Abstract

A Few Topics in Statistics

Xiaoqian (Dana) Yang

2019

This dissertation discusses several topics in statistics that I have worked on throughout my years as a graduate student. I have had the opportunity of working with multiple faculty members who have exposed me to a variety of research topics including Bayesian analysis, fairness in machine learning, and random graphs. Abstract for each individual piece of work can be found at the start of the chapters.

A Few Topics in Statistics

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Xiaoqian (Dana) Yang

Dissertation Directors:
Prof. David Pollard
Prof. John Lafferty
Prof. Yihong Wu

December 2019

Acknowledgments

Towards finishing this dissertation, I have received a great deal of support from my advisors: Prof. David Pollard, Prof. John Lafferty and Prof. Yihong Wu. Being able to work with them three truly opened up my perspective, as they exposed me not only different areas in statistics, but also various approaches to research. For that I consider myself extremely lucky. This is going to sound very repetitive but I would like to thank them one by one.

I would first like to thank Prof. Pollard, who sparked my interest in theoretical statistics and guided me through the research process with his expert knowledge. Being an excellent teacher, Prof. Pollard taught me a lot of things that I will sure find helpful even long after I finish my PhD.

I would like to thank Prof. Lafferty, for introducing me to a world different from purely theoretical statistics. Prof. Lafferty showed me a way of doing research I had not experienced before. The constant guidance and feedback I received from him made me enjoy every stage of the research project.

I would like to thank Prof. Wu, who is always willing to discuss any research ideas I had anytime, even if most of them do not work. His passionate and hardworking attitude inspires me to become a better scholar.

I am also grateful for all the help I received from the people I did not have the opportunity to work with. I would like to thank Prof. Harrison Zhou for offering his expert advice on a few of my research projects. For the training in statistical computing I would like to thank Prof. Jay Emerson, whose voice I still hear every time I work in *R*.

Finally, even though I am lucky to not have struggled too much, there have been some difficult moments as a graduate student. I would like to thank my family and friends, whose love and support helped me through. This work would not have been possible without them.

Contents

List of Figures

List of Tables

Chapter 1

Posterior Asymptotic Normality for an Individual Coordinate in High-dimensional Linear Regression

Joint work with Prof. David Pollard

Abstract

It is well known that high-dimensional procedures like the LASSO provide biased estimators of parameters in a linear model. In a 2014 paper Zhang and Zhang showed how to remove this bias by means of a two-step procedure. We show that de-biasing can also be achieved by a one-step estimator, the form of which inspires the development of a Bayesian analogue of the frequentists' de-biasing techniques.

1.1 Introduction

Consider the regression model

$$Y = Xb + \epsilon, \quad \epsilon \sim \mathcal{N}(0, I_n). \tag{1.1}$$

The design matrix X is of dimension $n \times p$. The vector $Y \in \mathbb{R}^n$ is the response and $b \in \mathbb{R}^p$ is the unknown parameter. We are particularly interested in the case where $p > n$, for which b itself is not identifiable. In such a setting identifiability can be attained by adding a sparsity constraint, an upper bound on $||b||_0$, the number of nonzero b_i's. That is, the model consists of a family of probability measures $\{\mathbb{P}_b : b \in \mathbb{R}^p, ||b||_0 \leq s^*\}$, and the observation Y is distributed $\mathcal{N}(Xb, I_n)$ under $\mathbb{P}_b$.

We are interested in posterior inference on the vector b, when Y is actually distributed $N(X\beta, I_n)$ for some true sparse β. Throughout this paper the notation β is reserved for the truth, a p-dimensional deterministic vector. The notation b stands for the random vector with marginal distribution μ (*a.k.a.* the prior) and conditional distribution μ_Y (*a.k.a.* the posterior) given Y.

If p were fixed and X were full rank, classical theorems (the Bernstein-von Mises theorem, as in (Van der Vaart, 2000, page 141)) gives conditions under which the posterior distribution of b is asymptotically normal centered at the least squares estimator, with covariance matrix $(X^T X)^{-1}$ under $\mathbb{P}_\beta$.

The classical theorem fails when $p > n$. Although sparse priors have been proposed that give good posterior contraction rates Castillo et al. (2015) Gao et al. (2015), posterior normality of b is only obtained under strong signal-to-noise ratio (SNR) conditions, such as those of (Castillo et al., 2015, Corollary 2), which forced the posterior to eventually have the same support as β. Effectively, their conditions reduce the problem to the classical, fixed dimensional case. However that is arguably not the most interesting scenario. Without the SNR condition, (Castillo et al., 2015, Theorem 6) pointed out that under the sparse prior, the posterior distribution of b behaves like a mixture of Gaussians.

There is hope to obtain posterior normality results without the SNR condition if one considers the situation where only one component of b is of interest, say b_1, without loss of generality. All the other components are viewed as nuisance parameters. As shown by Zhang and Zhang (2014) in a non-Bayesian setting, it is possible to construct asymptotically unbiased estimators such that

$$\widehat{\beta}_1 = \beta_1 + \frac{X_1^T \epsilon}{\|X_1\|_2^2} + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{1.2}$$

Here and subsequently $o_p(\cdot)$ is a shorthand for a stochastically small order term under $\mathbb{P}_\beta$ and X_i denotes the i'th column of X. Similarly X_{-i} denotes the $n \times (p-1)$ matrix formed by all columns of X except for X_i. For $J \subset [p]$ denote by b_J the vector $(b_j)_{j \in J}$ in $\mathbb{R}^{|J|}$. Write b_{-1} for $b_{[p]\setminus\{1\}}$. The $\|\cdot\|_2$ norm on a vector refers to the Euclidean norm.

Approximation (1.2) is useful when $\|X_1\|_2 = O(\sqrt{n})$, in which case the expansion (1.2) implies weak convergence (Pollard, 2002, page 171):

$$\|X_1\|_2(\widehat{\beta}_1 - \beta_1) \rightsquigarrow \mathcal{N}(0,1) \quad \text{under } \mathbb{P}_\beta.$$

Such behavior for $\|X_1\|$ is obtained with high probability when the entries of X are generated *i.i.d.* from the standard normal distribution. More precisely, Zhang and Zhang (2014) proposed a two-step estimator $\widehat{\beta}_1^{(ZZ)}$ that satisfies (1.2) under some regularity assumptions on X and no SNR conditions. The exact form of the estimator $\widehat{\beta}_1^{(ZZ)}$ will be given in section 1.2.1. Zhang and Zhang required the following behavior for X.

Assumption 1.1. *Let $\gamma_i = X_1^T X_i / \|X_1\|_2^2$, and $\lambda_n = \sqrt{\frac{\log p}{n}}$. There exists a constant $c_1 > 0$ for which*

$$\max_{2 \le i \le p} |\gamma_i| \le c_1 \lambda_n.$$

In addition, $\max_{i \le p} \|X_i\|_2 = O(\sqrt{n})$.

Assumption 1.2. *(REC($3s^*, c_2$)) There exist constants $c' > 0$ and $c_2 > 2$ for which*

$$\kappa(3s^*, c_2) = \min_{\substack{J \subset [p], \\ |J| \le 3s^*}} \inf_{\substack{b \ne 0, \\ \|b_{J^C}\|_1 \le c_2 \|b_J\|_1}} \frac{\|Xb\|_2}{\sqrt{n}\|b_J\|_2} > c' > 0. \tag{1.3}$$

Assumption 1.3. *The model dimension satisfies*

$$s^* \log p = o(\sqrt{n}).$$

Remark 1.1. *Assumption 1.2 is known as the restricted eigenvalue condition (Bickel et al., 2009, page 1710) required for penalized regression estimators such as the LASSO estimator (Tibshirani, 1996, page 1) and the Dantzig selector (Candes and Tao, 2007, page 1) to enjoy optimal l^1 and l^2 convergence rates. Note that assumption 1.2 forces $\|X_i\|_2 > c' \sqrt{n}$ for all $i \leq p$. Therefore assumptions 1.1 and 1.2 imply that the lengths of all columns of X are of order $\Theta(\sqrt{n})$.*

Remark 1.2. *Assumptions 1.1 and 1.2 are satisfied with high probability when the $n \times p$ entries of X are generated i.i.d. from a sub-Gaussian random variable with a fixed sub-Gaussian parameter. Assumption 1.1 can be easily proved via the Markov inequality. For the proof of assumption 1.2 see Mendelson et al. (2008) and Zhou (2009).*

Zhang and Zhang (2014) provided the following theorem.

Theorem 1.1 ((Zhang and Zhang, 2014, Section 2.1, 3.1))**.** *Under assumptions 1.1, 1.2 and 1.3, the estimator $\widehat{\beta}_1^{(ZZ)}$ has expansion* (1.2).

The goal of this paper is to give a Bayesian analogue for Theorem 1.1, in the form of a prior distribution on b such that as $n, p \to \infty$, the posterior distribution of b_1 starts to resemble a normal distribution centered around an estimator in the form of (1.2). We provide the following bias corrected version of the sparse prior proposed by Gao, van der Vaart, and Zhou (2015).

The bias corrected prior distribution on $b \in \mathbb{R}^p$:

1. *Let the sparsity level s of b_{-1} obey the probability mass function $\pi(s) \propto \frac{\Gamma(s)}{\Gamma(s/2)} \exp(-2Ds \log \frac{e(p-1)}{s})$ for a positive constant D.*
2. *Denote the projection matrix onto span(X_1) by H. Write $W = (I - H)X$. Let $S|s \sim Unif\,(Z_s := \{S \subset \{2, ..., p\} : |S| = s, W_S \text{ is full rank}\})$.*
3. *Given S, let b_S have density $f_S(b_S) \propto \exp(-\eta \|W_S b_S\|_2)$ for a positive constant η. Set $b_{S^c} = 0$.*
4. *Let $b_1|b_{-1} \sim \mathcal{N}(-\sum_{i\geq 2} \gamma_i b_i, \sigma_n^2)$ where $\sigma_n^2 \gg \|\beta\|_1 \lambda_n / \|X_1\|_2$ and γ_i, λ_n are as defined in assumption 1.1.*

The following is the main result of this paper.

Theorem 1.2. *Under assumptions 1.1, 1.2 and 1.3, for each constant η, there exists a large enough constant $D > 0$ for which the prior distribution on b described above gives a posterior distribution of $\|X_1\|_2(b_1 - \widehat{\beta}_1)$ that satisfies*

$$\left\| \mathcal{L}\left(\|X_1\|_2 (b_1 - \widehat{\beta}_1) | Y \right) - \mathcal{N}(0,1) \right\|_{BL} \to 0 \text{ in } \mathbb{P}_\beta, \tag{1.4}$$

where $\widehat{\beta}_1$ is an estimator of β_1 with expansion (1.2).

Here $\|\cdot\|_{BL}$ denotes the bounded Lipschitz norm, which metrizes the topology of weak convergence (Dudley, 1972, page 323). The bounded Lipschitz norm between two probability measures P and Q on $\mathcal{X}$ is defined as $\|P - Q\|_{BL} = \sup_f |Pf - Qf|$ where the supremum is over all functions $f : \mathcal{X} \to [-1, 1]$ with Lipschitz constant at most 1.

