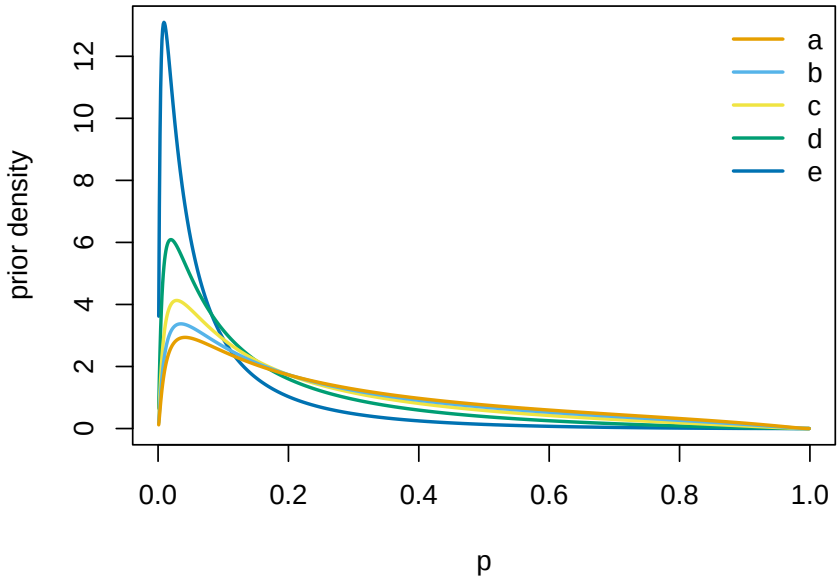
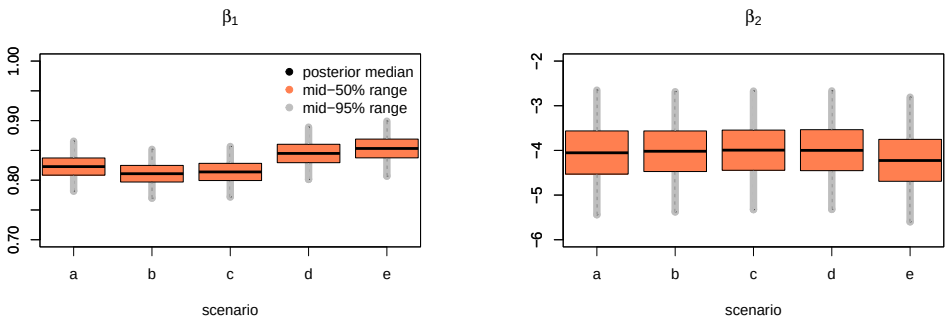


Figure A.5: Prior distributions on the reporting probability p under scenarios (a) - (e) (with model coefficients α_1, α_2, α_3, and α_4 set to zero).

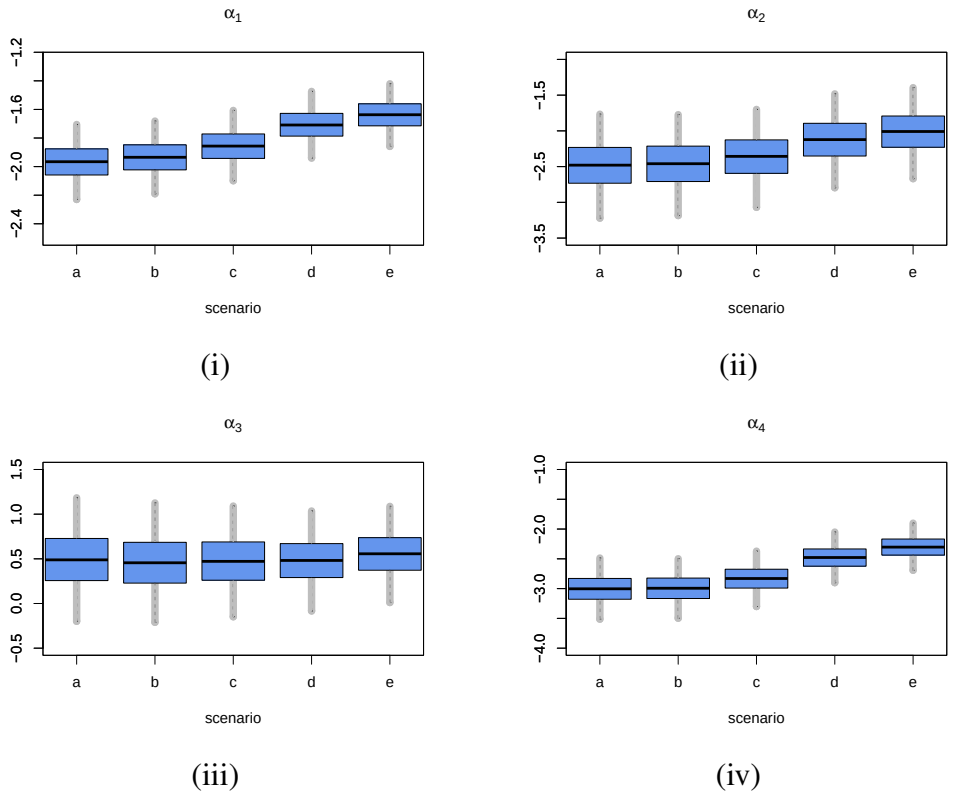
In line with the discussion in Section 1.3.5, we find that inferences about the role of covariates are relatively more stable across scenarios (as depicted in Figures A.6 and A.7), while inferences about the true incidence and reporting rate exhibit more sensitivity. As shown in Figure A.8,

prior belief in a higher incidence of sexual assault produces higher posterior estimates of the total number of assaults occurring. Likewise, Figure A.9 demonstrates that prior belief in a lower probability of reporting sexual assault produces lower posterior estimates of the reporting rate. However, zooming in on scenario (e) (which represents the most extreme change to the priors) in Figure A.10, results still suggest that an increase in reporting rates is a more plausible explanation for the increase in reported assaults over 2014-2018, as opposed to an increase in the true incidence of assault. Note that material variability in NCVS survey respondents' truthfulness over time could call this result into question as well; interested readers may find discussion of a similar issue in the context of marijuana consumption surveys in Cuellar (2018).



(i) (ii)

Figure A.6: Coefficients relating covariates to the total number of assaults do not display much practical difference under moderate changes to the prior distributions envisioned in scenarios a-e. Panel (i) depicts β_1, the coefficient corresponding to student population. Panel (ii) depicts β_2, the coefficient corresponding to gender imbalance of student body.



(i) (ii)

(iii) (iv)

Figure A.7: Estimates of coefficients relating covariates to the reporting probability remain similar under moderate changes to the prior distributions, and begin to shrink toward zero under more extreme changes to the prior distribution. Panel (i) depicts α_1, the coefficient corresponding to whether a school is a junior college. Panel (ii) depicts α_2, the coefficient corresponding to whether a school is substantially engaged in religious instruction. Panel (iii) depicts α_3, the coefficient corresponding to the proportion of a school's student body composed of women. Panel (iv) depicts α_4, the coefficient corresponding to the proportion of a school's student body receiving Pell grants.

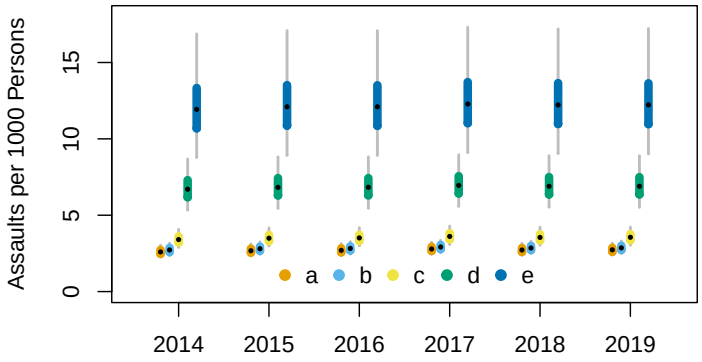


Figure A.8: Prior belief in a higher incidence of sexual assault produces higher posterior estimates of the total number of assaults occurring.

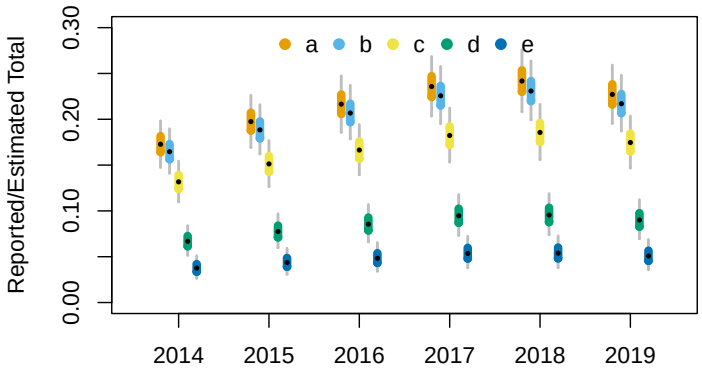
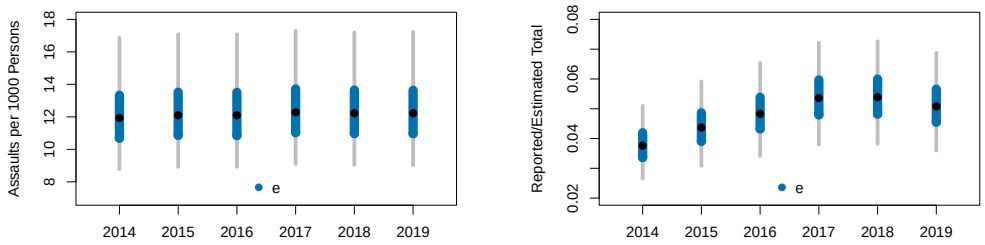


Figure A.9: Prior belief in a lower probability of reporting sexual assault produces lower posterior estimates of the reporting rate.



(i) (ii)

Figure A.10: Under scenario (e), posterior estimates of the incidence rate do not exhibit a clear increasing or decreasing trend, while posterior estimates of the reporting rate are more suggestive of an increase over 2014-2018. Panel (i) depicts the posterior distribution of the incidence rate; panel (ii) depicts the posterior distribution of the reporting rate.

Given the sensitivity to prior information, we positively note that in 2011 the Bureau of Justice Statistics convened a panel of statisticians, sociologists, and criminal justice experts to review its methodology and recommend optimal practices for measuring sexual assault (Kruttschnitt et al., 2014). The panel's recommendations led to the development and pilot testing of new survey approaches to measuring sexual assault within the NCVS context (Cantor et al., 2021). These findings fed into a broader multi-year redesign of the overall survey. After further field testing to assess the impact of proposed changes, the redesigned survey is being phased in beginning in 2024 (further detail is available in Truman and Brotsos (2021)).

Appendix B: Supplementary Material for Chapter 2

The following 61 parliamentary constituencies in Ghana were flagged by the randomization test in Chapter 2, Section 2.2. Among polling places in each of these constituencies, the observed correlation between turnout rate and incumbent share of the vote was significantly higher than would be expected if those two quantities were truly independent of each other.