An estimator $\widehat{\beta}_1$ with expansion (1.2) is the appropriate centering for the posterior distribution of b_1 given Y. To see that, take the two-sided α-credible interval as an example.

Recall that μ_Y stands for the posterior distribution of b given Y. It is an easy consequence of (1.4) that (by taking a sequence of bounded Lipschitz functions approaching an indicator function):

$$\left|\mu_Y\left\{\|X_1\|_2\left|b_1-\widehat{\beta}_1\right|\leq\Phi^{-1}\left(\frac{1+\alpha}{2}\right)\right\}-\alpha\right|\to 0 \text{ in } \mathbb{P}_\beta, \quad \text{or}$$

$$\mu_Y\left\{b_1\in\left[\widehat{\beta}_1-\frac{\Phi^{-1}((1+\alpha)/2)}{\|X_1\|_2},\widehat{\beta}_1+\frac{\Phi^{-1}((1+\alpha)/2)}{\|X_1\|_2}\right]\right\}=\alpha+o_p(1).$$

On the other hand, for any estimator $\widehat{\beta}_1$ with expansion (1.2), under the assumption that $\|X_1\|_2=O(\sqrt{n})$,

$$\mathbb{P}_\beta\left\{\beta_1\in\left[\widehat{\beta}_1-\frac{\Phi^{-1}((1+\alpha)/2)}{\|X_1\|_2},\widehat{\beta}_1+\frac{\Phi^{-1}((1+\alpha)/2)}{\|X_1\|_2}\right]\right\}=\alpha+o(1).$$

That is, the Bayesian's credible interval and the frequentist's confidence interval are both $\left[\widehat{\beta}_1-\Phi^{-1}((1+\alpha)/2)/\|X_1\|_2,\widehat{\beta}_1+\Phi^{-1}((1+\alpha)/2)/\|X_1\|_2\right]$, which covers the truth β_1 roughly α proportion of the time. In other words, Theorem 1.2 implies that the Bayesian inference on b_1 and frequentist inference on β_1 are aligned in the asymptotics.

We would like to point out that although our Bayesian analogue of bias correction matches the frequentist's treatment in terms of statistical performance, the from of posterior distribution involves up to 2^p integrations and is therefore very expensive to compute.

The paper is organized as follows. We begin by discussing the frequentists' de-biasing techniques in section 1.2.1, including the two-step procedure developed by Zhang and Zhang (2014) and a one-step estimator. We show that the one-step estimator also achieves de-biasing. In section 1.2.2 we use the form of the one-step estimator to illustrate the intuition behind the construction of the bias corrected prior distribution. The proof of our main result Theorem 1.2 is given in section 3.5.

1.2 Main results

1.2.1 How does de-biasing work?

This section describes the main idea behind the construction of the two-step de-biasing estimator proposed by Zhang and Zhang (2014). An estimator is proposed to provide another way of interpreting the two-step procedure. The success of these estimators inspired us to design a prior distribution that achieves de-biasing under the same set of assumptions.

In sparse linear regression, penalized likelihood estimators such as the LASSO are often used and tend to give good global properties, such as control of the l_1 loss:

$$\mathbb{P}_\beta \left\{ \|\widetilde{\beta} - \beta\|_1 > C s^* \lambda_n \right\} \to 0 \text{ as } n, p \to \infty \text{ for some } C > 0, \tag{1.5}$$

where λ_n is as defined in assumption 1.1. For example, (Bickel et al., 2009, Theorem 7.1) showed that under the REC condition (assumption 1.2) the LASSO estimator satisfies (1.5).

In general, penalized likelihood estimators introduce bias for the estimation of individual coordinates. To eliminate this bias, Zhang and Zhang (2014) proposed a two-step procedure using the following idea. First find a $\widetilde{\beta}$ that satisfies (1.5), perhaps via a LASSO procedure. Then define

$$\widehat{\beta}_1^{(ZZ)} = \arg\min_{b_1 \in \mathbb{R}} \left\| Y - X_{-1}\widetilde{\beta}_{-1} - b_1 X_1 \right\|_2^2. \tag{1.6}$$

Remark 1.3. *The estimator given by* (1.6) *is not exactly the same as the one that appears in Zhang and Zhang (2014). Note that* $\widehat{\beta}_1^{(ZZ)}$ *can be equivalently written as*

$$\widehat{\beta}_1^{(ZZ)} = \widetilde{\beta}_1 + \frac{X_1^T (Y - X\widetilde{\beta})}{X_1^T X_1}.$$

Compare with the estimator proposed by Zhang and Zhang (2014) which takes the form

$$\widehat{\beta}_1 = \widetilde{\beta}_1 + \frac{Z_1^T(Y - X\widetilde{\beta})}{Z_1^T X_1}, \tag{1.7}$$

where Z_1 is some pre-calculated vector, typically obtained by running penalized regression of X_1 on X_{-1} and taking the regression residual. Getting a Bayesian analogue for (1.7) *may be possible. But we choose to present our findings on the simpler version* (1.6) *to better illustrate the idea behind the prior design.*

Since $\widetilde{\beta}^{-1}$ is obtained via some penalized likelihood procedure, the estimator in (1.6) essentially penalizes the size of all coordinates except the one of interest. Under assumptions 1.1, 1.2 and 1.3, the two-step estimator $\widehat{\beta}_1^{(ZZ)}$ is asymptotically unbiased with expansion (1.2).

We show in the next theorem that the same asymptotic behavior can be obtained in a single step. The idea of penalizing all coordinates but one is seen more clearly here. By leaving one term out of the LASSO penalty, de-biasing is achieved. This observation inspired us to construct our bias corrected prior (see section 1.2.2) such that the parameter of interest is not penalized.

Theorem 1.3. *Define*

$$\widehat{\beta} = \arg\min_{b \in \mathbb{R}^p} \left(\|Y - Xb\|_2^2 + \eta_n \sum_{i \geq 2} |b_i| \right).$$

Under assumptions 1.1, 1.2 and 1.3, if η_n is a large enough multiple of $n\lambda_n$, the one-step de-biasing estimator $\widehat{\beta}$ achieves l_1 control (1.5) *and de-biasing of the first coordinate simultaneously. The estimator for β_1 satisfies*

$$\widehat{\beta}_1 = \beta_1 + \frac{X_1^T \epsilon}{\|X_1\|_2^2} + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{1.8}$$

Proof. We will first show that $\widehat{\beta}$ satisfies (1.5). It is well known that when the penalty involves all coordinates of b, then the bound on the l_1 norm is true (Bickel et al., 2009, Theorem 7.1). It turned out that leaving one term out the of penalty does not ruin that property.

As in the proof of (Bickel et al., 2009, Theorem 7.1), we compare the evaluation of the penalized log-likelihood function at $\widehat{\beta}$ and the truth β using the definition of $\widehat{\beta}$.

$$\left\|Y - X\widehat{\beta}\right\|_2^2 + \eta_n \left\|\widehat{\beta}_{-1}\right\|_1 \leq \|Y - X\beta\|_2^2 + \eta_n\|\beta_{-1}\|_1.$$

Plug in $Y = X\beta + \epsilon$, the above is reduced to

$$\left\|X(\widehat{\beta} - \beta)\right\|_2^2 \leq 2\sum_{i\leq p} \xi_i(\widehat{\beta}_i - \beta_i) + \eta_n\left(\|\beta_{-1}\|_1 - \left\|\widehat{\beta}_{-1}\right\|_1\right),$$

where $\xi_i = X_i^T\epsilon$. With high probability $|\max_{i\leq n}\xi_i| \leq R = C_2 n\lambda_n$, in which case we have

$$\left\|X(\widehat{\beta} - \beta)\right\|_2^2 \leq 2R\left\|\widehat{\beta} - \beta\right\|_1 + \eta_n\left(\|\beta_{-1}\|_1 - \left\|\widehat{\beta}_{-1}\right\|_1\right). \tag{1.9}$$

From here we can bound $\|\beta_{-1}\|_1 - \|\widehat{\beta}_{-1}\|_1$ by $\|(\widehat{\beta} - \beta)_{-1}\|_1$ using the triangle inequality. But since $\beta_{S_C} = 0$, we can obtain a much tighter bound:

$$\begin{aligned} \|\beta_{-1}\|_1 - \left\|\widehat{\beta}_{-1}\right\|_1 &\leq \left\|(\widehat{\beta} - \beta)_{S\setminus\{1\}}\right\|_1 - \left\|\widehat{\beta}_{S^C\setminus\{1\}}\right\|_1 \\ &= \left\|(\widehat{\beta} - \beta)_{S\setminus\{1\}}\right\|_1 - \left\|(\widehat{\beta} - \beta)_{S^C\setminus\{1\}}\right\|_1. \end{aligned}$$

Combine with (1.9) to deduce that

$$\left\|X(\widehat{\beta} - \beta)\right\|_2^2 \leq (\eta_n + 2R)\left\|(\widehat{\beta} - \beta)_{S\cup\{1\}}\right\|_1 - (\eta_n - 2R)\left\|(\widehat{\beta} - \beta)_{S^C\setminus\{1\}}\right\|_1.$$

By choosing η_n to be a large enough multiple of $n\lambda_n$, we have

$$\left\|X(\widehat{\beta}-\beta)\right\|_2^2 \leq c_3 n\lambda_n \left\|(\widehat{\beta}-\beta)_{S\cup\{1\}}\right\|_1 - c_4 n\lambda_n \left\|(\widehat{\beta}-\beta)_{S^C\setminus\{1\}}\right\|_1 \tag{1.10}$$

for some positive constants c_3, c_4 with $c_3/c_4 \leq 2 < c_2$. Since $\|X(\widehat{\beta}-\beta)\|_2$ is always nonnegative, the inequality above implies

$$\left\|(\widehat{\beta}-\beta)_{S^C\setminus\{1\}}\right\|_1 \leq \frac{c_3}{c_4}\left\|(\widehat{\beta}-\beta)_{S\cup\{1\}}\right\|_1. \tag{1.11}$$

Therefore under assumption 1.2, we have

$$\left\|(\widehat{\beta}-\beta)_{S\cup\{1\}}\right\|_2 \leq \frac{1}{c'\sqrt{n}}\left\|X(\widehat{\beta}-\beta)\right\|_2.$$

Combine with (1.10) to deduce that

$$\begin{aligned}\left\|X(\widehat{\beta}-\beta)\right\|_2^2 \leq & c_3 n\lambda_n \left\|(\widehat{\beta}-\beta)_{S\cup\{1\}}\right\|_1 \\ \leq & c_3 n\lambda_n \sqrt{s^*+1}\left\|(\widehat{\beta}-\beta)_{S\cup\{1\}}\right\|_2 \\ \leq & \frac{c_3}{c'}\sqrt{(s^*+1)\log p}\left\|X(\widehat{\beta}-\beta)\right\|_2.\end{aligned}$$

Hence

$$\left\|X(\widehat{\beta}-\beta)\right\|_2 \leq \frac{c_3}{c'}\sqrt{(s^*+1)\log p}.$$

Again by assumption 1.2, we can go back to bound the l_1 loss.

$$\begin{aligned}\left\|(\widehat{\beta}-\beta)_{S\cup\{1\}}\right\|_1 \leq & \sqrt{s^*+1}\left\|(\widehat{\beta}-\beta)_{S\cup\{1\}}\right\|_2 \leq \frac{1}{c'}\sqrt{\frac{s^*+1}{n}}\left\|X(\widehat{\beta}-\beta)\right\|_2 \\ \leq & \frac{2c_3}{(c')^2}s^*\lambda_n.\end{aligned}$$

From (1.11) we have

$$\left\|\widehat{\beta} - \beta\right\|_1 \leq 2\left(1 + \frac{c_3}{c_4}\right)\frac{c_3}{(c')^2} s^* \lambda_n.$$

That concludes the proof of (1.5). To show (1.8), observe that the penalty term does not involve b_1.

$$\begin{aligned}\widehat{\beta}_1 =& \arg\min_{b_1 \in \mathbb{R}} \left\|Y - X_{-1}\widehat{\beta}_{-1} - b_1 X_1\right\|_2^2 \\ =&\beta_1 + \sum_{i \geq 2} \gamma_i(\beta_i - \widehat{\beta}_i) + \frac{X_1^T \epsilon}{\|X_1\|^2}.\end{aligned} \tag{1.12}$$

We only need to show the second term in (1.12) is of order $o_p(1/\sqrt{n})$. Bound the absolute value of that term with

$$\max_{i \geq 2} |\gamma_i| \cdot \left\|\widehat{\beta}_S - \beta_S\right\|_1 \leq (c_1 \lambda_n)(C_1 s^* \lambda_n),$$

by assumption 1.1 and the l_1 control (1.5). That is then bounded by $O_p(s^* \lambda_n^2) = o_p(1/\sqrt{n})$ by assumption 1.3. □

Remark 1.4. *With some careful manipulation the REC($3s^*, c_2$) condition as in assumption 1.2 can be reduced to REC(s^*, c_2). The proof would require an extra step establishing that $|\widehat{\beta}_1 - \beta_1|$ is of order $o_p(\|\widehat{\beta}_S - \beta_S\|_1) + O_p(1/\sqrt{n})$.*

The ideas in the proofs for the two de-biasing estimators $\widehat{\beta}_1^{(ZZ)}$ and $\widehat{\beta}_1$ are similar. Ideally we want to run the regression

$$\arg\min_{b_1 \in \mathbb{R}} \|Y - X_{-1}\beta_{-1} - b_1 X_1\|^2. \tag{1.13}$$

That gives a perfectly efficient and unbiased estimator. However β_{-1} is not observed. It is natural to replace it with an estimator which is made globally close to the truth β_{-1} using a penalized likelihood approach. As seen in the proof of Theorem 1.3, most of the work goes into establishing global l_1 control (1.5). The de-biasing estimator is then obtained by running an ordinary least squares regression like (1.13), replacing β_{-1} by some estimator satisfying (1.5), so that the solution to the least squares optimization is close to the solution of (1.13) with high probability.

1.2.2 Bayesian analogue of de-biasing estimators

In the Bayesian regime, recall that b is the p-dimensional random vector obeying distribution μ under the prior and μ_Y under the posterior. For the Bayesian analogue to the de-biasing estimators, it is again essential to establish l_1 control on $b_{-1} - \beta_{-1}$, the deviation of b_{-1} from the truth. Such posterior contraction results were established by Castillo et al. (2015) and Gao et al. (2015), which already provide the preliminary steps for our Bayesian procedure. The following lemma in Gao et al. (2015) serves as a Bayesian analogue of (1.5). It gives conditions under which the sparse prior proposed by Gao et al. (2015) enjoys the l_1 minimax rate of posterior contraction.

Lemma 1.1. *(Corollary 5.4, Gao et al. (2015)) Under the following prior distribution,*

1. *Let s have the probability mass function $\pi(s) \propto \frac{\Gamma(s)}{\Gamma(s/2)} \exp(-2Ds \log \frac{ep}{s})$.*
2. *Let $S|s \sim Unif(Z_s := \{S \subset \{1, ..., p\} : |S| = s, X_S \text{ is full rank}\})$.*
3. *Given the subset selection S, let the coefficients b_S have density $f_S(b_S) \propto \exp(-\eta\|X_S b_S\|)$,*

if the design matrix X satisfies

$$\kappa_0((2+\delta)s^*, X) = \inf_{\|b\|_0 \le (2+\delta)s^*} \frac{\sqrt{s^*}\|Xb\|_2}{\sqrt{n}\|b\|_1} \ge c \tag{1.14}$$

for some positive constant c, δ, then for each positive constant η there exist constants $c_3 > 0$ and large enough $D > 0$ for which

$$\mu_Y\{\|b - \beta\|_1 > c_3 s^* \lambda_n\} \to 0 \quad \text{in } \mathbb{P}_\beta \text{ probability},$$

where μ_Y denotes the posterior distribution of b given Y.

Our bias corrected prior described in section 1.1 is obtained by slightly modify the sparse prior of Gao et al. (2015) to give good, asymptotically normal posterior behavior for a single coordinate. As discussed in the last section, classical approaches to de-biasing exploit the idea of penalizing all coordinates except the one of interest. The idea behind the construction of our bias corrected prior is to essentially put the sparse prior only on b_{-1}.

Recall that H is the matrix projecting $\mathbb{R}^n$ to $span(X_1)$. Under the model where $Y \sim \mathcal{N}(Xb, I_n)$, the likelihood function has the factorization

$$\begin{aligned}\mathcal{L}_n(b) &= \frac{1}{\sqrt{n}(2\pi)^{n/2}} \exp\left(-\frac{\|Y - Xb\|_2^2}{2}\right)\\ &= \frac{1}{\sqrt{n}(2\pi)^{n/2}} \exp\left(-\frac{\|HY - HXb\|_2^2}{2}\right)\\ &\quad \times \exp\left(-\frac{\|(I-H)Y - (I-H)Xb\|_2^2}{2}\right).\end{aligned}$$

Write $W = (I - H)X_{-1}$ and reparametrize $b_1^* = b_1 + \sum_{i\geq 2} \gamma_i b_i$ with γ_i as defined in assumption 1.1. The likelihood $\mathcal{L}_n(b)$ can be rewritten as a constant multiple of

$$\exp\left(-\frac{\|HY - b_1^* X_1\|_2^2}{2}\right) \exp\left(-\frac{\|(I-H)Y - Wb_{-1}\|_2^2}{2}\right).$$

The likelihood factorizes into a function of b_1^* and b_{-1}. Therefore if we make b_1^* and b_{-1} independent under the prior, they will be independent under the posterior. In the prior construction we made $b_1|b_{-1} \sim \mathcal{N}(-\sum_{i\geq 2} \gamma_i b_i, \sigma_n^2)$. Hence $b_1^* \sim \mathcal{N}(0, \sigma_n^2)$ and b_1^* is independent of b_{-1}. Note that under the prior distribution b_1 and b_{-1} are not necessarily independent.

The sparse prior put on b_{-1} is analogue to that of (Gao et al., 2015, section 3), using W as the design matrix in the prior construction. By lemma 1.1, b_{-1} is close to β_{-1} in l_1 norm with high posterior probability as long as $\kappa_o((2 + \delta)s^*, W)$ is bounded away from 0.

We main result (Theorem 1.2) states that the prior distribution we propose has the effect of correcting for the bias, in a fashion analogous to that of the two-step procedure $\widehat{\beta}_1^{(ZZ)}$. Let us first give an outline of the proof. The joint posterior distribution of b_1^* and b_{-1} factorizes into two marginals. In the X_1 direction, the posterior distribution of b_1^* is asymptotically Gaussian centered around $\frac{X_1^T Y}{\|X_1\|_2^2} = \beta_1^* + \frac{X_1^T \epsilon}{\|X_1\|_2^2}$. After we reverse the reparametrization we want the posterior distribution of b_1 to be asymptotically Gaussian centered around an efficient estimator $\widehat{\beta_1} = \beta_1 + \frac{X_1^T \epsilon}{\|X_1\|_2^2} + o_p(1/\sqrt{n})$. Therefore we need to show $b_1^* - b_1$ is very close to $\beta_1^* - \beta_1$. That can be obtained from the l_1 control on $b_{-1} - \beta_{-1}$ under the posterior. In the next section we will give the proof of Theorem 1.2 in detail.

1.3 Proof of Theorem 1.2

Since the prior and the likelihood of b_1^* are both Gaussian, the posterior distribution is also Gaussian:

$$b_1^*|Y \sim \mathcal{N}\left(\frac{\sigma_n^2}{1 + \|X_1\|_2^2\sigma_n^2} X_1^T Y, \frac{\sigma_n^2}{1 + \|X_1\|_2^2\sigma_n^2}\right).$$

Independence of b_1^* and b_{-1} under the posterior gives that the above is also the distribution of b_1^* given Y and b_{-1}. Take $\widehat{\beta_1}$ to be any estimator with expansion (1.2). The distribution of $\|X_1\|_2(b_1 - \widehat{\beta_1})$ given Y and b_{-1} is

$$\mathcal{N}\left(\|X_1\|_2\left(\frac{\sigma_n^2}{1 + \|X_1\|_2^2\sigma_n^2} X_1^T Y - \sum_{i\geq 2} \gamma_i b_i - \widehat{\beta_1}\right), \frac{\sigma_n^2\|X_1\|_2^2}{1 + \|X_1\|_2^2\sigma_n^2}\right). \tag{1.15}$$

Note that without conditioning on b_{-1}, the posterior distribution of b_1 is not necessarily Gaussian.

The main part of the proof of Theorem 1.2 is to show that the bounded-Lipschitz metric between the posterior distribution of b_1 and $\mathcal{N}(\widehat{\beta_1}, 1/\|X_1\|_2^2)$ goes to 0 under the truth. From Jensen's inequality and the definition of the bounded-Lipschitz norm we have

$$\begin{aligned}&\left\|\mathcal{L}(\|X_1\|_2(b_1-\widehat{\beta}_1)|Y)-\mathcal{N}(0,1)\right\|_{BL}\\ \leq&\mu_Y^{b_{-1}}\left\|\mathcal{L}(\|X_1\|_2(b_1-\widehat{\beta}_1)|Y,b_{-1})-\mathcal{N}(0,1)\right\|_{BL}. \end{aligned}\tag{1.16}$$

Here $\mu_Y^{b_{-1}}$ stands for the expected value operator under the posterior distribution of b given Y. The superscript is a reminder that the operator integrates over the randomness of b_{-1}.