Constituency	Region	Rank	Correlation
MANHYIA NORTH	ASHANTI	1 (tie)	0.79
BOSOME FREHO	ASHANTI	1 (tie)	0.64
NSUTA/KWAMANG/BEPOSO	ASHANTI	1 (tie)	0.61
MANHYIA SOUTH	ASHANTI	1 (tie)	0.56
SUBIN	ASHANTI	1 (tie)	0.54
MANSO ADUBIA	ASHANTI	1 (tie)	0.54
ASOKWA	ASHANTI	1 (tie)	0.52
JUABEN	ASHANTI	1 (tie)	0.49
NHYIAESO	ASHANTI	1 (tie)	0.48
SUAME	ASHANTI	1 (tie)	0.48
EJISU	ASHANTI	1 (tie)	0.46
KWADASO MUNICIPAL	ASHANTI	1 (tie)	0.44
ODOTOBRI	ASHANTI	1 (tie)	0.43
AFIGYA KWABRE SOUTH	ASHANTI	1 (tie)	0.42
ASANTE AKIM NORTH	ASHANTI	1 (tie)	0.41
OFORIKROM	ASHANTI	1 (tie)	0.40
BANTAMA	ASHANTI	1 (tie)	0.38
ATWIMA KWANWOMA	ASHANTI	1 (tie)	0.37
KWABRE EAST	ASHANTI	1 (tie)	0.34
WENCHI	BONO	1 (tie)	0.48
KINTAMPO SOUTH	BONO EAST	1 (tie)	0.44
ABUAKWA SOUTH	EASTERN	1 (tie)	0.53
ATIWA WEST	EASTERN	1 (tie)	0.52
KADE	EASTERN	1 (tie)	0.36
DOME/KWABENYA	GREATER ACCRA	1 (tie)	0.33
KPANDAI	NORTHERN	1 (tie)	0.39

Constituency	Region	Rank	Correlation
KRACHI WEST	OTI	1 (tie)	0.45
SALAGA SOUTH	SAVANNAH	1 (tie)	0.43
TARKWA NSUAEM	WESTERN	1 (tie)	0.43
EVALUE AJOMORO GWIRA	WESTERN	1 (tie)	0.42
AMENFI EAST	WESTERN	1 (tie)	0.37
PRESTEA HUNI-VALLEY	WESTERN	1 (tie)	0.32
OLD TAFO	ASHANTI	33 (tie)	0.39
MAMPONG	ASHANTI	33 (tie)	0.34
PRU WEST	BONO EAST	33 (tie)	0.50
BAWKU CENTRAL	UPPER EAST	33 (tie)	0.36
KETU NORTH	VOLTA	37	0.37
AFIGYA KWABRE NORTH	ASHANTI	38 (tie)	0.46
AKROPONG	EASTERN	38 (tie)	0.33
SAWLA-TUNA-KALBA	SAVANNAH	40 (tie)	0.34
CHIANA-PAGA	UPPER EAST	40 (tie)	0.37
AFIGYA SEKYERE EAST	ASHANTI	42 (tie)	0.31
BEKWAI	ASHANTI	42 (tie)	0.28
SISSALA WEST	UPPER WEST	42 (tie)	0.42
SEKYERE AFRAM PLAINS	ASHANTI	45	0.47
NAVRONGO CENTRAL	UPPER EAST	46	0.32
BOSOMTWE	ASHANTI	47	0.24
TALENSI	UPPER EAST	48	0.33
YENDI	NORTHERN	49 (tie)	0.27
AGOTIME ZIOPE	VOLTA	49 (tie)	0.38
YAGABA/ KUBORI	NORTH EAST	51	0.41
NKORANZA NORTH	BONO EAST	52	0.33
FANTEAKWA SOUTH	EASTERN	53	0.34
BOLE-BAMBOI	SAVANNAH	54	0.28
ZEBILLA	UPPER EAST	55	0.25
AWUTU SENYA WEST	CENTRAL	56	0.21
BIAKOYE	OTI	57	0.25
NKORANZA SOUTH	BONO EAST	58	0.24
TECHIMAN SOUTH	BONO EAST	59	0.18
SENE WEST	BONO EAST	60	0.35
TEMA CENTRAL	GREATER ACCRA	61	0.24

Appendix C: Supplementary Material for Chapter 3

The synthetic data used in Chapter 3, Section 3.3 was derived from the cigarette consumption data studied in Abadie et al. (2010). We began with the annual per-capita cigarette pack sales for 16 states over 30 years, during which time none of the 16 states implemented significant legislative measures to reduce cigarette consumption. The 16 states included were CO, CT, GA, IA, KY, LA, ME, MO, MT, NV, SD, TX, UT, VA, VT, and WI. Singular value thresholding was applied to the matrix of cigarette consumption data to produce a lower-rank approximation of the original data, referred to as $\mathbf{Z}$ in Section 3.3. Random noise terms $\varepsilon_{jt} \overset{iid}{\sim} N(0, 3^2)$ were then added to the entries of the singular value-thresholded matrix to produce the synthetic data, referred to as $\mathbf{Y}$ in Section 3.3.

The weights on each of the 16 states used to construct the artificial target unit were as follows:

State	coefficient
CO	0.00
CT	0.28
GA	0.00
IA	0.03
KY	0.05
LA	0.15
ME	0.00
MO	0.00
MT	0.00
NV	0.00
SD	0.14
TX	0.09
UT	0.00
VA	0.25
VT	0.01
WI	0.00

Appendix D: Supplementary Material for Chapter 4

D.1 Concentration Inequalities

The following two lemmas will be instrumental in the analysis of the differentially private algorithms proposed in Chapter 4.

Lemma 2. *Let $X \in \mathbb{R}^d$ be a sub-Gaussian random vector with variance proxy σ^2. For any $\alpha > 0$, with probability at least $1 - \alpha$,*

$$\|X\|_2 \leq 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\alpha)}.$$

Proof. This result can be found in (Rigollet and Hütter, 2017, Theorem 1.19). □

Lemma 3. *Let W be a symmetric $p \times p$ random matrix whose upper triangular elements, including the diagonal, are i.i.d. $N(0,1)$. For any $\alpha > 0$, with probability at least $1 - \alpha$,*

$$\|W\|_2 \leq \sqrt{2p\log(2p/\alpha)}.$$

Proof. Letting E_{jk} denote the matrix which has the value 1 in the $(j,k)^{\text{th}}$ position and 0 everywhere else, we see that

$$W = \sum_{1 \leq j \leq k \leq p} Z_{jk}(E_{jk} + E_{kj} - \mathbb{1}_{\{j=k\}}E_{jj}),$$

where $\{Z_{jk}\}_{1 \leq j,k \leq p}$ is an i.i.d. sequence of standard normal variables. It follows from (Tropp, 2015, Theorem 4.1.1) that with probability at least $1 - \alpha$,

$$\|W\|_2 \leq \sqrt{2v(Z)\log(2p/\alpha)},$$

where

$$v(Z) = \left\| \sum_{1 \le j \le k \le p} (E_{jk} + E_{kj} - \mathbb{1}_{\{j=k\}} E_{jj})^2 \right\|_2 = \|pI\|_2 = p.$$

□

D.2 Convex Analysis

The following results are standard, and proofs can be found in Boyd and Vandenberghe (2004) or Bubeck (2015).

Lemma 4. *Suppose $f : \mathbb{R}^p \to \mathbb{R}$ is strongly convex with parameter τ_1, i.e.*

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \tau_1 \|x - y\|_2^2, \qquad \forall x, y \in \mathbb{R}^p.$$

Then

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \ge 2\tau_1 \|x - y\|_2^2, \qquad \forall x, y \in \mathbb{R}^p,$$

and if f is twice-differentiable, then

$$\nabla^2 f(x) \ge 2\tau_1 I, \qquad \forall x \in \mathbb{R}^p.$$

Lemma 5. *Suppose $f : \mathbb{R}^p \to \mathbb{R}$ is convex and τ_2-smooth, i.e.,*

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \tau_2 \|x - y\|_2^2, \qquad \forall x, y \in \mathbb{R}^p.$$

Then

$$\frac{1}{2\tau_2} \|\nabla f(x) - \nabla f(y)\|_2^2 \le \langle \nabla f(x) - \nabla f(y), x - y \rangle, \qquad \forall x, y \in \mathbb{R}^p,$$

and

$$\|\nabla f(x) - \nabla f(y)\|_2 \le 2\tau_2 \|x - y\|_2, \qquad \forall x, y \in \mathbb{R}^p.$$

If f is twice-differentiable, then

$$\nabla^2 f(x) \preceq 2\tau_2 I, \qquad \forall x \in \mathbb{R}^p.$$

D.3 Proofs for Noisy Gradient Descent

Proofs due to authors Avella Medina and Loh (Avella-Medina et al., 2023). Included here for completeness.

This section contains proofs of the two main results on the convergence of noisy gradient descent, Theorem 2 and Proposition 1.

D.3.1 Proof of Theorem 2

We first present the main argument, followed by statements and proofs of supporting lemmas.

Main Argument

It is easy to see that the Gaussian mechanism and post-processing guarantee that every iteration of the algorithm is $\frac{\mu}{\sqrt{K}}$-GDP: Fixing $\theta^{(k)}$, the global sensitivity for the gradient iterates $\theta^{(k+1)}$ in equation (4.3) is clearly bounded by $\frac{2\eta B}{n}$, and then we simply apply the Gaussian mechanism (Theorem 1) to obtain the noisy gradient descent iterates (4.4). It follows from Corollary 2 in Dong et al. (2022) that the entire algorithm is μ-GDP after K iterations. This proves (i).

Let us now turn to the proof of (ii). We denote $N_k = \frac{B\sqrt{K}}{\mu n} Z_k$, so the noisy gradient updates can be rewritten as $\theta^{(k+1)} = \theta^{(k)} - \eta \nabla \mathcal{L}_n(\theta^{(k)}) + \eta N_k$. From Lemma 2 and a union bound, we have that with probability at least $1 - \xi$,

$$\|N_k\|_2 \leq \frac{\{4\sqrt{p} + 2\sqrt{2\log(K/\xi)}\} B\sqrt{K}}{\mu n} := r_{priv}, \tag{D.1}$$

for all $k < K$. For the remainder of the argument, we will assume the bound (D.1) holds, and prove

that the desired conditions hold deterministically.

From Lemmas 7 and 8 and the local strong convexity assumption, we have

$$\begin{aligned}\|\theta^{(k)} - \hat{\theta}\|_2^2 &\leq \frac{1}{\tau_1}\kappa^k \Delta_0 + \frac{3rr_{priv}}{2(1-\kappa)\tau_1} \\ &\leq \frac{3rr_{priv}}{(1-\kappa)\tau_1},\end{aligned} \tag{D.2}$$

where $r_{priv} = \frac{\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{\mu n}$ and the last inequality holds as long as $k \geq k_0 = \lceil k_0' \rceil$, where

$$k_0' = \frac{\log(1/\Delta_0) + \log\left(\frac{3r\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{2(1-\kappa)\mu n}\right)}{\log(\kappa)}.$$

We note that the bound (D.2) is looser than the desired result. The rest of our argument will improve this estimate for later iterates $k > K$ (with $K > k_0$) by refining the argument of Lemma 8, leveraging inequality (D.2).

As derived in inequality (D.20) in the proof of Lemma 8, we can show that

$$\Delta_{k+1} \leq (1-\gamma)\Delta_k + \|N_k\|_2 \left(\|\hat{\theta} - \theta^{(k)}\|_2 + \|\theta^{(k+1)} - \theta^{(k)}\|_2\right), \tag{D.3}$$

where $\Delta_k = \mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta})$ and $\gamma = 2\eta\tau_1$. Furthermore, note that smoothness, inequality (D.2), and $k \geq k_0$ imply that

$$\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2 = \|\nabla\mathcal{L}_n(\theta^{(k)}) - \nabla\mathcal{L}_n(\hat{\theta})\|_2 \leq 2\tau_2\|\theta^{(k)} - \hat{\theta}\|_2 \leq 2\tau_2\sqrt{\frac{3rr_{priv}}{(1-\kappa)\tau_1}},$$

Thus, the triangle inequality and inequality (D.1) give

$$\|\theta^{(k+1)} - \theta^{(k)}\|_2 = \eta\|\nabla\mathcal{L}_n(\theta^{(k)}) + N_k\|_2 \leq \eta r_{priv} + 2\tau_2\eta\sqrt{\frac{3r}{(1-\kappa)\tau_1}}r_{priv}^{1/2}. \tag{D.4}$$

Consequently, from inequalities (D.2), (D.3), and (D.4) and the fact that $k \geq k_0$, we obtain

$$\Delta_{k+1} \leq (1-\gamma)\Delta_k + r_{priv}\left(\eta r_{priv} + (2\tau_2\eta + 1)\sqrt{\frac{3r}{(1-\kappa)\tau_1}} r_{priv}^{1/2}\right).$$

Recalling that $\kappa = 1 - \gamma$, we have that for $j = k_0 + 1, \ldots, K$,

$$\begin{aligned}\Delta_{j+k_0} &\leq \kappa\Delta_{j-1+k_0} + (2\tau_2\eta + 1)\sqrt{\frac{3r}{(1-\kappa)\tau_1}} r_{priv}^{3/2} + \eta r_{priv}^2 \\ &\leq \kappa^j \Delta_{k_0} + \left(2\tau_2\eta + \frac{3}{2}\right)\sqrt{\frac{3r}{(1-\kappa)\tau_1}} r_{priv}^{3/2}, \end{aligned} \tag{D.5}$$

where the last inequality used the fact that $n \geq \frac{\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{\mu \cdot \frac{1}{4\eta^2}\frac{3r}{(1-\kappa)\tau_1}} \iff \eta r_{priv}^{1/2} \leq \frac{1}{2}\sqrt{\frac{3r}{(1-\kappa)\tau_1}}$. Taking $k = j + k_0$, local strong convexity and inequality (D.5) show that

$$\begin{aligned}\|\theta^{(k)} - \hat{\theta}\|_2^2 &\leq \frac{1}{\tau_1}\kappa^j \Delta_{k_0} + \left(2\tau_2\eta + \frac{3}{2}\right)\sqrt{\frac{3r}{(1-\kappa)\tau_1^3}} r_{priv}^{3/2} \\ &\leq (2\tau_2\eta + 2)\sqrt{\frac{3r}{(1-\kappa)\tau_1^3}} r_{priv}^{3/2}, \end{aligned} \tag{D.6}$$

where the last inequality holds as long as $k \geq k_1 + k_0$ with $k_1 = \lceil k_1' \rceil$ and

$$k_1' = \frac{\log(1/\Delta_{k_0}) + \log\left(\frac{1}{2}\sqrt{\frac{3r}{(1-\kappa)\tau_1}} r_{priv}^{3/2}\right)}{\log(\kappa)}.$$

Thus, we see $k_1 + k_0$ iterations of the algorithm improve the convergence rate of the iterates from $O\left(r_{priv}^{1/2}\right)$, obtained in (D.2) after k_0 iterations, to $O\left(r_{priv}^{3/4}\right)$. Repeating the same argument m times, we will show that successive iterates of the algorithm in fact improve the estimation error bound successively as $O\left(r_{priv}^{1/2}\right), O\left(r_{priv}^{3/4}\right), O\left(r_{priv}^{7/8}\right), \ldots, O\left(r_{priv}^{1-1/2^m}\right)$. This will be enough to prove the desired result, since we can take $m = \log_2 n$, which gives the rate $O\left(r_{priv}^{1-1/n}\right)$.

Letting $C_0 = \sqrt{\frac{3r}{(1-\kappa)\tau_1}}$, we see that inequality (D.2) says that for $k > k_0$, we have $\|\theta^{(k)} - \hat{\theta}\|_2 \leq$

$C_0 r_{priv}^{1/2}$. We then saw that for $k > k_1 + k_0$ this bound can be further refined to

$$\|\theta^{(k)} - \hat{\theta}\|_2 \leq \sqrt{\frac{2\tau_2\eta + 2}{\tau_1} C_0} r_{priv}^{3/4} := C_1 r_{priv}^{3/4}.$$

The same arguments used to obtain inequality (D.5) show that for $j = k_0 + k_1 + 1, \ldots, K$, we have

$$\begin{aligned} \Delta_{j+k_0+k_1} &\leq \kappa \Delta_{j-1+k_0} + (2\tau_2\eta + 1) C_1 r_{priv}^{1+3/4} + \eta r_{priv}^2 \\ &\leq \kappa^{j+k_1} \Delta_{k_0} + \left(2\tau_2\eta + \frac{3}{2}\right) C_1 r_{priv}^{7/4}, \end{aligned} \tag{D.7}$$

where the last inequality uses $n \geq \frac{\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{\mu \frac{1}{2^4} C_1^4} \eta^4$, which is implied by the minimum sample size assumption $n \geq \frac{\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{\mu \cdot \frac{1}{2^2} C_0^2} \eta^2$ used to show inequality (D.5), since $\frac{1}{2^2\eta^2} C_0^2 < \frac{1}{2^4\eta^4} C_1^4$. Similar to inequality (D.6), taking $k = j + k_0 + k_1$, local strong convexity and inequality (D.7) show that

$$\begin{aligned} \|\theta^{(k)} - \hat{\theta}\|_2^2 &\leq \frac{1}{\tau_1} \kappa^j \Delta_{k_0+k_1} + \frac{2\tau_2\eta + 3/2}{\tau_1} C_1 r_{priv}^{7/4} \\ &\leq \frac{2\tau_2\eta + 2}{\tau_1} C_1 r_{priv}^{7/4}, \end{aligned} \tag{D.8}$$

where the last inequality holds as long as $k \geq k_2 + k_1 + k_0$ with $k_2 = \lceil k_2' \rceil$ and

$$k_2' = \frac{\log(1/\Delta_{k_0+k_1}) + \log\left((2\tau_2\eta + 2) C_1 r_{priv}^{7/4}\right)}{\log(\kappa)}.$$

We conclude from inequality (D.8) that for $k > k_0 + k_1 + k_2$ we have

$$\|\theta^{(k)} - \hat{\theta}\|_2 \leq \sqrt{\frac{2\tau_2\eta + 2}{\tau_1} C_1} r_{priv}^{7/8} = C_2 r_{priv}^{1-1/2^3}.$$

Iterating this argument, a tedious but straightforward calculation shows that we can obtain the

following recurrence: Let $C_0 = \sqrt{\frac{3r}{(1-\kappa)\tau_1}}$, and define $C_m := \sqrt{\frac{2\tau_2\eta+2}{\tau_1}C_{m-1}}$, $k_m := \lceil k'_m \rceil$, and

$$k'_m = \frac{\log(1/\Delta_{\sum_{j=0}^{m-1} k_j}) + \log\left((2\tau_2\eta+2)C_{m-1}r_{priv}^{2-1/2^m}\right)}{\log(\kappa)}.$$

Note that we also have

$$C_m = \left(\frac{2\tau_2\eta+2}{\tau_1}\right)^{\sum_{j=1}^{m} 1/2^j} C_0^{1/2^m} = \left(\frac{2\tau_2\eta+2}{\tau_1}\right)^{1-1/2^m} C_0^{1/2^m}. \tag{D.9}$$

Then, taking $k > \sum_{j=0}^{m} k_j$ and $m = \log_2 n - 1$, we have

$$\|\theta^{(k)} - \hat{\theta}\|_2 \leq C_m r_{priv}^{1-1/2^{m+1}} = \left(\frac{2\tau_2\eta+2}{\tau_1}\right)^{1-2/n} C_0^{2/n} r_{priv}^{1-1/n} \leq C r_{priv}, \tag{D.10}$$

where C is some positive constant. We note that (D.10) implicitly used the fact that the minimum sample size requirement $n \geq \frac{\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{\mu\cdot\frac{1}{(2\eta)^2}C_0^2}$ also implies that $n \geq \frac{\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{\mu\cdot\frac{1}{(2\eta)^{2^{m+1}}}C_{m-1}^{2^m}}$ when $\eta \leq \frac{1}{2}$. To see this, it suffices to check that $C_0^2 \leq \frac{1}{(2\eta)^{2^m}}C_{m-1}^{2^m}$. The latter holds true since inequality (D.9) and $\tau_2 \geq \tau_1$ show that $C_{m-1} \geq (2\eta)^{1-1/2^{m-1}} C_0^{1/2^{m-1}}$. This last inequality implies the desired inequality, since clearly,

$$C_0^2 \leq \frac{1}{(2\eta)^{2^m}}\left((2\eta)^{1-1/2^{m-1}} C_0^{1/2^{m-1}}\right)^{2^m} = \frac{1}{(2\eta)^2}C_0^2.$$

This completes the proof.

Supporting Lemmas

Lemma 6. *Suppose $\mathcal{L}_n$ is convex in $\Theta \subseteq \mathbb{R}^p$ and locally τ_1-strongly convex in $\mathcal{B}_r(\theta_0)$. If $\hat{\theta} \in \mathcal{B}_{r/2}(\theta_0)$ and $\mathcal{L}_n(\theta) - \mathcal{L}_n(\hat{\theta}) \leq \frac{r^2}{4}\tau_1$, then $\|\theta - \hat{\theta}\|_2 \leq \frac{r}{2}$ and $\theta \in \mathcal{B}_r(\theta_0)$.*

Proof. Define

$$\Delta := \inf_{\theta \notin \mathcal{B}_{r/2}(\hat{\theta})} \mathcal{L}_n(\theta) - \mathcal{L}_n(\hat{\theta}).$$

By convexity of $\mathcal{L}_n$, the infimum must be achieved at some point θ_r^* on the boundary of the ball $\mathcal{B}_{r/2}(\hat{\theta})$. Therefore, for any parameter θ such that $\mathcal{L}_n(\theta) - \mathcal{L}_n(\hat{\theta}) < \Delta$, we must have $\|\theta - \hat{\theta}\|_2 \leq \frac{r}{2}$, and hence also $\theta \in \mathcal{B}_r(\theta_0)$.

We now claim that $\Delta \geq \frac{r^2}{4}\tau_1$, from which the desired result follows. By the triangle inequality, we have $\theta_r^* \in \mathcal{B}_r(\theta_0)$. Thus, by strong convexity,

$$\Delta = \mathcal{L}_n(\theta_r^*) - \mathcal{L}_n(\hat{\theta}) \geq \tau_1 \|\theta_r^* - \hat{\theta}\|_2^2 = \frac{r^2}{4}\tau_1,$$

as claimed. □

Lemma 7. *Suppose $\mathcal{L}_n(\theta)$ is twice-differentiable almost everywhere in $\mathcal{B}_r(\theta_0)$ and satisfies LSC and strong smoothness with parameters τ_1 and $\tau_2 \leq \frac{1}{2\eta}$. Also suppose the sample size satisfies $n = \Omega\left(\frac{\sqrt{K\log(K/\xi)}}{\mu}\right)$ and inequality (D.1) holds, and suppose $\hat{\theta} \in \mathcal{B}_{r/2}(\theta_0)$ and $\mathcal{L}_n(\theta^{(0)}) \leq \mathcal{L}_n(\hat{\theta}) + \tau_1\frac{r^2}{4}$.*

Then

$$\mathcal{L}_n(\theta^{(k)}) \leq \mathcal{L}_n(\hat{\theta}) + \tau_1\frac{r^2}{4} \quad \text{and} \quad \|\theta^{(k)} - \hat{\theta}\|_2 \leq \frac{r}{2}, \quad \forall k \leq K.$$

Proof. We will induct on k. Note that by Lemma 6, we have $\|\theta^{(0)} - \hat{\theta}\|_2 \leq \frac{r}{2}$.

For the inductive step, suppose $\mathcal{L}_n(\theta^{(k)}) \leq \mathcal{L}_n(\hat{\theta}) + \tau_1\frac{r^2}{4}$ and $\|\theta^{(k)} - \hat{\theta}\|_2 \leq \frac{r}{2}$ for some $0 \leq k < n^\alpha$. Since $\mathcal{L}_n$ is almost everywhere twice-differentiable, we have

$$\mathcal{L}_n(\theta - \eta\Delta) = \mathcal{L}_n(\theta) - \eta\langle\nabla\mathcal{L}_n(\theta), \Delta\rangle + \frac{\eta^2}{2}\langle\int_0^1 \nabla^2\mathcal{L}_n(\theta - t\eta\Delta)\mathrm{d}t\Delta, \Delta\rangle, \tag{D.11}$$

for any $\Delta \in \mathbb{R}^p$. Recalling that $N_k = \frac{B\sqrt{K}}{\mu n}Z_k$ and $\theta^{(k+1)} = \theta^{(k)} - \eta\nabla\mathcal{L}_n(\theta^{(k)}) + \eta N_k$, applying

equation (D.11) with $\theta = \theta^{(k)}$ and $\Delta = \nabla\mathcal{L}_n(\theta^{(k)}) - N_k$ leads to

$$\begin{aligned}\mathcal{L}_n(\theta^{(k+1)}) &= \mathcal{L}_n(\theta^{(k)}) - \eta\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2 + \eta\langle\nabla\mathcal{L}_n(\theta^{(k)}), N_k\rangle \\ &\quad + \frac{\eta^2}{2}\|\nabla\mathcal{L}_n(\theta^{(k)})\|_{H_k}^2 - \eta^2\langle N_k, H_k\nabla\mathcal{L}_n(\theta^{(k)})\rangle + \frac{\eta^2}{2}\|N_k\|_{H_k}^2 \\ &\leq \mathcal{L}_n(\theta^{(k)}) - \eta\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2 + \eta\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2\|N_k\|_2 \\ &\quad + \frac{\eta^2}{2}\|\nabla\mathcal{L}_n(\theta^{(k)})\|_{H_k}^2 + \eta^2\|N_k\|_2\|H_k\|_2\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2 + \frac{\eta^2}{2}\|N_k\|_{H_k}^2, \end{aligned} \tag{D.12}$$

where $H_k = \int_0^1 \nabla^2\mathcal{L}_n\left(\theta^{(k)} - t\eta\nabla\mathcal{L}_n(\theta^{(k)}) + t\eta N_k\right)\mathrm{d}t$. By inequality (D.12) and the Cauchy-Schwarz inequality, we can guarantee that $\mathcal{L}_n(\theta^{(k+1)}) \leq \mathcal{L}_n(\theta^{(k)})$ holds if the following inequality is satisfied:

$$\|\nabla\mathcal{L}_n(\theta^{(k)})\|_{I-\frac{\eta}{2}H_k}^2 \geq \|\nabla\mathcal{L}_n(\theta^{(k)})\|_2\|N_k\|_2 + \eta\|N_k\|_2\|H_k\|_2\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2 + \frac{\eta}{2}\|N_k\|_2^2\|H_k\|_2. \tag{D.13}$$

The descent condition $\mathcal{L}_n(\theta^{(k+1)}) \leq \mathcal{L}_n(\theta^{(k)})$, together with the inductive hypothesis, would then imply that

$$\mathcal{L}_n(\theta^{(k+1)}) - \mathcal{L}_n(\hat{\theta}) \leq \mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta}) \leq \tau_1\frac{r^2}{4},$$

so from Lemma 6, we could conclude that $\theta^{(k+1)} \in \mathcal{B}_{r/2}(\hat{\theta})$, as wanted.