For simplicity denote the posterior mean and variance in (1.15) as ν_n and τ_n^2 respectively. The bounded-Lipschitz distance is always upper bounded by the total variation distance, and it is at most 2. Therefore

$$\|\mathcal{N}(\mu_1,\sigma_1^2)-\mathcal{N}(\mu_2,\sigma_2^2)\|_{BL}\leq(|\mu_1-\mu_2|+|\sigma_1-\sigma_2|)\wedge 2.$$

Hence (1.16) is bounded by

$$\mu_Y^{b_{-1}}(|\nu_n|\wedge 2)+\mu_Y^{b_{-1}}((|\tau_n-1|)\wedge 2).$$

Therefore to obtain the desired convergence in (1.4), we only need to show

$$\mathbb{P}_\beta\mu_Y^{b_{-1}}(|\nu_n|\wedge 2)\rightarrow 0,\quad\text{and}\tag{1.17}$$

$$\mathbb{P}_\beta\mu_Y^{b_{-1}}((|\tau_n-1|)\wedge 2)\rightarrow 0.\tag{1.18}$$

To show (1.17), notice that the integrand is bounded. Hence it is equivalent to show convergence in probability. Write

$$
\begin{aligned}
|\nu_n| =&\frac{\sigma_n^2\|X_1\|_2^3}{1+\sigma_n^2\|X_1\|_2^2}\left(\beta_1+\frac{X_1^T\epsilon}{\|X_1\|_2^2}+\sum_{i\geq 2}\gamma_i\beta_i\right)\\
&-\|X_1\|_2\sum_{i\geq 2}\gamma_i b_i-\|X_1\|_2\left(\beta_1+\frac{X_1^T\epsilon}{\|X_1\|_2^2}+o_p\left(\frac{1}{\sqrt{n}}\right)\right)\\
\leq&\frac{\|X_1\|_2}{1+\sigma_n^2\|X_1\|_2^2}\left|\beta_1+\frac{X_1^T\epsilon}{\|X_1\|_2}+\sum_{i\geq 2}\gamma_i\beta_i\right|+\sum_{i\geq 2}\gamma_i(\beta_i-b_i)+o_p(1). \qquad (1.19)
\end{aligned}
$$

The first term is no longer random in b, and it can be made as small as we wish now that it is decreasing in σ_n. If we set $\sigma_n^2 \gg \|\beta\|_1\lambda_n/\|X_1\|_2$, this term is of order $o_p(1)$.

For the second term, we will apply lemma 1.1 to deduce that this term also goes to 0 in $\mathbb{P}_\beta\mu_Y^{b_{-1}}$ probability. To apply the posterior contraction result we need to establish the compatibility assumption (1.14) on W.

Lemma 1.2. *Under assumption 1.1, 1.2, 1.3, the matrix* $W = (I - H)X_{-1}$ *satisfies*

$$
\kappa_0((2+\delta)s^*, W) = \inf_{\|b\|_0\leq(2+\delta)s^*}\frac{\sqrt{s^*}\|Wb\|_2}{\sqrt{n}\|b\|_1}\geq c
$$

for some $c, \delta > 0$.

We will prove the lemma after the proof of Theorem 1.2.

To show (1.18), note that the integrand is not a random quantity. It suffices to show

$$
|\tau_n - 1| = \left|\frac{\sigma_n^2\|X_1\|_2}{1+\sigma_n^2\|X_1\|_2^2}-\frac{1}{\|X_1\|_2}\right|\to 0.
$$

That is certainly true for a $\{\sigma_n\}$ sequence chosen large enough. Combine (1.17), (1.18) and the bound on the bounded Lipschitz distance, we have shown

$$
\mathbb{P}_\beta\left\|\mathcal{L}\left(\|X_1\|_2(b_1-\widehat{\beta}_1)|Y\right)-\mathcal{N}(0,1)\right\|_{BL}\to 0.
$$

Proof of lemma 1.2. We will justify the compatibility assumption on W in two steps. First we will show that the compatibility assumption of the X matrix follows from the REC assumption 1.2. Then we will show that the compatibility constant of X and W are not very far apart.

Let us first show that under assumption 1.2, there exist constants $0 < \delta < 1$ and $c > 0$, for which

$$\kappa_0((2+\delta)s^*, X) = \inf_{\|b\|_0 \leq (2+\delta)s^*} \frac{\sqrt{s^*}\|Xb\|_2}{\sqrt{n}\|b\|_1} \geq c.$$

Denote the support of g as S. We have

$$\begin{aligned}
\kappa_0((2+\delta)s^*, X) \geq & \inf_{\|b\|_0 \leq (2+\delta)s^*} \frac{1}{\sqrt{2+\delta}} \frac{\|Xb\|_2}{\sqrt{n}\|b_S\|_2} \\
\geq & \min_{\substack{J \subset [p], \\ |J| \leq 3s^*}} \inf_{\substack{b \neq 0, \\ \|b_{J^C}\|_1 \leq c_2 \|b_J\|_1}} \frac{\|Xb\|_2}{\sqrt{n}\|b_J\|_2} \\
= & \kappa(3s^*, c_2) > 0.
\end{aligned}$$

Now, under assumptions 1.1, 1.2 and 1.3, we will show that there exist constants $0 < \delta' < 1$ and $c' > 0$, for which

$$\kappa_0((2+\delta')s^*, W) \geq \kappa_0((2+\delta)s^*, X) + o(1).$$

For $g \in [R]^{p-1}$, we have

$$\begin{aligned}
\|Wg\|_2 = & \left\| X \begin{bmatrix} 0 \\ g \end{bmatrix} - \sum_{i \geq 2} \gamma_i g_i \right\|_2 \\
\geq & \left\| X \begin{bmatrix} 0 \\ g \end{bmatrix} \right\|_2 - \lambda_n \|g_1\|_2
\end{aligned}$$

by assumption 1.1. Deduce that

$$\begin{aligned}\kappa_0((2+\delta')s^*, W) = &\inf_{|b|_0 \leq (2+\delta)s^*} \frac{\sqrt{s^*}\|Wb\|_2}{\sqrt{n}\|b\|_1}\\ &\geq \kappa_0((2+\delta')s^* + 1, X) - \sqrt{\frac{s^*}{n}}\lambda_n\\ &= \kappa_0((2+\delta')s^* + 1, X) - \frac{\sqrt{s^* \log p}}{n}.\end{aligned}$$

The second term is of order $o(1)$ under assumption 1.3. □

Bibliography

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics 37*(4), 1705–1732. 4, 7, 9

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics 35*(35), 2313–2351. 4

Castillo, I., J. Schmidt-Hieber, A. Van der Vaart, et al. (2015). Bayesian linear regression with sparse priors. *Annals of Statistics 43*(5), 1986–2018. 2, 12

Dudley, R. (1972). Speeds of metric probability convergence. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 22*(4), 323–332. 5

Gao, C., A. W. van der Vaart, and H. H. Zhou (2015). A general framework for Bayes structured linear models. *arXiv:1506.02174*. 2, 4, 12, 13

Mendelson, S., A. Pajor, and N. Tomczak-Jaegermann (2008). Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation 28*(3), 277–289. 4

Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability*, Volume 8 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press. 3

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288. 4

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, Volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge university press. 2

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–242. 2, 3, 4, 6, 7, 8

Zhou, S. (2009). Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*. 4

Chapter 2

Rapid mixing of a Markov chain for an exponentially weighted aggregation estimator

Joint work with Prof. David Pollard

Abstract

The Metropolis-Hastings method is often used to construct a Markov chain with a given π as its stationary distribution. The method works even if π is known only up to an intractable constant of proportionality. Polynomial time convergence results for such chains (rapid mixing) are hard to obtain for high dimensional probability models where the size of the state space potentially grows exponentially with the model dimension. In a Bayesian context, Yang, Wainwright, and Jordan (2016) (=YWJ) recently used the path method to prove rapid mixing for high dimensional linear models.

We propose a modification of the YWJ approach that simplifies the theoretical argument and improves the rate of convergence. The new approach is illustrated by an application to an exponentially weighted aggregation estimation.

2.1 Introduction

Markov chain Monte Carlo (MCMC) is a well-known technique for generating observations from a fixed probability distribution π on the finite state space. For Bayesians $\pi = \pi(\cdot \mid Y)$ is often a posterior distribution.

Yang, Wainwright, and Jordan (2016) (= YWJ) applied the method to a posterior distribution for a model where the observed Y is modeled as having a $N(Xb, \sigma^2 I_n)$ distribution for X an observed $n \times p$ matrix, with p possibly much larger than n. They followed the tradition of studying behavior of the posterior under a model $\mathbb{P}_\theta$ for which $Y \sim N(X\theta, \sigma^2 I_n)$ for a sparse vector θ, that is, a vector whose support set $T := \{j \in [p] : \theta_j \neq 0\}$ was assumed to have cardinality no greater than some pre-specified (small) value s^*. One of their main concerns was to determine how the rate of convergence of the Markov chain to its stationary distribution depended on s^*, n, p, θ, the prior, and the various assumed properties of the matrix X. Thy noted a dearth of literature on this topic.

The rate of convergence for a time reversible Markov chain $\{Z_n\}$ with transition matrix P depends critically on the gap between the largest and second largest singular of P. YWJ employed the path method developed by Diaconis and Stroock (1991) and Sinclair (1992) to provide lower bounds for the size of the eigengap, which translated easily into bounds on the mixing times, the number of steps of the chain $\{Z_n\}$ needed before the total variation distance between π and the distribution of Z_n becomes smaller than any specified $\epsilon > 0$.

The YWJ chains ran on a state space whose elements they identified with subsets of columns of X. More precisely, they assumed a prior distribution concentrated on a set $\mathcal{S} = \{S \in \{0, 1\}^p : |S| \leq s_0\}$, that is it concentrated on vectors b in $\mathbb{R}^p$ for which the size of $\{j \in [p] : b_j \neq 0\}$ was at most some (suitably small) s_0.

The restriction to small sets of columns was natural, given the assumption of a sparse θ for the model assumed to generate Y. However it had some unfortunate complicating effects on construction of the Markov chain and the paths that determine the mixing rate. The main difficulties arose for sets S on the "boundary" of $\mathcal{S}$ as a subset of $\{0,1\}^p$, the sets of size s_0.

In this note we describe a modification of the YWJ approach that eliminates the difficulties caused by the boundary. Instead of reanalyzing the YWJ problem (which would actually improve the mixing rate), we illustrate our approach by an application to the method of aggregation described by Rigollet and Tsybakov (2012). The setting is similar to that considered by YWJ. The observed Y is modeled as a sparse linear combination Xb plus a noise vector ϵ. The estimator for the mean is taken as a convex combination $\sum_S \pi(S)\Phi_S Y$ of least squares estimators, where Φ_S denotes the matrix for orthogonal projection onto the subspace of $\mathbb{R}^n$ spanned by the columns of the $n \times |S|$ submatrix $X_S = (X_j : j \in S)$ of X. The vector π is defined to be of the form

$$\pi(S) \propto \mu(S) \exp\left(-\frac{\|(I-\Phi_S)Y\|^2 + 2\text{trace}(\Phi_S)}{\beta}\right).$$

For $\beta = 2$, the vector π can be interpreted as a posterior distribution on the space of least squares projections $\{\Phi_S Y\}_{S\subset[p]}$.