Returning to inequality (D.13), note that $\|\nabla\mathcal{L}_n(\theta^{(k)})\|_{I-\frac{\eta}{2}H_k}^2 \geq \|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2(1-\eta\tau_2)$, since by τ_2-smoothness, we have $\lambda_{\max}(H_k) \leq 2\tau_2$. Using this bound on the left-hand side of inequality (D.13) and the upper bounds $\sup_\theta \|\nabla\mathcal{L}_n(\theta)\|_2 \leq B$ and $\lambda_{\max}(H_k) \leq 2\tau_2$ on the right-hand side of inequality (D.13), we have the stronger sufficient condition

$$\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2(1-\eta\tau_2) \geq \|N_k\|_2 B\,(1+2\eta\tau_2) + \eta\tau_2\|N_k\|_2^2,$$

which is guaranteed to be satisfied when

$$\|N_k\| \leq \frac{-B(1+2\eta\tau_2) + \sqrt{B^2(1+2\eta\tau_2)^2 + 4\eta\tau_2(1-\eta\tau_2)\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2}}{2\eta\tau_2},$$

or equivalently,

$$\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \geq \sqrt{\frac{(2\eta\tau_2\|N_k\|_2 + B(1+2\eta\tau_2))^2 - B^2(1+2\eta\tau_2)^2}{4\eta\tau_2(1-\eta\tau_2)}}$$
$$= \sqrt{\frac{\eta\tau_2\|N_k\|_2^2 + B(1+2\eta\tau_2)\|N_k\|_2}{1-\eta\tau_2}}. \tag{D.14}$$

Combining inequality (D.14) with the upper bound (D.1) on $\|N_k\|_2$, we see that inequality (D.13) holds with high probability, provided

$$\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \geq \sqrt{\frac{\eta\tau_2 r_{priv}^2 + B(1+2\eta\tau_2) r_{priv}}{1-\eta\tau_2}} = \bar{r}_{priv}. \tag{D.15}$$

We now argue that if $\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 < \bar{r}_{priv}$, we still have $\|\theta^{(k+1)} - \hat{\theta}\|_2 \leq \frac{r}{2}$ and $\mathcal{L}_n(\theta^{(k+1)}) \leq \mathcal{L}_n(\hat{\theta}) + \tau_1 \frac{r^2}{4}$. Indeed, since $\theta^{(k)} \in \mathcal{B}_r(\theta_0)$ by assumption, local strong convexity implies that

$$\tau_1\|\theta^{(k)} - \hat{\theta}\|_2^2 \leq \mathcal{L}_n(\hat{\theta}) - \mathcal{L}_n(\theta^{(k)}) - \langle \nabla \mathcal{L}_n(\theta^{(k)}), \hat{\theta} - \theta^{(k)} \rangle \leq \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \|\theta^{(k)} - \hat{\theta}\|_2,$$

so

$$\bar{r}_{priv} > \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \geq \tau_1\|\theta^{(k)} - \hat{\theta}\|_2.$$

Hence, by the triangle inequality and inequality (D.1), we have

$$\begin{aligned}
\|\theta^{(k+1)} - \hat{\theta}\|_2 &\leq \|\theta^{(k)} - \hat{\theta}\|_2 + \|\theta^{(k+1)} - \theta^{(k)}\|_2 \\
&\leq \frac{\bar{r}_{priv}}{\tau_1} + \eta\|\nabla \mathcal{L}_n(\theta^{(k)}) + N_k\|_2 \\
&\leq \left(\frac{1}{\tau_1} + \eta\right)\bar{r}_{priv} + \eta r_{priv} \\
&\leq \left(\frac{\tau_1}{\tau_2}\right)^{1/2} \frac{r}{2} \leq \frac{r}{2}
\end{aligned} \tag{D.16}$$

with high probability, when n is sufficiently large. Consequently, $\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 < \bar{r}_{priv}$ also

implies that $\theta^{(k+1)} \in \mathcal{B}_{r/2}(\hat{\theta})$; furthermore, by smoothness, we also have

$$\mathcal{L}_n(\theta^{(k+1)}) - \mathcal{L}_n(\hat{\theta}) \leq \tau_2 \|\theta^{(k+1)} - \hat{\theta}\|_2^2 \leq \tau_2 \cdot \frac{\tau_1}{\tau_2}\frac{r^2}{4} \leq \tau_1 \frac{r^2}{4}.$$

Note that the penultimate inequality in the chain (D.16) holds under the sample size condition $n = \Omega\left(\frac{\sqrt{K\log(K/\xi)}}{\mu}\right)$.

Indeed, $\eta r_{priv} \leq \frac{r}{4}\sqrt{\frac{\tau_1}{\tau_2}}$ if $n \geq \eta \frac{4\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{r\mu\sqrt{\tau_1/\tau_2}}$ and $\left(\frac{1}{\tau_1}+\eta\right)\bar{r}_{priv} \leq \frac{r}{4}\sqrt{\frac{\tau_1}{\tau_2}}$ if

$$r_{priv} \leq \frac{\frac{-B(1+2\eta\tau_2)}{1-\eta\tau_2} + \sqrt{\frac{B^2(1+2\eta\tau_2)^2}{(1-\eta\tau_2)^2} + \frac{r^2}{4}\frac{\eta\tau_1}{(1-\eta\tau_2)(\frac{1}{\tau_1}+\eta)^2}}}{\frac{2\eta\tau_2}{1-\eta\tau_2}}.$$

□

In fact, as the following result shows, the assumptions in Lemma 7 also imply that the sub-optimality gap of successive iterates decreases at a geometric rate (up to an error term).

Lemma 8. *Let $\Delta_k = \mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta})$, and suppose the conditions of Lemma 7 hold. Then*

$$\Delta_k = \kappa^k \Delta_0 + \frac{3r\{4\sqrt{p}+2\sqrt{2\log(K/\xi)}\}B\sqrt{K}}{2(1-\kappa)\mu n}, \quad \forall k \leq K,$$

where $\kappa = 1 - 2\eta\tau_1 \in [0,1)$.

Proof. Note that $\theta^{(k+1)}$ can be viewed as the minimizer of

$$Q_k(\theta) := \mathcal{L}_n(\theta^{(k)}) + \langle \nabla\mathcal{L}_n(\theta^{(k)}) - N_k, \theta - \theta^{(k)}\rangle + \frac{1}{2\eta}\|\theta - \theta^{(k)}\|_2^2.$$

Denote $\theta_\gamma = \gamma\hat{\theta} + (1-\gamma)\theta^{(k)}$, for a parameter $\gamma \in (0,1)$ to be chosen later. By optimality of $\theta^{(k+1)}$

and local strong convexity, we have

$$\begin{aligned} Q_k(\theta^{(k+1)}) &\leq Q_k(\theta_\gamma) \\ &= \mathcal{L}_n(\theta^{(k)}) + \gamma\langle \nabla\mathcal{L}_n(\theta^{(k)}) - N_k, \hat{\theta} - \theta^{(k)}\rangle + \frac{\gamma^2}{2\eta}\|\theta^{(k)} - \hat{\theta}\|_2^2 \\ &\leq \mathcal{L}_n(\theta^{(k)}) - \gamma\left(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta})\right) + \left(\frac{\gamma^2}{2\eta} - \tau_1\right)\|\theta^{(k)} - \hat{\theta}\|_2^2 - \gamma\langle N_k, \hat{\theta} - \theta^{(k)}\rangle \\ &= \mathcal{L}_n(\theta^{(k)}) - \gamma\Delta_k + \left(\frac{\gamma^2}{2\eta} - \tau_1\right)\|\theta^{(k)} - \hat{\theta}\|_2^2 - \gamma\langle N_k, \hat{\theta} - \theta^{(k)}\rangle. \end{aligned} \tag{D.17}$$

Furthermore, by local smoothness and the fact that $\tau_1 \leq \tau_2 \leq \frac{1}{2\eta}$, we have

$$\begin{aligned} \mathcal{L}_n(\theta^{(k+1)}) &\leq \mathcal{L}_n(\theta^{(k)}) + \langle \nabla\mathcal{L}_n(\theta^{(k)}), \theta^{(k+1)} - \theta^{(k)}\rangle + \tau_2\|\theta^{(k+1)} - \theta^{(k)}\|_2^2 \\ &\leq Q_k(\theta^{(k+1)}) + \langle N_k, \theta^{(k+1)} - \theta^{(k)}\rangle. \end{aligned} \tag{D.18}$$

Combining inequalities (D.17) and (D.18), we obtain

$$\mathcal{L}_n(\theta^{(k+1)}) \leq \mathcal{L}_n(\theta^{(k)}) - \gamma\Delta_k + \left(\frac{\gamma^2}{2\eta} - \tau_1\right)\|\theta^{(k)} - \hat{\theta}\|_2^2 - \gamma\langle N_k, \hat{\theta} - \theta^{(k)}\rangle + \langle N_k, \theta^{(k+1)} - \theta^{(k)}\rangle. \tag{D.19}$$

Furthermore, subtracting $\mathcal{L}_n(\hat{\theta})$ from both sides of inequality (D.19), we see that local strong convexity, taking $\gamma = 2\eta\tau_1$ so that $\gamma^2 \leq 2\eta\tau_1$, and using the bound (D.1) implies that

$$\begin{aligned} \Delta_{k+1} &\leq \Delta_k(1-\gamma) + \left(\frac{\gamma^2}{2\eta} - \tau_1\right)\|\theta^{(k)} - \hat{\theta}\|_2^2 - \gamma\langle N_k, \hat{\theta} - \theta^{(k)}\rangle + \langle N_k, \theta^{(k+1)} - \theta^{(k)}\rangle \\ &\leq (1-\gamma)\Delta_k - \gamma\langle N_k, \hat{\theta} - \theta^{(k)}\rangle + \langle N_k, \theta^{(k+1)} - \theta^{(k)}\rangle \\ &\leq (1-\gamma)\Delta_k + \|N_k\|_2\left(\|\hat{\theta} - \theta^{(k)}\|_2 + \|\theta^{(k+1)} - \theta^{(k)}\|_2\right) \\ &\leq (1-\gamma)\Delta_k + r_{priv}\frac{3r}{2}, \end{aligned} \tag{D.20}$$

where we have used the conclusion of Lemma 7 in the last inequality. Iterating the bound (D.20)

and writing $\kappa = 1 - \gamma$, we see that

$$\begin{aligned}\Delta_k &\leq \kappa^k \Delta_0 + \frac{3 r_{priv} r}{2}(1 + \kappa + \kappa^2 + \cdots + \kappa^{k-1}) \\ &\leq \kappa^k \Delta_0 + \frac{3 r_{priv} r}{2(1-\kappa)}.\end{aligned}$$

This shows the desired result. □

D.3.2 Proof of Proposition 1

We again begin by presenting the main argument, followed by statements and proofs of supporting lemmas.

Main Argument

Let $N_k = \frac{B\sqrt{K}}{\mu n} Z_k$. By Lemma 2 and a union bound, we have

$$\max_{k < K_0} \|N_k\|_2 \leq \frac{\{4\sqrt{p} + 2\sqrt{2\log(K_0/\xi_0)}\} B\sqrt{K_0}}{\mu n} := r_{priv}, \tag{D.21}$$

with probability at least $1 - \xi_0$. We will assume this inequality holds for the rest of the proof, and argue that the desired conditions hold deterministically.

Recall the form of the updates (4.4). By strong smoothness and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathcal{L}_n(\theta^{(k+1)}) &\leq \mathcal{L}_n(\theta^{(k)}) - \eta \langle \nabla \mathcal{L}_n(\theta^{(k)}), \nabla \mathcal{L}_n(\theta^{(k)}) - N_k \rangle + \tau_2 \eta^2 \|\nabla \mathcal{L}_n(\theta^{(k)}) - N_k\|_2^2 \\ &\leq \mathcal{L}_n(\theta^{(k)}) - \eta\,(1 - \tau_2 \eta)\,\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2 + \eta\,(1 + 2\tau_2 \eta)\,\|N_2\|_2 \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \\ &\quad + \tau_2 \eta^2 \|N_k\|_2^2.\end{aligned}$$

Using the bound (D.21) and the fact that $\eta \leq \frac{1}{2\tau_2}$, we then have

$$\mathcal{L}_n(\theta^{(k+1)}) \leq \mathcal{L}_n(\theta^{(k)}) - \frac{\eta}{2}\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2 + Cr_{priv}, \tag{D.22}$$

for all $k \leq K_0$,

where $C \geq 2\eta B + \frac{\eta}{2}r_{priv}$.

Furthermore, convexity implies that

$$\mathcal{L}_n(\theta^{(k)}) \leq \mathcal{L}_n(\hat{\theta}) + \langle\nabla\mathcal{L}_n(\theta^{(k)}), \theta^{(k)} - \hat{\theta}\rangle. \tag{D.23}$$

Combining inequalities (D.22) and (D.23), we obtain

$$\mathcal{L}_n(\theta^{(k+1)}) \leq \mathcal{L}_n(\hat{\theta}) + \langle\nabla\mathcal{L}_n(\theta^{(k)}), \theta^{(k)} - \hat{\theta}\rangle - \frac{\eta}{2}\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2 + Cr_{priv}.$$

This in turn shows that

$$\begin{aligned}
\mathcal{L}_n(\theta^{(k+1)}) - \mathcal{L}_n(\hat{\theta}) &\leq \frac{1}{2\eta}\left(2\eta\langle\nabla\mathcal{L}_n(\theta^{(k)}), \theta^{(k)} - \hat{\theta}\rangle - \eta^2\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2^2\right) + Cr_{priv} \\
&= \frac{1}{2\eta}\left(\|\theta^{(k)} - \hat{\theta}\|_2^2 - \|\theta^{(k)} - \eta\nabla\mathcal{L}_n(\theta^{(k)}) - \hat{\theta}\|_2^2\right) + Cr_{priv} \\
&= \frac{1}{2\eta}\left(\|\theta^{(k)} - \hat{\theta}\|_2^2 - \|\theta^{(k+1)} - \hat{\theta} - \eta N_k\|_2^2\right) + Cr_{priv} \\
&= \frac{1}{2\eta}\left(\|\theta^{(k)} - \hat{\theta}\|_2^2 - \|\theta^{(k+1)} - \hat{\theta}\|_2^2 - \eta^2\|N_k\|_2^2 + 2\eta\langle N_k, \theta^{(k+1)} - \hat{\theta}\rangle\right) + Cr_{priv} \\
&\leq \frac{1}{2\eta}\left(\|\theta^{(k)} - \hat{\theta}\|_2^2 - \|\theta^{(k+1)} - \hat{\theta}\|_2^2\right) + \|N_k\|_2\|\theta^{(k+1)} - \hat{\theta}\|_2 + Cr_{priv}.
\end{aligned}$$

Summing over k, we obtain

$$\begin{aligned}
\sum_{k=1}^{K_0}\left(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta})\right) &\leq \frac{1}{2\eta}\left(\|\theta^{(0)} - \hat{\theta}\|_2^2 - \|\theta^{(K_0)} - \hat{\theta}\|_2^2\right) + K_0 \max_{k<K_0}\|N_k\|_2\|\theta^{(k+1)} - \hat{\theta}\|_2 \\
&\quad + K_0 Cr_{priv} \\
&\leq \frac{1}{2\eta}\|\theta^{(0)} - \hat{\theta}\|_2^2 + K_0 r_{priv}(\|\theta^{(0)} - \hat{\theta}\|_2 + 1) + K_0 Cr_{priv},
\end{aligned} \tag{D.24}$$

where the last inequality is a consequence of Lemma 9.

By inequality (D.22), we have

$$\mathcal{L}_n(\theta^{(K_0)}) \leq \mathcal{L}_n(\theta^{(k)}) + CK_0 r_{priv}$$

for all $k < K_0$, so

$$K_0 \left(\mathcal{L}_n(\theta^{(K_0)}) - \mathcal{L}_n(\hat{\theta}) \right) \leq \sum_{k=1}^{K_0} \left(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta}) \right) + CK_0^2 r_{priv}.$$

Combined with inequality (D.24), this gives

$$\begin{aligned}
\mathcal{L}_n(\theta^{(K_0)}) - \mathcal{L}_n(\hat{\theta}) &\leq \frac{1}{K_0} \sum_{k=1}^{K_0} \left(\mathcal{L}_n(\theta^{(k)}) - \mathcal{L}_n(\hat{\theta}) \right) + CK_0 r_{priv} \\
&\leq \frac{1}{K_0} \frac{1}{2\eta} \|\theta^{(0)} - \hat{\theta}\|_2^2 + r_{priv}(\|\theta^{(0)} - \hat{\theta}\|_2 + 1) + (K_0 + 1) C r_{priv} \\
&\leq \frac{1}{K_0} \frac{1}{2\eta} R^2 + (R + C(K_0 + 1) + 1) r_{priv} \\
&\leq \Delta,
\end{aligned}$$

where we have taken $K_0 = \frac{R^2}{\eta\Delta}$ and

$$n \geq \frac{2(R + C(K_0 + 1) + 1)\{4\sqrt{p} + 2\sqrt{2\log(K_0/\xi_0)}\} B\sqrt{K_0}}{\Delta\mu}.$$

This last inequality establishes the desired result.

Supporting lemmas

Lemma 9. *Suppose $\mathcal{L}_n$ is convex and τ_2-smooth and let $\theta^{(k)}$ denote the updates (4.4), with $\eta \leq \frac{1}{2\tau_2}$. Let $\|\theta^{(0)} - \hat{\theta}\|_2 = R$. Also suppose inequality (D.21) holds and $(K_0 C' + 2(K_0 - 1)\eta)\, r_{priv} \leq 1$, where*

$$C' = 2\eta R + 2\eta^2 (B + 1).$$

Then

$$\|\theta^{(k+1)} - \hat{\theta}\|_2^2 \leq R^2 + \left((k+1)C' r_{priv} + 2k\eta\right) r_{priv} \leq R^2 + 1,$$

for all $k < K_0$.

Proof. We first note that since $\mathcal{L}_n$ is convex and τ_2-smooth, we have

$$\langle \nabla \mathcal{L}_n(\theta_1) - \nabla \mathcal{L}_n(\theta_2), \theta_1 - \theta_2 \rangle \geq \frac{1}{2\tau_2} \|\nabla \mathcal{L}_n(\theta_1) - \nabla \mathcal{L}_n(\theta_2)\|_2^2, \quad \forall \theta_1, \theta_2 \in \Theta \subseteq \mathbb{R}^p. \tag{D.25}$$

Applying inequality (D.25) to the pair $(\theta^{(k)}, \hat{\theta})$, we then have

$$\frac{1}{2\tau_2} \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2 \leq \langle \nabla \mathcal{L}_n(\theta^{(k)}), \theta^{(k)} - \hat{\theta} \rangle.$$

Then using the Cauchy-Schwarz inequality and the fact that $\eta \leq \frac{1}{2\tau_2}$, we obtain

$$\begin{aligned}
\|\theta^{(k+1)} - \hat{\theta}\|_2^2 &= \|\theta^{(k)} - \eta \nabla \mathcal{L}_n(\theta^{(k)}) + \eta N_k - \hat{\theta}\|_2^2 \\
&= \|\theta^{(k)} - \hat{\theta}\|_2^2 - 2\eta \langle \theta^{(k)} - \hat{\theta}, \nabla \mathcal{L}_n(\theta^{(k)}) - N_k \rangle + \eta^2 \|\nabla \mathcal{L}_n(\theta^{(k)}) - N_k\|_2^2 \\
&\leq \|\theta^{(k)} - \hat{\theta}\|_2^2 - \frac{\eta}{\tau_2} \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2 + 2\eta \|\theta^{(k)} - \hat{\theta}\| \|N_k\|_2 + \eta^2 \|\nabla \mathcal{L}_n(\theta^{(k)}) - N_k\|_2^2 \\
&\leq \|\theta^{(k)} - \hat{\theta}\|_2^2 - \frac{\eta}{\tau_2} \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2 + 2\eta \|\theta^{(k)} - \hat{\theta}\| \|N_k\|_2 \\
&\quad + \eta^2 \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2 + 2\eta^2 \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \|N_k\|_2 + \eta^2 \|N_k\|_2^2 \\
&\leq \|\theta^{(k)} - \hat{\theta}\|_2^2 + 2\eta \|\theta^{(k)} - \hat{\theta}\|_2 \|N_k\|_2 - \frac{\eta}{2\tau_2} \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2 + 2\eta^2 B \|N_k\|_2 + \eta^2 \|N_k\|_2^2.
\end{aligned} \tag{D.26}$$

It is easy to see from the condition (D.21) and inequality (D.26) that

$$\|\theta^{(1)} - \hat{\theta}\|_2^2 \leq R^2 + 2\eta R r_{priv} + 2\eta^2 B r_{priv} + \eta^2 r_{priv}^2 \leq R^2 + C' r_{priv}, \tag{D.27}$$

where $C' = 2\eta R + 2\eta^2(B+1) \geq 2\eta R + 2\eta^2 B + \eta^2 r_{priv}$.

Furthermore, inequalities (D.26) and (D.27) also imply that

$$\begin{aligned}
\|\theta^{(2)} - \hat{\theta}\|_2^2 &\leq \|\theta^{(1)} - \hat{\theta}\|_2^2 + 2\eta\|\theta^{(1)} - \hat{\theta}\|_2 r_{priv} + 2\eta^2 B r_{priv} + \eta^2 r_{priv}^2 \\
&\leq (R^2 + C' r_{priv}) + 2\eta r_{priv}(R + \sqrt{C' r_{priv}}) + 2\eta^2 B r_{priv} + \eta^2 r_{priv}^2 \\
&\leq R^2 + 2C' r_{priv} + 2\eta r_{priv}\sqrt{C' r_{priv}} \\
&\leq R^2 + (2C' + 2\eta) r_{priv},
\end{aligned}$$

where the last inequality uses the fact that $C' r_{priv} \leq 1$. A similar argument shows that

$$\begin{aligned}
\|\theta^{(3)} - \hat{\theta}\|_2^2 &\leq \|\theta^{(2)} - \hat{\theta}\|_2^2 + 2\eta\|\theta^{(2)} - \hat{\theta}\|_2 r_{priv} + 2\eta^2 B r_{priv} + \eta^2 r_{priv}^2 \\
&\leq (R^2 + 2C' r_{priv}) + 2\eta r_{priv} + 2\eta r_{priv}(R + \sqrt{(2C' + 2\eta) r_{priv}}) + 2\eta^2 B r_{priv} + \eta^2 r_{priv}^2 \\
&\leq R^2 + 3C' r_{priv} + 2\eta r_{priv}(1 + \sqrt{(2C' + 2\eta) r_{priv}}) \\
&\leq R^2 + 3C' r_{priv} + 4\eta r_{priv},
\end{aligned}$$

where the last inequality uses the assumption $(2C' + 2\eta) r_{priv} \leq 1$. More generally, one can recursively verify that as long as $((k+1)C' + 2k\eta)\, r_{priv} \leq 1$, we also have

$$\|\theta^{(k+1)} - \hat{\theta}\|_2^2 \leq R^2 + \left((k+1)C' r_{priv} + 2k\eta\right) r_{priv}.$$

□

D.4 Proofs for Noisy Newton's Method

Proofs due to authors Avella Medina and Loh (Avella-Medina et al., 2023). Included here for completeness.

This section contains a proof of Theorem 3. It begins with the main argument, followed by statements and proofs of supporting lemmas.

Main Argument

The arguments for part (i) are very similar to those presented for Theorem 2. In particular, every iteration of the algorithm is $\frac{\mu}{\sqrt{K}}$-GDP by the Gaussian mechanism, post-processing, and the composition theorem of (Dong et al., 2022, Corollary 1). The same composition result shows that the whole algorithm is μ-GDP after K iterations. This proves (i).

We now turn to the proof of (ii). Let $\tilde{W}_k = \frac{2\bar{B}\sqrt{2K}}{\mu n} W_k$ and $\tilde{Z}_k = \frac{2B\sqrt{2K}}{\mu n} Z_k$, and note that Lemmas 2 and 3 and a union bound imply that with probability at least $1-\xi$, we have $\max_{k<K} \|Z_k\|_2 \leq 4\sqrt{p}+2\sqrt{2\log(2K/\xi)}$ and $\max_{k<K} \|W_k\|_2 \leq \sqrt{2p\log(4Kp/\xi)}$. We will assume that these inequalities hold and argue deterministically for the rest of the proof.