We propose a time-reversible Markov chain with state space $\mathcal{S}$ equal to the whole of $\{0,1\}^p$ with π as its unique stationary distribution. Our main result is stated in Section 2.5. Roughly speaking, we show that:

> *Under suitable regularity conditions on the design matrix X and noise distribution, if we start our proposed Markov chain from a well chosen initial state $\widehat{T}$, then the $\epsilon-$mixing time of our Markov chain*
>
> $$\tau_\epsilon(\widehat{T}) = \inf_t \left\{\|P^t(\widehat{T}) - \pi\|_{TV} \le \epsilon, \forall t' \ge t\right\}.$$
>
> *is bounded by a constant multiple of $s^{*2} p \log p$.*

Compare with Theorem 2 of YWJ that gave a mixing time of the order $O(s_0^2 p(n +$

$s_0) \log p$). We are able to speed up the convergence of the Markov chain by avoiding the hard boundary of the state space. Instead we use a chain that encourages jumps from non-sparse S to a $\widehat{T}$ that is suitably close to the unknown T. One possible $\widehat{T}$ could be the thresholded lasso estimator studied by Zhou (2010). We also start the chain from $\widehat{T}$. Together these modifications lead to both a simpler analysis and a faster rate of convergence than the YWJ chain.

2.2 Metropolis-Hastings

Suppose π is a probability measure defined on a finite set $\mathcal{S}$. There are simple ways to construct a Markov chain via a transition matrix P for which π is the unique stationary distribution by virtue of the time reversibility condition

$$\pi(S)P(S, S') = \pi(S')P(S', S).$$

Equivalently, P corresponds to a random walk on a graph with edge weights $Q(\mathfrak{e})$ for $\mathfrak{e} = \{S, S'\}$ such that

$$\pi(S) = \sum_{S'} Q\{S, S'\} \qquad \text{and} \qquad P(S, S') = Q\{S, S'\}/\pi(S).$$

For the Metropolis-Hastings method one starts with a "proposal chain" given by a transition matrix R then defines P via acceptance/rejection of the proposal. For distinct S and S' one defines

$$P(S, S') = R(S, S') \min\left\{1, \frac{\pi(S')R(S', S)}{\pi(S)R(S, S')}\right\}$$

with $P(S,S)$ defined so that $\sum_{S'} P(S,S') = 1$. The edge weight then becomes

$$Q\{S,S'\} = \min\{\pi(S)R(S,S'), \pi(S')R(S',S)\} \qquad \text{for } S \neq S'.$$

Provided P is irreducible and aperiodic, it has eigenvalues $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_N > -1$ where N is the cardinality of the state space $\mathcal{S}$. Diaconis and Stroock (1991, Proposition 3) proved that, for a P-chain started in state S_0,

$$\left\| P^t(S_0, \cdot) - \pi(\cdot) \right\|_{\text{TV}} \leq \tfrac{1}{2}\pi(S_0)^{-1/2}\beta^t \qquad \text{where } \beta := \min(|\lambda_2|, |\lambda_N|).$$

Here $\|P^n(S_0,\cdot) - \pi(\cdot)\|_{\text{TV}} = \sum_S |P^n(S_0,S) - \pi(S)|/2$.

The upper bound can be simplified by running the lazy version of the P-chain, with transition matrix $\widetilde{P} = (I_N + P)/2$, which has nonnegative eigenvalues $(1+\lambda_i)/2$. The corresponding β equals $(1+\lambda_2)/2 \leq e^{-(1-\lambda_2)/2}$, so that

$$\left\| \widetilde{P}^t(S_0, \cdot) - \pi(\cdot) \right\|_{\text{TV}} \leq \tfrac{1}{2}\pi(S_0)^{-1/2}e^{-t(1-\lambda_2)/2}.$$

The quantity $1 - \lambda_2$ is called the spectral gap for the matrix P, which we denote by $\textsc{gap}(P)$. It is traditional to invert the last bound to see that $\left\| \widetilde{P}^t(S_0, \cdot) - \pi(\cdot) \right\|_{\text{TV}} \leq \epsilon$ when

$$t \geq \tau_\epsilon(S_0) \geq \frac{2\log(1/2\epsilon) + \log(1/\pi(S_0))}{\textsc{gap}(P)}.$$

For both the YWJ problem and our aggregation example the challenge is to design chains for which $\textsc{gap}(P)$, does not decrease too rapidly to zero. We follow YWJ in using the path method of Sinclair (1992) to get an upper bound for $1/\textsc{gap}(P)$.

2.3 Paths

The path method provides a lower bound for the spectral gap of a transition matrix for a time reversible Markov chain on a finite state space $\mathcal{S}$. The method requires construction of a set of directed paths connecting different states, one path for each pair (S, S') with $S \neq S'$. The path $\gamma(S, S')$ connecting S and S' should consist of distinct elements $S_0 = S, S_1, \ldots, S_m = S'$ of the state space with edge weights $Q\{S_j, S_{j+1}\} > 0$ for each j. The path can also be thought of as a sequence of directed edges, (S_j, S_{j+1}) for $j = 0, \ldots, m$. The path has length $\textsc{len}(\gamma(S, S')) = m$. The loading of an edge $\mathfrak{e}$ in $\mathcal{S} \times \mathcal{S}$ is defined as

$$\rho(\mathfrak{e}) = \sum_{\gamma(S,S')\ni\mathfrak{e}} \pi(S)\pi(S')/Q(\mathfrak{e})$$

where the sum runs over all paths γ with $\mathfrak{e}$ as one of the edges. Sinclair (1992, Corollary 6) showed that

$$1/\textsc{gap}(P) \leq \left(\max_{S,S'} \textsc{len}((\gamma(S, S'))\right) \times \left(\max_{\mathfrak{e}} \rho(\mathfrak{e})\right). \tag{2.1}$$

It is important to note that the paths are a theoretical construct that can depend on information about a Markov chain not known to the MCMC practitioner. For example, the paths defined by YWJ were allowed to depend on the unknown mean $X\theta$ for the $N(X\theta, \sigma^2 I_n)$ distribution that generates Y. Indeed they designed paths that involved knowledge of the support set $T = \{j : \theta_j \neq 0\}$.

For the YWJ implementation of the Sinclair method one first defines a map $\mathcal{G} : \mathcal{S}\backslash\{T\} \to \mathcal{S}$ which decreases the Hamming distance to T: that is, $h(S, T) > h(\mathcal{G}(S), T)$ for each $S \in \mathcal{S}\backslash\{T\}$. One defines $\gamma(S, S')$ as the unique path joining S to S' with (undirected) edges that are all of the form $\{W, \mathcal{G}(W)\}$ for some W in $\mathcal{S}$. More descriptively, one first joins S to T by following the direction indicated by $\mathcal{G}$, and similarly for S'. If the $S \to T$ and $S' \to T$ paths first meet at state W then $\gamma(S, S')$ follows $\mathcal{G}$ from S to W then reverses the direction of $\mathcal{G}$ to get from W to S'. We adapt this construction to the aggregation setting.

2.4 Comparison of MCMC stategies

YWJ (equation A2 in the supplement) had a posterior distribution $\pi = \pi_Y$ for which

$$\pi(S) = \exp\left(G(S, Y) - m(S)\right)/\mathcal{Z}(Y)$$

with

$$G(S, Y) = -\frac{n}{2}\log\left(1 + g(1 - \|\Phi_S Y\|^2/\|Y\|^2)\right),$$

an increasing function of $\|\Phi_S Y\|^2$, and

$$m(S) = \kappa|S|\log p + |S|\log(1+g)/2,$$

a function that increases as S gets larger. The constants g and κ depended on the YWJ prior. The factor $\mathcal{Z}(Y)$ was just a normalizing constant.

For the aggregation problem, inspired by work by Castillo et al. (2015) and Gao et al. (2015) on posterior contraction in the setting of high dimensional linear regression, for S with size no more than $s_1 + s^*$, we set the weight $\mu(S)$ equal to $\exp(-D|S|\log p)$ for dimension penalization. For even larger S, we need a heavier penalization. Set

$$\mu(S) = \exp\left(-D|S|\log p - \frac{4n}{\beta}\mathbb{1}\{|S| > s_1 + s^*\}\right).$$

we have a similar form for the weights $\pi(S)$ as in YWJ, with

$$G(S, y) = \|\Phi_S Y\|^2/\beta \qquad \text{AND} \qquad m(S) = D|S|\log p + 2\text{trace}(\Phi_S)/\beta.$$

2.4.1 The YWJ proposal chain

For their proposal R-chain YWJ (page 2502) introduced two types of edges on their state space $\mathcal{S} = \{S \in \{0,1\}^p : |S| \le s_0\}$:

1. single flips that connect a state S to a state S' by changing either a single 1 to a 0 or a single 0 to a 1
2. double flips that change a single 1 to a 0 as well as changing a different 0 to a 1

The single flips join S to another state S' at Hamming distance 1 from S. The double flips join S to another state S' at Hamming distance 2 from S, with $h(S,\emptyset) = h(S',\emptyset)$. The YWJ proposal chain is designed as follows.

1. with probability $1/2$ move via a single flip from S to a site chosen uniformly at random from the set of neighbors in $\mathcal{S}$ at Hamming distance 1 from S
2. with probability $1/2$ move via a double flip from S to a site chosen uniformly at random from the set of neighbors in $\mathcal{S}$ at Hamming distance 2 from S

2.4.2 Our proposal chain

Our method uses only single flips and jumps to a state $\widehat{T}$, which we will soon assume has size at most s^*, with high $\mathbb{P}_\theta$ probability. Define $\mathcal{S}_k = \{S \in \mathcal{S} : |S| = k\}$. We replace the hard boundary at s_0 by a soft boundary at $s_1 = 3s^*$. Our proposal R-chain allows these moves:

1. If $|S| \le s_1$ and $S \neq \widehat{T}$ then move via a single flip from S to a site chosen uniformly at random from the set of neighbors at Hamming distance 1 from S.
2. If $|S| > s_1$ then, with probability $1/2$ move to $\widehat{T}$ and with probability $1/2$ move via a single flip from S to a site chosen uniformly at random from the set of neighbors at Hamming distance 1 from S.

3. For a move from $\widehat{T}$, with probability $1/2$ move via a single flip to a site chosen uniformly at random from the set of neighbors at Hamming distance 1 from S, and with probability $1/2$ first choose uniformly at random an integer k with $s_1 < k \leq p$, then jump to an S' chosen uniformly at random from $\mathcal{S}_k$.

It will be important to have $\pi(\widehat{T})$ not too small—the choice $\widehat{T} = \emptyset$ does not work for our approach. The steps of the chain involving $\widehat{T}$ are much easier to handle than the double flips of the YWJ method.

2.4.3 The YWJ paths

YWJ defined paths by means of a map $\mathcal{G} : \mathcal{S}\backslash\{T\} \to S$, constructed as follows.

1. If $S \supset T$ define $\mathcal{G}(S)$ by flipping a single 1 from $S\backslash\{T\}$ to a 0. (The choice of the particular bit from $S\backslash\{T\}$ is not important.)

2. If T is not a subset of S and $|S| < s_0$ change a 0 in $T\backslash S$ to a 1 by the single flip that gives the largest $||\Phi_{S'} X\theta||$.

3. If T is not a subset of S and $|S| = s_0$, let $\mathcal{S}(S)$ be the result of a double flip to an S' for which $||\Phi_{S'} X\theta||$ is the largest.

The general YWJ strategy is to first build S up to a superset of T by single flips then reduce to T by single flips. The sets S on the boundary complicate the idea because a flip of a single 0 would lead to a $\mathcal{G}(S)$ outside the state space. The double flips are needed to keep $|\mathcal{G}(S)| \leq s_0$. The map $\mathcal{G}$ takes an S on the boundary to another S' on the boundary.

2.4.4 Our choice of paths

We follow YWJ in constructing paths by means of a a map $\mathcal{G} : \mathcal{S}\backslash\{T\} \to \mathcal{S}$, but with a slightly different choice for $\mathcal{G}$. Our choice avoids the difficulties with the boundary.

Define $\mathcal{U} = \{S \in \mathcal{S} : |S\backslash T| \leq s_1\}$. Notice that $|S| > s_1$ for each S in $\mathcal{S}\backslash\mathcal{U}$.

1. If $S \supset T$ and $S \in \mathcal{U}$ define $\mathcal{G}(S)$ by flipping a single 1 from $S\backslash\{T\}$ to a zero. (The choice of the particular bit from $S\backslash\{T\}$ is not important.)

2. If T is not a subset of S and $S \in \mathcal{U}$ change a 0 in $T\backslash S$ to a 1 by the single flip that gives the largest $\|\Phi_{S'}X\theta\|$.

3. If $S \in \mathcal{S}\backslash\mathcal{U}$ define $\mathcal{G}(S) = \widehat{T}$.

Our $\mathcal{G}$ also ensures that $h(\mathcal{G}(S), T) < h(S, T)$ for all $S \in \mathcal{S}\backslash\{T\}$.

2.5 Our main theorem

Remember that the aggregation weights are given by

$$\pi(S) = \exp\left(G(S, Y) - m(S)\right)/\mathcal{Z}(Y)$$

with

$$G(S, Y) = \|\Phi_S Y\|^2/\beta \qquad \text{AND} \qquad m(S) = D|S|\log p + 2\text{trace}(\Phi_S)/\beta, \tag{2.2}$$

for positive constants β and D that need to be specified.

Define sets:

$$\mathcal{A}_n = \left\{|\widehat{T}| \le s_1 - s^*, \left\|(I - \Phi_{\widehat{T}})X\theta\right\|^2 \le cs^*\log p\right\}.$$

$$\mathcal{E}_n = \left\{\max_{|S|\le s_1, j\notin S}\left|\left\langle(I - \Phi_S)X_j, \epsilon\right\rangle\right|^2 \le nL\nu\log p, \|\epsilon\|^2 \le 2n\right\}.$$

We need both $\mathcal{A}_n$ and $\mathcal{E}_n$ to occur with high probability. YWJ assumed that

$$\mathbb{E}_\theta \max_{|S|\le s_0, j\notin S}\left|\left\langle(I - \Phi_S)X_j, \epsilon\right\rangle\right| \le \sqrt{nL\nu\log p}/2,$$

which ensures that the first condition in $\mathcal{E}_n$ occurs with probability at least $1-\exp(-L\nu\log p/8)$. See Section 2.9 for assumptions that ensure $\mathcal{A}_n$ occurs with high probability.

Theorem 2.1. *Assume*

1. *All columns of the design matrix have length* $\sqrt{n}$*. There exists a constant* $\nu > 0$ *such that*

$$\|X_S w\|^2 \geq n\nu \|w\|^2 \qquad \text{for each } w \in \mathbb{R}^S \text{ and } S \subset [p] \text{ with } |S| \leq s_1 + s^*.$$

2. $\min_{j \in T} |\theta_j|^2 \geq \theta_{\min}^2 \geq (8\beta D \log p)/(n\nu^2)$.

If D is chosen to be greater than $4 + (4L + 2c)/\beta$*, then on the set* $\mathcal{A}_n \cap \mathcal{E}_n$*, we have*

$$1/\textsc{gap}(P) \leq 12p(s_1 + s^*).$$

Corollary 2.1. *Under the assumptions of Theorem 2.1, on the set* $\mathcal{A}_n \cap \mathcal{E}_n$*,*

$$\tau_\epsilon(\widehat{T}) \leq 12p(s_1 + s^*)\left(\log \frac{1}{2\epsilon} + 2Ds^* \log p\right).$$

YWJ assumption B required $\max_{|S| \leq s_0} \lambda_{\min}(X_S^T X_S/n) \geq \nu$, which is equivalent to our assumption 1 except that they had the much larger s_0 in place of our $s_1 + s^*$. YWJ needed the larger s_0 value to accommodate their double flips. We are able to weaken their assumption by avoiding the difficulty around the boundary of the state space.

2.6 Proof of Theorem 2.1

As with the YWJ construction, our map $\mathcal{G}$ (described in Section 2.4.4) defines a directed tree on $\mathcal{S}$ with T as the root. Following YWJ, we denote the collection of states on the subtree with root S (including S) as $\Lambda(S)$. That is, $\Lambda(S)$ consists of all S' for which the $\mathcal{G}$-path from S' to T passes through S.

To bound the length of the longest path, notice that the length of a path from a state I to a state F is at most the length of $\gamma_{I,T}$ and $\gamma_{T,F}$ combined. For all $S \in \mathcal{U}$, $h(S,T) \le 3s^*$. All other states are pulled to $\widehat{T} \in \mathcal{U}$ in one step. It follows that

$$\max_{S,S'} \textsc{len}(\{\gamma(S,S')) \le 2(s_1 + s^*).$$

The argument for the loadings requires more work. It is easy to see that a path from I to F passes through a directed edge $\mathfrak{e} = (S,S')$ if and only if $I \in \Lambda(S)$ and $F \notin \Lambda(S)$. Therefore the loading of edge $\mathfrak{e}$ can be written as

$$\rho(\mathfrak{e}) = \frac{\sum_{I,F} \pi(I)\pi(F) : \mathfrak{e} \in \gamma\{I,F\}}{Q(\mathfrak{e})} = \frac{\pi(\Lambda(S))(1 - \pi(\Lambda(S)))}{\pi(S)P(S,S')}. \tag{2.3}$$

We need an upper bound for the ratio $\pi(\Lambda(S)/\pi(S))$ and a lower bound for $P(S,S')$. To that end we will establish three inequalities:

(a) $P(S,\mathcal{G}(S)) \ge 1/(2p)$ for each S in $\mathcal{S}\backslash\{T\}$.

(b) If $S \in \mathcal{S}\backslash\mathcal{U}$ then $\pi(\Lambda(S)/\pi(S)) \le 1$.

(c) If $S \in \mathcal{U}^c$ then $\pi(\Lambda(S)/\pi(S)) \le 3$.

From these three facts it follows that

$$\rho(\mathfrak{e} = \{S,S'\}) \le \frac{\pi(\Lambda(S))}{\pi(S)} \cdot \frac{1}{P(S,S')} \le 6p$$

so that, by (2.1),

$$1/\textsc{gap}(P) \le 2(s_1 + s^*) \times 6p.$$

Our proofs of the claims requires control of the ratio $\pi(S)/\pi(S')$ for various pairs S, S'. The necessary facts are contained in the following lemma, whose proof appears in Section 2.7. It is there that the main technical differences between the YWJ argument and ours appear.

Lemma 2.1. *Under the assumptions of Theorem 2.1, for each S in $\mathcal{S}\backslash\{T\}$,*

$$\frac{\pi(S)}{\pi(\widehat{T})} \leq \exp\left(-D(|S|-2s^*)\log p + \frac{4s^*}{\beta} + \frac{\left(\sqrt{cs^*} + \sqrt{L|S|}\right)^2 \log p}{\beta}\right) \tag{2.4}$$

and

$$\frac{\pi(S)}{\pi(\mathcal{G}(S))} \leq \begin{cases} \exp\left(-\frac{1}{2}D\log p\right) & \textit{if } S \supset T \text{ and } S \in \mathcal{U} & (2.5) \\ \exp\left(-D\log p + 2/\beta\right) & \textit{if } T\backslash S \neq \emptyset \text{ and } S \in \mathcal{U} & (2.6) \\ \exp\left(-2|S|\log p\right) & \textit{if } S \in \mathcal{U}^c. & (2.7) \end{cases}$$

Proof of claim (a). We claim that $P(S, \mathcal{G}(S)) = R(S, \mathcal{G}(S))$ for all S. Recall that the R matrix contains our proposal probabilities in the Metropolis-Hasting algorithm. We are claiming that all proposals from S to $\mathcal{G}(S)$ get accepted with probability 1. This is not true for every edge in the Markov chain. But when constructing $\mathcal{G}$ and the paths, we deliberately chose to only use the "clear acceptance" edges to bear the weight. For $S \in \mathcal{U}$, because the proposal from $S \in \mathcal{U}$ to $\mathcal{G}(S) \in \mathcal{U}$ is symmetric, we have

$$P(S, \mathcal{G}(S)) = R(S, \mathcal{G}(S)) \min\left\{1, \frac{\pi(\mathcal{G}(S))}{\pi(S)}\right\}.$$

By lemma 2.1, we have $\pi(\mathcal{G}(S)) \geq \pi(S)$ for all $S \in \mathcal{U}$. Therefore the proposal always gets accepted.

For $S \in \mathcal{S}\backslash\mathcal{U}$, $\mathcal{G}(S) = \widehat{T}$. We have asymmetric proposal probabilities between S and $\widehat{T}$.

$$P(S, \widehat{T}) = \frac{1}{2}\min\left\{1, \frac{\pi(\widehat{T})R(\widehat{T}, S)}{\pi(S)R(S, \widehat{T})}\right\}.$$

From our proposal scheme, $R(\widehat{T}, S) = \frac{1}{2}\left((p-2s^*)\binom{p}{|S|}\right)^{-1}$ and $R(S, \widehat{T}) = 1/2$. From (2.4), we have

$$\frac{\pi(\widehat{T})R(\widehat{T},S)}{\pi(S)R(S,\widehat{T})} \geq \exp\left(|S|\log p - |S| - \log p\right) \geq 1.$$

Therefore $P(S,\widehat{T}) = 1/2$. For all $S \in \mathcal{U}$, it is easy to see that $R(S,\mathcal{G}(S)) \geq 1/(2p)$. Deduce that $P(S,\mathcal{G}(S)) \geq 1/(2p)$ for all $S \neq T$.