For a sufficiently large n the Neumann series formula is well defined and leads to the identity

$$\left\{\nabla^2 \mathcal{L}_n(\theta^{(k)}) + \tilde{W}_k\right\}^{-1} = \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1} \sum_{j=0}^{\infty} \left[-\tilde{W}_k \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1}\right]^j.$$

Hence,

$$\begin{aligned}
\theta^{(k+1)} &= \theta^{(k)} - \eta \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1} \sum_{j=0}^{\infty} \left[-\tilde{W}_k \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1}\right]^j \left\{\nabla \mathcal{L}_n(\theta^{(k)}) + \tilde{Z}_k\right\} \\
&= \theta^{(k)} - \eta \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1} \nabla \mathcal{L}_n(\theta^{(k)}) + \eta \tilde{N}_k,
\end{aligned}$$

where

$$\tilde{N}_k = \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1} \left(\tilde{Z}_k + \sum_{j=1}^{\infty} \left[-\tilde{W}_k \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1}\right]^j \left\{\nabla \mathcal{L}_n(\theta^{(k)}) + \tilde{Z}_k\right\} \right). \tag{D.28}$$

Lemma 10 is our main workhorse. The bound (D.30), together with the starting value condition $\|\nabla \mathcal{L}_n(\theta^{(0)})\|_2 \leq \frac{\tau_1^2}{L}$, implies that

$$\frac{L}{2\tau_1^2} \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 < \left(\frac{L}{2\tau_1^2} \|\nabla \mathcal{L}_n(\theta^{(0)})\|_2 \right)^{2^k} + 3C\tilde{r}_{priv}, \tag{D.29}$$

for all $k \leq K$, $\tilde{r}_{priv} := \frac{\sqrt{Kp\log(Kp/\xi)}}{\mu n}$. Indeed, we can prove inequality (D.29) by induction: The base case $k = 1$ holds by inequality (D.30), and if we assume the bound holds for some $k \geq 1$, then

$$\begin{aligned}
\frac{L}{2\tau_1^2}\|\nabla\mathcal{L}_n(\theta^{(k+1)})\|_2 &\leq \left(\frac{L}{2\tau_1^2}\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2\right)^2 + C\tilde{r}_{priv} \\
&\leq \left\{\left(\frac{L}{2\tau_1^2}\|\nabla\mathcal{L}_n(\theta^{(0)})\|_2\right)^{2^k} + 3C\tilde{r}_{priv}\right\}^2 + C\tilde{r}_{priv} \\
&\leq \left(\frac{L}{2\tau_1^2}\|\nabla\mathcal{L}_n(\theta^{(0)})\|_2\right)^{2^{k+1}} + C\tilde{r}_{priv}\left(\frac{3}{2} + 9C\tilde{r}_{priv}\right) + C\tilde{r}_{priv} \\
&< \left(\frac{L}{2\tau_1^2}\|\nabla\mathcal{L}_n(\theta^{(0)})\|_2\right)^{2^{k+1}} + 3C\tilde{r}_{priv},
\end{aligned}$$

where the first inequality comes from inequality (D.30); the second inequality follows by the induction hypothesis; and the last inequality uses the fact that the sample size condition $n = \Omega\left(\frac{\sqrt{Kp\log(Kp/\xi)}}{\mu}\right)$ guarantees that $C\tilde{r}_{priv} < \frac{1}{18}$. This completes the inductive step.

Note that inequality (D.29) implies that

$$\frac{L}{2\tau_1^2}\|\nabla\mathcal{L}_n(\theta^{(K)})\|_2 < \left(\frac{1}{2}\right)^{2^K} + 3C\tilde{r}_{priv}.$$

Hence, for $K \geq \frac{1}{\log(2)}\log\left(\frac{\log(C\tilde{r}_{priv})}{\log(1/2)}\right)$, we have

$$4C\tilde{r}_{priv} \geq \frac{L}{2\tau_1^2}\|\nabla\mathcal{L}_n(\theta^{(K)})\|_2 \geq \frac{L}{2\tau_1^2}\cdot 2\tau_1\|\theta^{(K)} - \hat{\theta}\|_2,$$

where the last inequality holds by LSC. Rearranging yields the desired result.

Supporting Lemmas

In the statements and proofs of these lemmas, we work under the assumption that

$$\max_{k<K} \|Z_k\|_2 \leq 4\sqrt{p} + 2\sqrt{2\log(2K/\xi)} \text{ and}$$
$$\max_{k<K} \|W_k\|_2 \leq \sqrt{2p\log(4Kp/\xi)},$$

in addition to the assumptions stated in Theorem 3.

Lemma 10. *For all* $0 \leq k < K$*, we have*

$$(a)\ \|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \leq \min\left\{\tau_1 r, \frac{\tau_1^2}{L}\right\},$$
$$(b)\ \|\theta^{(k)} - \hat{\theta}\|_2 \leq r, \text{ and}$$
$$(c)\ \frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(k+1)}\|_2 \leq \left(\frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2\right)^2 + C\tilde{r}_{priv}, \tag{D.30}$$

where $\tilde{r}_{priv} := \frac{\sqrt{Kp\log(Kp/\xi)}}{\mu n}$ *and* C *depends on* τ_1*,* B *and* $\bar{B}$*.*

In addition, $\|\theta^{(K)} - \hat{\theta}\|_2 \leq r$*.*

Proof. We induct on k. For the inductive hypothesis, consider the case $k = 0$. Note that (a) holds by assumption. Lemma 11 then implies that (b) holds, and then Lemma 12 implies that (c) holds, as well.

Turning to the inductive step, assume that (a), (b), and (c) all hold for some $k \geq 0$. We will show that $\|\nabla \mathcal{L}_n(\theta^{(k+1)})\|_2 \leq \min\left\{\tau_1 r, \frac{\tau_1^2}{L}\right\}$; then Lemmas 11 and 12 imply that statements (b) and (c) hold, as well. Note that by the inductive hypothesis, we have

$$\frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 \leq \frac{1}{2}.$$

Then by (a) and (c), we have

$$\begin{aligned}\frac{L}{2\tau_1^2}\|\nabla\mathcal{L}_n(\theta^{(k+1)}\|_2 &\leq \left(\frac{L}{2\tau_1^2}\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2\right)^2 + C\tilde{r}_{priv} \\ &\leq \frac{1}{2}\left(\frac{L}{2\tau_1^2}\|\nabla\mathcal{L}_n(\theta^{(k)})\|_2\right) + C\tilde{r}_{priv} \\ &\leq \frac{1}{2}\cdot\frac{L}{2\tau_1^2}\min\left\{\tau_1 r, \frac{\tau_1^2}{L}\right\} + C\tilde{r}_{priv} \\ &\leq \frac{L}{2\tau_1^2}\min\left\{\tau_1 r, \frac{\tau_1^2}{L}\right\}, \end{aligned} \tag{D.31}$$

using the sample size assumption $n = \Omega\left(\frac{\sqrt{Kp\log(Kp/\xi)}}{\mu}\right)$. This completes the inductive step.

Finally, note that the argument used to establish inequality (D.31) above shows that we can also deduce statements (a) and (b) for $k = K$. □

Lemma 11. *Suppose* $\|\nabla\mathcal{L}_n(\theta)\|_2 < 2\tau_1 r$. *Then* $\|\theta - \hat{\theta}\|_2 \leq r$.

Proof. We prove the contrapositive. Suppose $\|\theta - \hat{\theta}\|_2 > r$, and let $\widetilde{\theta}$ denote the point on the boundary of $\mathcal{B}_r(\hat{\theta})$ lying on the segment between θ and $\hat{\theta}$. By the LSC condition, we have

$$\langle\nabla\mathcal{L}_n(\widetilde{\theta}), \widetilde{\theta} - \hat{\theta}\rangle \geq 2\tau_1\|\widetilde{\theta} - \hat{\theta}\|_2^2.$$

Rearranging and defining $v = \frac{\widetilde{\theta}-\hat{\theta}}{\|\widetilde{\theta}-\hat{\theta}\|_2}$, we then have

$$\langle\nabla\mathcal{L}_n(\widetilde{\theta}), v\rangle \geq 2\tau_1 r.$$

Further note that if we define $f(t) = \langle\nabla\mathcal{L}_n(\hat{\theta}+tv), v\rangle$ for $t \geq 0$, then $f'(t) = v^T\nabla^2\mathcal{L}_n(\hat{\theta}+tv)v \geq 0$, so $f(t)$ is increasing. This implies that

$$\|\nabla\mathcal{L}_n(\theta)\|_2 \geq \langle\nabla\mathcal{L}_n(\theta), v\rangle \geq \langle\nabla\mathcal{L}_n(\widetilde{\theta}), v\rangle \geq 2\tau_1 r,$$

giving the desired result. □

Lemma 12. *Suppose* $\theta^{(k)} \in \mathcal{B}_r(\hat{\theta})$. *Then inequality* (D.30) *holds.*

Proof. We begin by establishing a bound on $\|\tilde{N}_k\|_2$. By LSC, we have $\| \left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1} \|_2 \leq \frac{1}{2\tau_1}$. Furthermore, by our sample size assumption, we have $\max_{k<K} \|\tilde{W}_k\|_2 \leq \tau_1$. Then the series expansion (D.28) and the bounds on $\|Z_k\|_2$ and $\|W_k\|_2$ imply that for all $k < K$, we have

$$\begin{aligned}
\|\tilde{N}_k\|_2 &\leq \frac{1}{2\tau_1}\|\tilde{Z}_k\|_2 + \sum_{j=1}^{\infty} \frac{\|\tilde{W}_k\|_2^j}{(2\tau_1)^{(j+1)}} \left\{\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 + \|\tilde{Z}_k\|_2\right\} \\
&\leq \frac{1}{2\tau_1}\|\tilde{Z}_k\|_2 + \frac{1}{2\tau_1}\frac{\|\tilde{W}_k\|_2}{2\tau_1 - \|\tilde{W}_k\|_2} \left\{\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2 + \|\tilde{Z}_k\|_2\right\} \\
&\leq \frac{2B\sqrt{2K}}{2\mu n \tau_1}\|Z_k\|_2 + \frac{1}{2\tau_1}\frac{\|\tilde{W}_k\|_2}{2\tau_1 - \|\tilde{W}_k\|_2} \left\{B + \frac{2B\sqrt{2K}}{\mu n}\|Z_k\|_2\right\} \\
&\leq \frac{B\sqrt{2K}(4\sqrt{p} + 2\sqrt{2\log(2K/\xi)})}{\mu n \tau_1} \\
&\quad + \frac{2\bar{B}\sqrt{2K}}{2\tau_1 \mu n}\frac{\|W_k\|_2}{\tau_1}\left(B + \frac{2B\sqrt{2K}(4\sqrt{p} + 2\sqrt{2\log(2K/\xi)})}{\mu n}\right) \\
&\leq C_0 \frac{B\sqrt{K}\{\sqrt{p} + \sqrt{\log(2K/\xi)}\} + \bar{B}B\tau_1^{-1}\sqrt{Kp\log(Kp/\xi)}}{\mu n \tau_1} = C'\tilde{r}_{priv}, \qquad \text{(D.32)}
\end{aligned}$$

where $C_0 > 0$ is a constant.

To establish inequality (D.30), we write $\theta^{k+1} = \theta^{(k)} + \Delta\theta^{(k)} + \tilde{N}_k$, where

$$\Delta\theta^{(k)} = -\left\{\nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}^{-1} \nabla \mathcal{L}_n(\theta^{(k)}).$$

Using Condition 4, the LSC condition, and the triangle inequality, we then obtain

$$
\begin{aligned}
&\|\nabla \mathcal{L}_n(\theta^{(k+1)})\|_2 \\
&= \left\|\nabla \mathcal{L}_n\left(\theta^{(k)} + \Delta\theta^{(k)} + \tilde{N}_k\right)\right\|_2 \\
&= \left\|\nabla \mathcal{L}_n\left(\theta^{(k)} + \Delta\theta^{(k)} + \tilde{N}_k\right) - \nabla \mathcal{L}_n(\theta^{(k)}) - \nabla^2 \mathcal{L}_n(\theta^{(k)})\left(\Delta\theta^{(k)} - \tilde{N}_k + \tilde{N}_k\right)\right\|_2 \\
&= \left\|\int_0^1 \left\{\nabla^2 \mathcal{L}_n\left(\theta^{(k)} + t(\Delta\theta^{(k)} + \tilde{N}_k)\right) - \nabla^2 \mathcal{L}_n(\theta^{(k)})\right\}(\Delta\theta^{(k)} + \tilde{N}_k)\mathrm{d}t + \nabla^2 \mathcal{L}_n(\theta^{(k)})\tilde{N}_k\right\|_2 \\
&\leq \left\|\int_0^1 Lt\|\Delta\theta^{(k)} - \tilde{N}_k\|_2^2 \mathrm{d}t + \nabla^2 \mathcal{L}_n(\theta^{(k)})\tilde{N}_k\right\|_2 \\
&\leq \frac{L}{2}\|\Delta\theta^{(k)} + \tilde{N}_k\|_2^2 + \|\nabla^2 \mathcal{L}_n(\theta^{(k)})\tilde{N}_k\|_2 \\
&\leq \frac{L}{2}\|\Delta\theta^{(k)}\|_2^2 + \left(L\|\Delta\theta^{(k)}\|_2 + \bar{B}\right)\|\tilde{N}_k\|_2 + \frac{L}{2}\|\tilde{N}_k\|_2^2 \\
&\leq \frac{L}{2\tau_1^2}\|\nabla \mathcal{L}_n(\theta^{(k)})\|_2^2 + \left(L\|\Delta\theta^{(k)}\|_2 + \bar{B}\right)\|\tilde{N}_k\|_2 + \frac{L}{2}\|\tilde{N}_k\|_2^2.
\end{aligned}
$$

Since LSC guarantees that $\|\Delta\theta^{(k)}\|_2 \leq \frac{B}{\tau_1}$, the desired result follows from the bound (D.32) on $\|\tilde{N}_k\|_2$ and the sample size condition. □

D.5 Numerical Illustrations

In this section, we explore the performance of our algorithms on several real data sets.