□

Proof of claim (b). In this case S is a leaf node on the tree, and $\Lambda(S) = \{S\}$. Therefore $\pi(\Lambda(S))/\pi(S) = 1$. □

Proof of claim (c). Split the set $\Lambda(S)$ in two parts: $\Lambda(S) \cap \mathcal{U}$ and $\Lambda(S)\backslash\mathcal{U}$. If the second part is nonempty, then $\widehat{T}$ must also be an offspring of (or is equal to) S, in which case $\pi(S) \geq \pi(\widehat{T})$. We have

$$\frac{\pi(\Lambda(S))}{\pi(S)} \leq \frac{\pi(\Lambda(S)\cap\mathcal{U})}{\pi(S)} + \frac{\pi(\mathcal{U}^c)}{\pi(\widehat{T})}.$$

The set $\Lambda(S) \cap \mathcal{U}$ can be split into layers on the tree. Denote $\mathcal{S}_k(S) = \{\widetilde{S} \in \mathcal{S} : d_H(S,\widetilde{S}) = k\}$.

$$\begin{aligned}
\frac{\pi(\Lambda(S)\cap\mathcal{U})}{\pi(S)} &= \sum_{k\geq 0} \frac{\pi(\Lambda(S)\cap\mathcal{U}\cap\mathcal{S}_k(S))}{\pi(S)} \\
&\leq \sum_{k\geq 0} |\Lambda(S)\cap\mathcal{U}\cap\mathcal{S}_k(S)| \exp\left(-\frac{Dk\log p}{2}\right) \\
&\leq \sum_{k\geq 0} \exp\left(k + k\log p - \frac{Dk\log p}{2}\right) \\
&= \frac{1}{1-\exp\left(-(D/2-1)\log p + 1\right)} \leq 2.
\end{aligned}$$

(2.7) implies

$$\begin{aligned}\frac{\pi(\mathcal{U}^c)}{\pi(\widehat{T})} &= \sum_{k>2s^*} \frac{\pi((\mathcal{S}\backslash\mathcal{U}) \cap \mathcal{S}_k)}{\pi(\widehat{T})} \\ &\leq \sum_{k>2s^*} \exp\left(-k \log p + k\right) \leq 1.\end{aligned}$$

Deduce that $\pi(\Lambda(S))/\pi(S) \leq 3$ for $S \in \mathcal{U}$.

□

2.7 Proof of Lemma 2.1

Each of the assertions of the lemma relies on a simple consequence of assumption 1 from Theorem 2.1. If A and B are disjoint subsets of $[p]$ with $|A| + |B| \leq s_1 + s^*$ then

$$\|X_A r + X_B t\|^2 \geq n\nu(\|r\|^2 + \|t\|^2)$$

for all $r \in \mathbb{R}^A$ and $b \in \mathbb{R}^B$. If we choose r so that $X_A r = -\Phi_A X_B t$ then ignore the $\|r\|^2$ on the right-hand side we get

$$\|(I - \Phi_A)X_B t\| \geq n\nu\|t\|^2 \qquad \text{for each } t \text{ in } \mathbb{R}^B,$$

which implies that the smallest singular value of the $n \times |B|$ matrix $(I - \Phi_A)X_B$ is no less than $\sqrt{n\nu}$. It follows that

$$\left\|X_B^T(I - \Phi_A)X_B t\right\|^2 \geq (n\nu)^2\|t\|^2 \qquad \text{for each } t \text{ in } \mathbb{R}^B. \tag{2.8}$$

Proof of (2.5). By construction $\mathcal{G}(S) = S\backslash\{j\}$ for some j in $S\backslash T$ and

$$\begin{aligned} G(S,Y) - G(\mathcal{G}(S),Y) &= \left(\|\Phi_S Y\|^2 - \left\|\Phi_{\mathcal{G}(S)} Y\right\|^2\right)/\beta \\ &= \left\|\left(\Phi_S - \Phi_{\mathcal{G}(S)}\right) Y\right\|^2/\beta = \left\|\left(\Phi_S - \Phi_{\mathcal{G}(S)}\right)\epsilon\right\|^2/\epsilon \end{aligned}$$

because $\Phi_{\mathcal{G}(S)}X\theta = \Phi_S X\theta = X\theta$. The difference $\Phi_S - \Phi_{\mathcal{G}(S)}$ projects orthogonally onto the subspace spanned by $z = (I - \Phi_{\mathcal{G}(S)})X_j$, so that

$$\left(\Phi_S - \Phi_{\mathcal{G}(S)}\right)\epsilon = \langle z, \epsilon\rangle/\|z\|.$$

Inequality (2.8) with $A = \mathcal{G}(S)$ and $B = \{j\}$ gives

$$\|z\|^2 = \left\|(I - \Phi_{\mathcal{G}(S)})X_j\right\|^2 \geq n\nu.$$

Thus

$$\left\|\left(\Phi_S - \Phi_{\mathcal{G}(S)}\right)\epsilon\right\|^2 \leq \left|\left\langle (I - \Phi_{\mathcal{G}(S)})X_j, \epsilon\right\rangle\right|^2/(n\nu),$$

which is bounded above by $L\log p$ on the set $\mathcal{E}_n$.

The dimension penalization contributes

$$m(S) - m(\mathcal{G}(\mathcal{S})) = D\log p + \frac{2}{\beta}\left(\text{trace}(\Phi_S) - \text{trace}(\Phi_{\mathcal{G}(S)})\right) \geq D\log p.$$

Deduce that

$$\begin{aligned} \frac{\pi(S)}{\pi(\mathcal{G}(S))} &= \exp\left(G(S,Y) - G(\mathcal{G}(S),Y) - (m(S) - m(\mathcal{G}(\mathcal{S})))\right) \\ &\leq \exp\left(-D\log p + \frac{L\log p}{\beta}\right) \\ &\leq \exp\left(-\frac{D\log p}{2}\right) \qquad \text{if } D > \frac{2L}{\beta}. \end{aligned}$$

□

Proof of (2.6). Here $T\backslash S$ is nonempty. By construction $\mathcal{G}(S) = S \cup \{j_S\}$ where $j_S = \arg\max_{j\in T\backslash S} \left\|\Phi_{S\cup\{j_S\}}X\theta\right\|$.

This time $\Phi_{S\cup\{j\}} - \Phi_S$ projects orthogonally onto the space spanned by $z_j := (I - \Phi_S)X_j$. Once again inequality (2.8) implies that $\|z_j\|^2 \geq n\nu$ for each j in $T\backslash S$, so that

$$\left\|\left(\Phi_{S\cup\{j\}} - \Phi_S\right)\epsilon\right\|^2 \leq \left|\left\langle\left(I - \Phi_{\mathcal{G}(S)}\right)X_j, \epsilon\right\rangle\right|^2 /(n\nu) \qquad \text{for } j \in T\backslash S.$$

In particular, with $j = j_S$, we get

$$\left\|\left(\Phi_{\mathcal{G}(S)} - \Phi_S\right)\epsilon\right\|^2 \leq L\log p \qquad \text{on the set } \mathcal{E}_n. \tag{2.9}$$

To control the G contribution to $\pi(S)/\pi(\mathcal{G}(S))$ we first note that

$$\left\|\left(\Phi_{\mathcal{G}(S)} - \Phi_S\right)Y\right\| \geq \left\|\left(\Phi_{\mathcal{G}(S)} - \Phi_S\right)X\theta\right\| - \left\|\left(\Phi_{\mathcal{G}(S)} - \Phi_S\right)\epsilon\right\|. \tag{2.10}$$

YWJ (Lemma 8) showed that, under our assumption 1,

$$\left\|\left(\Phi_{\mathcal{G}(S)} - \Phi_S\right)X\theta\right\| \geq \sqrt{n\nu}\min_{j\in T}|\theta_j|. \tag{2.11}$$

For completeness we will prove why that is true. By the choice of j_S,

$$\begin{aligned}\left\|\left(\Phi_{\mathcal{G}(S)} - \Phi_S\right)X\theta\right\|^2 \geq& \frac{1}{|T\backslash S|}\sum_{j\in T\backslash S}\left\|\left(\Phi_{S\cup\{j\}} - \Phi_S\right)X\theta\right\|^2\\ \geq& \frac{1}{|T\backslash S|}\sum_{j\in T\backslash S}|\langle z_j, X\theta\rangle|^2/\|z_j\|^2.\end{aligned} \tag{2.12}$$

To simplify notation write B for $T\backslash S$. We bound $\|z_j\|^2$ from above by $\|X_j\|^2 = n$. The inner product term simplifies slightly because

$$(I - \Phi_S)X\theta = (I - \Phi_S)X_B\theta_B,$$

The projection killing off contributions from X_j's with $j \in S$. The z_j's can be thought of as the columns of the matrix $Z_B = (I - \Phi_S)X_B$. The sum $\sum_{j\in B} z_j z_j^T$ equals $Z_B Z_B^T$. Thus

$$\begin{aligned}\sum_{j\in B} \langle z_j, X_B\theta_B\rangle^2 =& (X_B\theta_B)^T Z_B Z_B^T X_B\theta_B \\ =& \left\| X_B^T (I - \Phi_S) X_B\theta_B \right\|^2 \\ \geq& (n\nu)^2 \|\theta_B\|^2 \qquad \text{by inequality (2.8).} \qquad (2.13)\end{aligned}$$

Inequalities (2.12) and (2.13) combine to give

$$\left\| (\Phi_{\mathcal{G}(S)} - \Phi_S)\, X\theta \right\|^2 \geq \frac{n^2\nu^2 \|\theta_B\|^2}{n|B|} \geq n\nu^2 \min_{j\in B} \theta_j^2,$$

which implies (2.11).

Together, inequalities (2.9), (2.10) and (2.11) imply

$$g(S, Y) - g(\mathcal{G}(S), Y) \leq -\left(\sqrt{8\beta D \log p} - \sqrt{L \log p} \right)^2 / \beta \leq -2D \log p.$$

We also have

$$m(S) - m(\mathcal{G}(S)) = -D \log p - \frac{2}{\beta} (\mathrm{trace}(\Phi_S) - \mathrm{trace}(\Phi_{\mathcal{G}(S)})) \geq -D \log p - \frac{2}{\beta}.$$

Deduce that

$$\frac{\pi(S)}{\pi(\mathcal{G}(S))} \leq \exp\left(-D \log p + \frac{2}{\beta} \right).$$

□

Proof of (2.4). Write

$$G(S, Y) - G(\widehat{T}, Y) = \frac{1}{\beta}\left(\|\Phi_S Y\|^2 - \left\|\Phi_{\widehat{T}} Y\right\|^2\right).$$

The set $\widehat{T}$ may not be contained in the set S. We use a little trick to remedy the problem.

$$\begin{aligned}
&\|\Phi_S Y\|^2 - \|\Phi_{\widehat{T}} Y\|^2 \\
&\leq \left\|\Phi_{S\cup\widehat{T}} Y\right\|^2 - \|\Phi_{\widehat{T}} Y\|^2 \\
&= \left\|\left(\Phi_{S\cup\widehat{T}} - \Phi_{\widehat{T}}\right) Y\right\|^2 \\
&\leq \left(\left\|\left(\Phi_{S\cup\widehat{T}} - \Phi_{\widehat{T}}\right) X\theta\right\| + \left\|\left(\Phi_{S\cup\widehat{T}} - \Phi_{\widehat{T}}\right)\epsilon\right\|\right)^2 \\
&\leq \left(\left\|\left(I - \Phi_{\widehat{T}}\right) X\theta\right\| + \left\|\left(\Phi_{S\cup\widehat{T}} - \Phi_{\widehat{T}}\right)\epsilon\right\|\right)^2.
\end{aligned}$$

On the set $\mathcal{A}_n$, we have $\left\|\left(I - \Phi_{\widehat{T}}\right) X\theta\right\|^2 \leq cs^* \log p$.