D.5.1 Linear Regression

To further explore the performance of our proposed methods, we consider fitting a linear regression model to a housing price data set utilized in Lei (2011). The data consist of price, square footage, year of sale, and county for 348,189 homes sold in the San Francisco Bay Area from 2003 to 2006. We excluded all records with missing values, and combined several counties such that the number of county categories was reduced from 9 to 5. The remaining data set contains 286,537 records. Additionally, price, square footage, and year were standardized by subtracting their medians and dividing by median absolute deviation. (Note that differentially private estimates of the median and MAD of these variables could also be returned. This comes at an additional cost in terms of privacy budget, but may be desirable for interpretability or prediction purposes.)

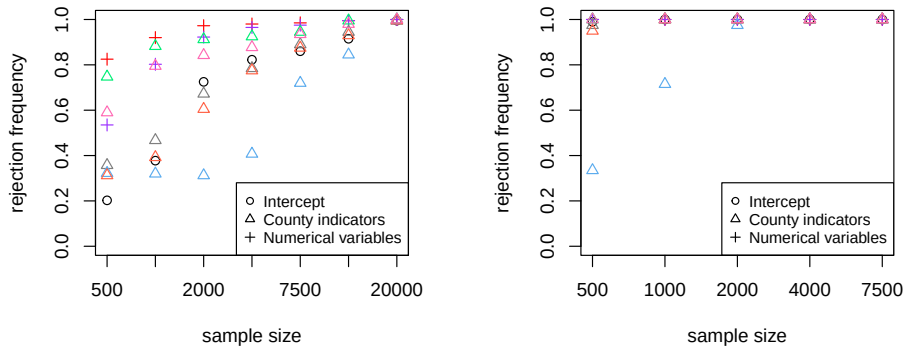
With this data set, we used noisy gradient descent with loss function as in equation (4.5) to estimate a linear regression model for predicting the home price from the remaining variables. The total privacy budget for estimation and corresponding inference was $\mu = 0.25$, the number of iterations K was 100, and optimization was initialized at $\beta^{(0)} = 0$ and $\sigma^{(0)} = 1$. The estimated coefficients of this regression model are in Table D.1.

	Value	Std. Error	z value	p value
(Intercept)	-0.184	0.002	-88.8	0
bsqft	0.643	0.001	456.0	0
date	0.458	0.002	269.8	0
Contra Costa	-0.274	0.003	-88.9	0
MSS	0.873	0.004	209.8	0
NS	-0.156	0.004	-41.8	0
Santa Clara	0.239	0.003	80.4	0
Solano	-0.672	0.003	-223.4	0

Table D.1: Coefficients estimated via noisy gradient descent ($\mu = 0.25$)

The large number of observations in this data set means that only a small amount of noise

must be added to guarantee privacy, and consequently, the results are practically indistinguishable from the non-private case. The inference procedures outlined in Section 4.4 indicate that all of the private regression coefficients are statistically significant at any reasonable threshold. To obtain an approximate notion of performance on smaller data sets, we repeatedly applied noisy gradient descent to random subsamples of the original data set. In each private implementation, the total privacy budget was $\mu = 0.25$, and optimization was initialized at $\beta^{(0)} = 0$ and $\sigma^{(0)} = 1$. The number of iterations K was set to 50 for sample size 500; 75 for sample sizes 750, 1,000, and 2,000; and 100 for all larger sample sizes.

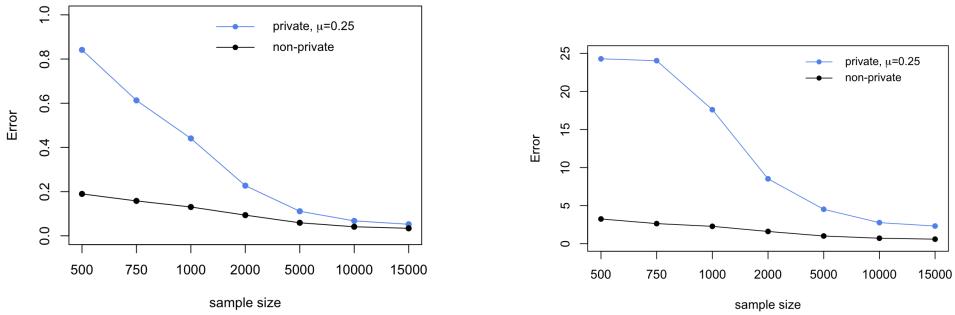


(a) (b)

Figure D.1: Frequency with which the hypothesis β_j was rejected, for each of $j = 1, \ldots, 8$, under (a) differentially private gradient descent and inference, and (b) the corresponding non-private estimation and inference

The non-private method finds most coefficients significant at the 0.05 level over 90% of the time for all sample sizes considered. The only exception is the indicator for the combined Napa and Sonoma counties ("NS"), which catches up to that 90% frequency by a sample size of 2000. At the smaller sample sizes, the private method finds most of the coefficients significant at a materially lower frequency, although results vary widely across the different predictors in the model. By a sample size of 4000, the private method finds most of the coefficients (except for the Napa-Sonoma county indicator) significant approximately 80% of the time, and at a sample size of 20,000 the

private method returns the same results as its non-private counterpart. Note that this is not exactly an analysis of statistical power, as we do not know that the model coefficients are truly nonzero. Nonetheless, this gives some sense of how the privacy constraints affect inference at a range of sample sizes.



(a) Linear regression with housing data (b) Logistic regression with bank data

Figure D.2: Subsampled estimates of $\mathbb{E}(\|\hat{\beta} - \beta\|_2)$.

Figure D.2(a) demonstrates the error of private and non-private estimates of the parameter vector β calculated from subsamples of the housing data set. Error is calculated with respect to the non-private estimator obtained from the full data set and averaged over 400 repetitions at each sample size. The number of iterations K was set to 50 for sample size 500; 75 for sample sizes 750, 1000, and 2000; and 100 for all larger sample sizes. The privacy mechanism has a relatively large impact on the estimation error for β in small samples, but the gap between private and non-private estimators closes with increasing sample size.

D.5.2 Logistic regression

In this example, we fit a logistic regression model to a data set from Moro et al. (2014), describing customer relationships with a bank in Portugal. To fit this model, we used noisy gradient descent to minimize a version of cross-entropy loss modified to include Mallows-style weights:

$$\mathcal{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(-y_i \log\left(\frac{1}{1+\exp(x_i^\top\beta)}\right) + (1-y_i)\log\left(\frac{\exp(x_i^\top\beta)}{1+\exp(x_i^\top\beta)}\right)\right) w(x_i), \tag{D.33}$$

where $w(x_i) = \min\left(1, \frac{25}{\|x_i\|_2^2}\right)$.

This data set contains 45,211 records consisting of customer attributes such as age, job type, and types of business conducted with the bank. The response is whether the customer subscribed to a term deposit. For preprocessing, numeric covariates were standardized and categorical covariates were converted to one-hot encoding (i.e., representing a categorical variable with k levels, with $k-1$ binary indicator variables). One covariate (days since previous marketing contact with customer) was excluded as it was undefined for over 80% of observations. After this preprocessing, the data set had 41 covariates.

Figure D.2(b) shows the error of private and non-private estimates of the parameter vector β. As in the previous section, error is calculated with respect to the non-private estimator obtained from the full data set. Results are averaged over 400 repetitions at each sample size. The number of iterations K was set to 50 for sample size 500; 75 for sample sizes 750, 1000, and 2000; and 100 for all larger sample sizes. All optimizations were initialized at $\beta^{(0)} = 0$.

D.6 Benchmark Experiments

In this appendix, we provide some additional numerical comparisons between our proposed noisy gradient descent algorithm and existing methods in the literature.

D.6.1 Linear Regression

We compare the noisy gradient descent approach (4.4) with a clipped gradient descent approach and with Sufficient Statistic Perturbation (SSP) (Vu and Slavkovic, 2009; Wang, 2018), which requires a priori bounds on $|y_i|$ and $\|x_i\|_2^2$. We consider two data scenarios: In the first, data are generated from a model obeying these constraints, and all methods are applied to the raw data. A key advantage of the noisy gradient descent method proposed in Section 4.2 is that it is not subject to the data constraints of SSP. Nevertheless, in Figure D.3, we see that noisy gradient descent remains a competitive estimation method in this constrained data regime. In addition to competitive estimation, noisy gradient descent can easily be extended to produce inference.

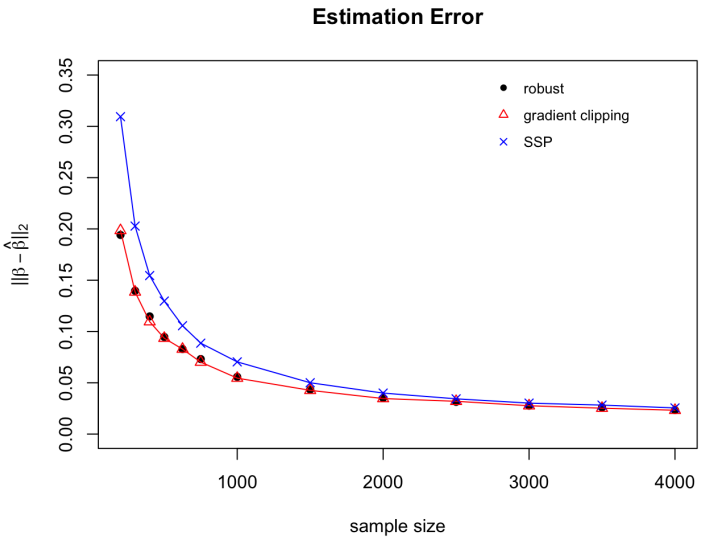


Figure D.3: Linear regression with bounded covariates and responses

Data for Figure D.3 were generated as follows:

1. Generate $z_i \in \mathbb{R}^3$ by drawing $z_{i,j} \overset{i.i.d.}{\sim} Unif(-2, 2)$ for $j = 1, 2, 3$ and $i = 1, ..., n$.

2. Impose an upper bound on $\|z_i\|_2$ by taking $\tilde{z}_i = z_i \cdot \min\left(1, \frac{\sqrt{3}}{\|z_i\|_2}\right)$.

3. Set $x_i = (1, \tilde{z}_i)^T$. By construction, $\|x_i\|_2 \leq 2$ for all i.

4. Draw i.i.d. noise terms u_i from a truncated normal distribution,

$$u_i \sim TN(0, 0.75^2, q_{0.025}, q_{0.975}),$$

where $q_{0.025}$ and $q_{0.975}$ are the 2.5^{th} and 97.5^{th} quantiles of a $N(0, 0.75^2)$ distribution, respectively.

5. Generate responses $y_i = x_i^T\beta + u_i$ for $i = 1, \ldots, n$, with $\beta = (1, -1, 0.5, -0.5)^T$. By construction, $|y_i| \leq 1 + \sqrt{2} + \sqrt{\frac{1}{2}} + q_{0.975}$ for all i.

Implementation details:

- **Noisy gradient descent ("robust"):** This method estimates β using the noisy gradient descent iterations of equation (4.4), with the loss function specified in equation (4.5), in which $c = 1.345$ and $w(x_i) = \min\left(1, \frac{2}{\|x_i\|_2}\right)$. For each sample size n, the number of iterations is $6 \log n$, rounded to the nearest integer. The step size at each iteration is taken to be $\frac{1}{L}$, where L is the Lipschitz constant of the gradient. This implementation is calibrated for μ-GDP with $\mu = 1$.

- **Gradient clipping:** This method performs gradient descent in which gradients are clipped to a user-specified maximum magnitude h. These simulations use a clipped version of the usual squared loss, such that the clipped gradients are of the form

$$\nabla_\beta \mathcal{L}_{n,clip}(\beta) = \frac{-1}{n} \sum_{i=1}^{n} (y_i - x_i^T\beta) x_i \min\left(1, \frac{h}{\left\|(y_i - x_i^T\beta)x_i\right\|_2}\right).$$

Specifically, the clipping threshold h is set equal to $2c$ (c is the tuning parameter of the Huber loss in the example above), so that the maximum magnitude of the gradient is the same for the robust and clipped algorithms. The step size at each iteration is equal in value to the step

size used for noisy gradient descent, and the number of iterations also remains the same. The privacy budget is $\mu = 1$.

- **Sufficient Statistic Perturbation ("SSP"):** This method uses the known bounds on $\|x_i\|_2$ and $|y_i|$ to privately estimate $X^T X$ and $X^T y$, and subsequently produce a private estimate of the ordinary least squares estimator $\hat{\beta} = (X^T X)^{-1} X^T y$. The private estimator is

$$\hat{\beta}_{priv} = \left(X^T X + \frac{x_{max}^2}{\mu/\sqrt{2}} W \right)^{-1} \left(X^T y + \frac{2 x_{max} y_{max}}{\mu/\sqrt{2}} Z \right),$$

  where $x_{max} = \sup_x \|x\|_2 = 2$, $y_{max} = \sup_y |y| = 1 + \sqrt{2} + \sqrt{\frac{1}{2}} + q_{0.975}$, $Z \sim N(0, \mathbb{I}_4)$, and W is a symmetric matrix whose upper-triangular elements, including the diagonal, are i.i.d. $N(0, 1)$. The privacy budget in these simulations is $\mu = 1$.

Next, we consider a second data scenario, in which data are generated from a linear model with Gaussian errors, and do not obey the SSP constraints. Both gradient-based methods are applied to the raw data, while SSP is applied to data which have been non-privately preprocessed to satisfy the requirement of bounded Y and X. In Figure D.4, we see that the two gradient-based methods outperform SSP, particularly in smaller samples. Unlike SSP, noisy gradient descent can naturally address regression problems with unbounded covariates and responses.

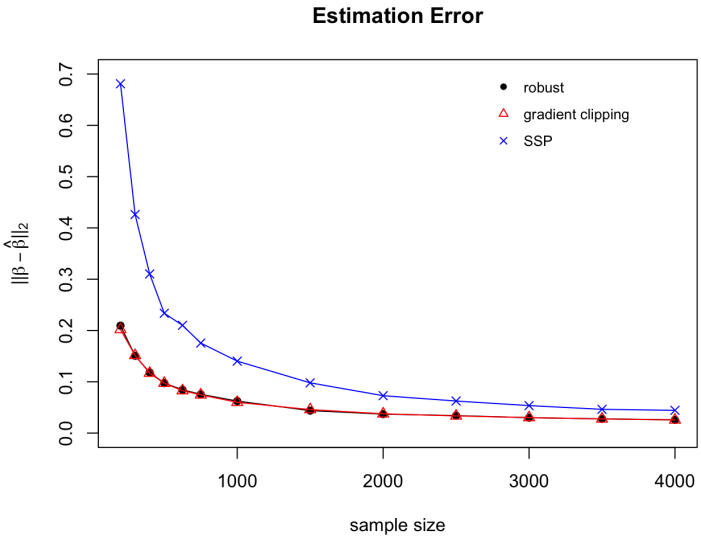


Figure D.4: Linear regression with unbounded covariates and responses

Data for Figure D.4 were generated as follows:

1. Draw $x_i \overset{i.i.d.}{\sim} N(0, \mathbb{I}_4)$.
2. Draw i.i.d. noise terms $u_i \sim N(0, 0.75^2)$.
3. Generate responses $y_i = x_i^T \beta + u_i$ for $i = 1, \ldots, n$, with $\beta = (1, -1, 0.5, -0.5)^T$.
4. Define $\alpha = \max(\max_i \|x_i\|_2, \max_i |y_i|)$.
5. Define $\tilde{X} = \frac{1}{\alpha} X$ and $\tilde{y} = \frac{1}{\alpha} y$.

Noisy gradient descent and clipped gradient descent are applied to the raw data (X, y), with implementation details unchanged from the previous scenario. SSP is applied to the (non-privately) preprocessed data $(\tilde{X}, \tilde{y})$. This is motivated by the observation that

$$(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{y} = \left(\frac{X^T}{\alpha} \cdot \frac{X}{\alpha} \right)^{-1} \frac{X^T}{\alpha} \cdot \frac{y}{\alpha} = (X^T X)^{-1} X^T y.$$

D.6.2 Logistic Regression

We now compare the noisy gradient descent approach (4.4) with clipped gradient descent and Objective Perturbation. Objective Perturbation for regression tasks, as proposed in Chaudhuri et al. (2011) and Kifer et al. (2012), requires a bound on $\|x_i\|_2$ be known a priori.[1] Neither noisy gradient descent with the updates (4.4) nor clipped gradient descent have this constraint on the covariate space. We first compare these methods in the case of a known bound on $\|x_i\|_2$. Within this case, we consider two scenarios:

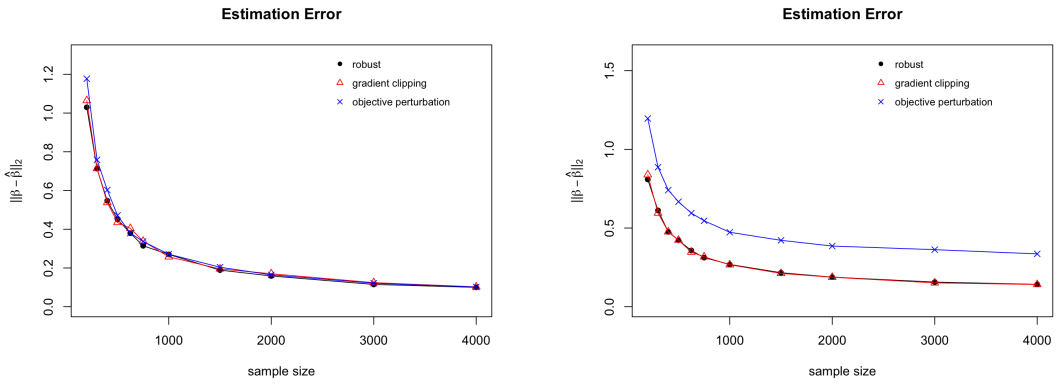
- Scenario 1: Many data points have $\|x_i\|_2$ close to the upper bound. The known bound on the covariates means that clipping or downweighting is not necessary to achieve bounded global sensitivity. Clipping/downweighting thresholds can be set such that no clipping or downweighting is performed, effectively reducing noisy gradient descent and gradient clipping to the same algorithm. Data for this scenario were generated as follows:

  1. Draw $z_i \overset{i.i.d.}{\sim} N(0, \mathbb{I}_3)$ for $i = 1, ..., n$.
  2. Impose an upper bound on $\|z_i\|_2$ by taking $\tilde{z}_i = z_i \cdot \min\left(1, \frac{\sqrt{3}}{\|z_i\|_2}\right)$.
  3. Set $x_i = (1, \tilde{z}_i)^T$. By construction, $\|x_i\|_2 \leq 2$ for all i.
  4. Generate response $y_i \sim \text{Bernoulli}\left(\frac{1}{1+\exp(-x_i^T\beta)}\right)$, with $\beta = (1, -1, 0.5, -0.5)^T$.

- Scenario 2: Few data points have $\|x_i\|_2$ close to the upper bound. While clipping or downweighting is not strictly necessary, it may be advantageous to perform some clipping/ downweighting, because it affects only a small fraction of the data points, in exchange for reducing the amount of added noise. Data for this scenario were generated in the same manner as above, modifying step 3 to instead set $\tilde{z}_i = z_i \min(1, \frac{\sqrt{15}}{\|z_i\|_2})$.

Algorithm performance in these two scenarios is depicted in Figure D.5.

[1] More specifically, the requirements are a bounded gradient of the loss function, and a bounded maximum eigenvalue of the Hessian of the loss function. Due to the special structure of logistic regression, these requirements can be met via a constraint on the norm of x, although in general, this is not the case. For example, in linear regression, one would need to impose additional constraints on the responses y and the radius of the parameter space for β.



(a) (b)

Figure D.5: (a) Scenario 1: Many data points near $\sup_i \|x_i\|_2$. (b) Scenario 2: Few data points near $\sup_i \|x_i\|_2$.

In Scenario 1 (Figure D.5(a)), all methods achieve similar performance. In Scenario 2 (Figure D.5(b)), noisy gradient descent and gradient clipping achieve lower estimation error than Objective Perturbation. Only a small fraction (roughly 5%) of data points had their contributions clipped/downweighted, in exchange for a smaller amount of added noise, allowing these methods to outperform Objective Perturbation. Note that because the amount of clipping/downweighting is small, noisy gradient descent and gradient clipping perform comparably. As discussed in Section 4.2.1, gradient clipping can result in biased estimates, but this effect is less pronounced when the amount of clipping is small.

Implementation details:

- **Noisy gradient descent ("robust"):** This method estimates β using the noisy gradient descent iterates of equation (4.4), with the robust cross-entropy loss function specified in equation (4.6). In Scenario 1, we take $w(x_i) = \min\left(1, \frac{2}{\|x_i\|_2}\right)$, which results in no downweighting, since $\|x_i\|_2 \leq 2$ for all i. In Scenario 2, we take $w(x_i) = \min\left(1, \frac{3}{\|x_i\|_2}\right)$, which does result in downweighting, as the upper bound in this scenario is $\|x_i\|_2 \leq 4$. For each sample size n, the number of iterations is $6 \log n$, rounded to the nearest integer. The step size at each iteration is taken to be $\frac{1}{L}$, where L is the Lipschitz constant of the gradient. This implementation is calibrated for μ-GDP with $\mu = 0.5$.

- **Gradient clipping:** This method performs gradient descent with a clipped version of the usual cross-entropy loss, such that the clipped gradients are of the form

$$\nabla \mathcal{L}_{n,clip}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left((1+e^{-x_i^T\beta})^{-1} - y_i\right)x_i \min\left(1, \frac{h}{\left\|((1+e^{-x_i^T\beta})^{-1} - y_i)x_i\right\|_2}\right).$$

  Specifically, the clipping threshold h is set equal to 2 in Scenario 1 and to 3 in Scenario 2, so that the maximum magnitude of the gradient is the same for the robust and clipped algorithms. The step size at each iteration is taken to be $\frac{1}{L}$, where L is the Lipschitz constant of the gradient. The number of iterations is the same as above, and the privacy budget is $\mu = 0.5$.

- **Objective Perturbation:** This method follows the (ϵ, δ)-differentially private objective perturbation procedure outlined in Algorithm 1 of Kifer et al. (2012). The loss function is the cross-entropy loss, i.e., equation (4.6) with no weight terms $\{w(x_i)\}$. Because this algorithm satisfies (ϵ, δ)-DP and we wish to make a relevant comparison to μ-GDP methods, we refer to Corollary 2.13 of Dong et al. (2022): a mechanism is μ-GDP if and only if it is $(\epsilon, \delta(\epsilon))$-DP for all $\epsilon \geq 0$, where $\delta(\epsilon) = \Phi(-\epsilon/\mu + \mu/2) - e^{\mu}\Phi(-\epsilon/\mu - \mu/2)$. In these simulations, we choose a pair $(\epsilon, \delta(\epsilon))$ such that $\delta < 1/n$ for each sample size n, with $\mu = 0.5$ in the expression for $\delta(\epsilon)$. The step size at each iteration is taken to be $\frac{1}{L}$, where L is the Lipschitz constant of the gradient.

Next we consider logistic regression without a priori bounds on $\|x_i\|_2$. Objective perturbation does not naturally address this data context, but can be extended to such data by including Mallows weights in the objective function (i.e. taking Equation 4.6 as the objective function). We have included this version of Objective Perturbation in our comparison. As seen in Figure D.6, noisy gradient descent and Objective Perturbation applied to the robustified version of the cross-entropy loss achieve comparable performance in this unbounded data regime. Here, we see gradient clipping results in biased estimates, a behavior detailed in Section 4.2.1.

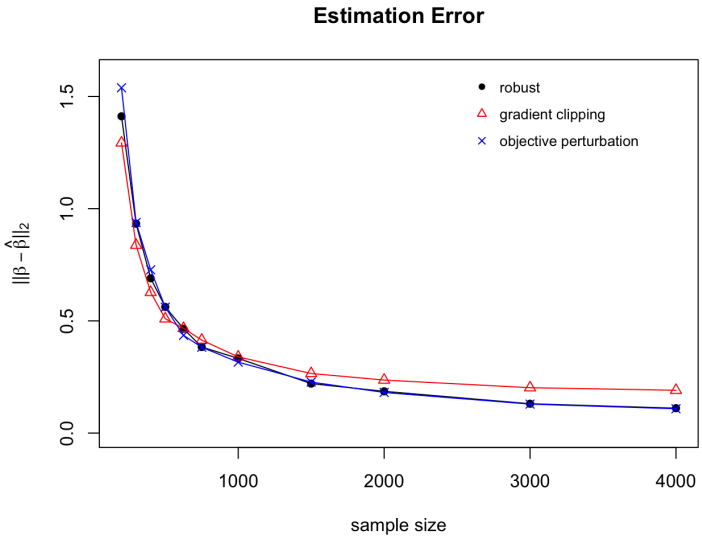


Figure D.6: Logistic regression with unbounded covariates

Data for Figure D.6 were generated in a similar manner to the bounded case above, without the rescaling of covariates in step 3. That is, we set $x_i = (1, z_i)^T$ directly. The gradient clipping threshold is $h = \sqrt{2}$, and noisy gradient descent uses weights $w(x_i) = \min\left(1, \frac{\sqrt{2}}{\|x_i\|_2}\right)$. Objective Perturbation requires a known upper bound on the largest eigenvalue of the Hessian of the loss function. To achieve such a bound for this problem, the weights in its objective function must downweight based on $\|x_i\|_2^2$. Specifically, we take these weights to be $\tilde{w}_i = \min\left(1, \frac{2}{\|x_i\|_2^2}\right)$. All other implementation details are unchanged from the bounded-data case above.