For the ϵ term, if $|S| > s_1 + s^*$, we bound it by $\|\epsilon\| \leq \sqrt{2n}$. If $|S| \leq s_1 + s^*$, we break the term into a sum. Suppose $(S \cup \widehat{T})\backslash\widehat{T} = \{k[1], ..., k[m]\}$, where $m \leq |S|$. Temporarily write B_j for $\widehat{T} \cup \{k[1], ..., k[j]\}$, with $B_0 = \widehat{T}$. Define $z_j := (I - \Phi_{B_{j-1}})X_{k[j]}$. Then $\Phi_{B_{m-1}} - \Phi_{B_0}$ projects orthogonally onto the subspace spanned by the orthogonal vectors $z_1, ..., z_m$. It follows that

$$\left\|(\Phi_{B_{m-1}} - \Phi_{B_0})\,\epsilon\right\|^2 = \sum_{j=1}^{m} \langle z_j, \epsilon\rangle^2 / \left\|z_j\right\|^2.$$

Each summand can be handled in almost the same way as for the proof of (2.5), leading to the bound

$$\left\|\left(\Phi_{S\cup\widehat{T}-\Phi_{\widehat{T}}}\right)\epsilon\right\| \leq \sqrt{L|S|\log p} + \sqrt{2n}\mathbb{1}\left\{|S| > s_1 + s^*\right\}.$$

The dimension penalization terms give

$$\begin{aligned}&m(S) - m(\mathcal{G}(S))\\ =&D(|S| - |\widehat{T}|)\log p + \frac{2}{\beta}(\mathrm{trace}(\Phi_S) - \mathrm{trace}(\Phi_{\widehat{T}})) + \frac{4n}{\beta}\mathbb{1}\{|S| > s_1 + s^*\}\\ \geq&D(|S| - 2s^*)\log p - \frac{4s^*}{\beta} + \frac{4n}{\beta}\mathbb{1}\{|S| > s_1 + s^*\}.\end{aligned}$$

We have

$$\frac{\pi(S)}{\pi(\widehat{T})} \leq \exp\left(-D(|S| - 2s^*)\log p + \frac{4s^*}{\beta} + \frac{\left(\sqrt{cs^*} + \sqrt{L|S|}\right)^2 \log p}{\beta}\right)$$

for all $S \in \mathcal{S}$.

□

Proof of (2.7). For $S \in \mathcal{U}^C$, the size of S is larger than $s_1 = 3s^*$. Apply (2.4) to deduce that

$$\begin{aligned}\frac{\pi(S)}{\pi(\widehat{T})} \leq& \exp\left(-\left(D - \frac{2L}{\beta}\right)|S|\log p + \left(2D + \frac{2c}{\beta}\right)s^*\log p + \frac{4s^*}{\beta}\right)\\ \leq& \exp\left(-\left(\frac{D}{3} - \frac{2L}{\beta} - \frac{2c}{3\beta} - \frac{4}{3\beta}\right)|S|\log p\right).\end{aligned}$$

Choose $D > (6L + 2c + 6\beta + 4)/\beta$ so that $\pi(S)/\pi(\widehat{T}) \leq \exp(-|S|\log p)$. □

2.8 Proof of Corollary 2.1

Theorem 2.1 provides us with a bound on the eigengap of P. We only need to show that $\pi(\widehat{T})$ is not too small.

Recall that $\mathcal{S}_k = \{S \in \mathcal{S} : |S| = k\}$, from (2.4),

$$\frac{\pi(\mathcal{N}_k)}{\pi(\widehat{T})} \leq \exp\left(k \log p + \frac{4s^*}{\beta} - D(k - 2s^*) \log p + \frac{(2cs^* + 2Lk) \log p}{\beta}\right).$$

If $k > s_1$, we could use the assumption $D > (6L + 2c + 6\beta + 4)/\beta$ to further bound the ratio by $\exp(-k \log p)$. Deduce that

$$\frac{\pi(\cup_{k>s_1} \mathcal{N}_k)}{\pi(\widehat{T})} \leq \sum_{k>s_1} \exp(-k \log p) \leq \exp(-s_1 \log p). \tag{2.14}$$

If $k \leq s_1$, we have

$$\frac{\pi(\mathcal{N}_k)}{\pi(\widehat{T})} \leq \exp\left(\left(2D + \frac{2c}{\beta}\right) s^* \log p + \frac{4s^*}{\beta}\right). \tag{2.15}$$

Combine (2.14) and (2.15) and it follows that

$$\begin{aligned} \log\left(\frac{1}{\pi(\widehat{T})}\right) &= \log\left(\sum_{k=0}^{p} \frac{\pi(\mathcal{N}_k)}{\pi(\widehat{T})}\right) \\ &\leq \log\left(\exp(-s_1 \log p) + s_1 \exp\left(\left(2D + \frac{2c}{\beta}\right) s^* \log p + \frac{4s^*}{\beta}\right)\right) \\ &\leq 3Ds^* \log p \qquad \text{for } D \text{ large enough.} \end{aligned}$$

2.9 Choice for the initializer

Theorem 2.1 holds for all initializers in the set

$$\mathcal{A}_n = \left\{|\widehat{T}| \leq s_1 - s^*, \left\|(I - \Phi_{\widehat{T}})X\theta\right\|^2 \leq cs^* \log p\right\}.$$

When s_1 is taken to be a constant multiple of s^*, Zhou (2010) showed that the thresholded LASSO estimator falls in $\mathcal{A}_n$ with high probability under mild assumptions. For completeness we will give the form of the estimator and the proof for controlling its prediction risk here.

Write λ_n for $\sqrt{\log p/n}$. The LASSO estimator is defined as

$$\widehat{\theta} = \arg\min_t \|Y - Xt\|^2 + \alpha\lambda_n\|t\|_1.$$

Define $\delta = \widehat{\theta} - \theta$. Bickel, Ritov, and Tsybakov (2009, Theorem 7.2) showed that under the restricted eigenvalue condition $\kappa = \kappa(s^*, 3) > 0$, on a set with probability at least $1 - p^{1-\alpha^2/32}$,

$$\|\delta\|_1 \leq \frac{8\alpha\lambda_n}{\kappa^2}s^*, \quad \|\delta_T\| \leq \frac{2\alpha\lambda_n}{\kappa^2}\sqrt{s^*}. \tag{2.16}$$

We define $\widehat{T}$ to be $\{j : |\widehat{\theta}_j| > 8\alpha\lambda_n/\kappa^2\}$. The following theorem restates a result of Zhou (2010, Theorem 1.3). It provides a theoretical guarantee that the event $\mathcal{A}_n$ occurs with high probability for this choice of $\widehat{T}$.

Theorem 2.2. *Under the REC*$(s^*, 3)$ *condition, on a set with probability at least* $1 - p^{1-\alpha^2/32}$, *we have* $|\widehat{T}| \leq 2s^*$ *and*

$$\left\|(I - \Phi_{\widehat{T}})X\theta\right\| \leq \frac{2\alpha\sqrt{\Lambda_{\max}(s^*)}}{\kappa^2}\sqrt{s^*\log p},$$

where $\Lambda_{\max} = \max_{|S|\leq s^*}\lambda_{\max}(X_S^TX_S/n)$*.*

Proof. To handle the size of $\widehat{T}$, note that

$$|\widehat{T}| = |\widehat{T} \cap T| + |\widehat{T}\backslash T|.$$

The first of the inequalities in (2.16) implies that

$$\frac{8\alpha\lambda_n}{\kappa^2}s^* \geq \|\delta\|_1 \geq \sum_{j\in\widehat{T}\backslash T}|\delta_j| > |\widehat{T}\backslash T| \cdot \frac{8\alpha\lambda_n}{\kappa^2},$$

the last inequality comes from the fact that $|\delta_j| = |\widehat{\theta}_j| > 8\alpha\lambda_n/\kappa^2$ for all $j \in \widehat{T}\backslash T$. It follows that $|\widehat{T}\backslash T| \le s^*$, and therefore $\widehat{T} \le |T| + s^* \le 2s^*$.

For the prediction risk:

$$\left\|(I - \Phi_{\widehat{T}})X\theta\right\| = \left\|(I - \Phi_{\widehat{T}})X_{T\backslash\widehat{T}}\theta_{T\backslash\widehat{T}}\right\| \le \left\|X_{T\backslash\widehat{T}}\theta_{T\backslash\widehat{T}}\right\| \le \sqrt{n\Lambda_{\max}}\|\theta_{T\backslash\widehat{T}}\|.$$

From the second inequality in (2.16), we have $\|\theta_{T\backslash\widehat{T}}\| \le \|\delta_T\| \le 2\alpha\lambda_n\sqrt{s^*}/\kappa^2$. Conclude that

$$\left\|(I - \Phi_{\widehat{T}})X\theta\right\| \le \sqrt{n\Lambda_{\max}}\frac{2\alpha\lambda_n}{\kappa^2}\sqrt{s^*} = \frac{\sqrt{2\alpha\Lambda_{\max}(s^*)}}{\kappa^2}\sqrt{s^* \log p}.$$

□

Bibliography

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics 37*(4), 1705–1732. 42

Castillo, I., J. Schmidt-Hieber, A. Van der Vaart, et al. (2015). Bayesian linear regression with sparse priors. *Annals of Statistics 43*(5), 1986–2018. 27

Diaconis, P. and D. Stroock (1991). Geometric bounds for eigenvalues of markov chains. *The Annals of Applied Probability 1*(1), 36–61. 22, 25

Gao, C., A. W. van der Vaart, and H. H. Zhou (2015). A general framework for Bayes structured linear models. *arXiv:1506.02174*. 27

Rigollet, P. and A. B. Tsybakov (2012). Sparse estimation by exponential weighting. *Statistical Science 27*(4), 558–575. 23

Sinclair, A. (1992). Improved bounds for mixing rates of markov chains and multicommodity flow. In *Latin American Symposium on Theoretical Informatics*, Volume 583, pp. 474–487. Springer. 22, 25, 26

Yang, Y., M. J. Wainwright, and M. I. Jordan (2016). On the computational complexity of high-dimensional bayesian variable selection. *The Annals of Statistics 44*(6), 2497–2532. 21, 22

Zhou, S. (2010). Thresholded lasso for high dimensional variable selection and statistical estimation. Technical report, arXiv:1002.1583. 24, 41, 42

Chapter 3

Fair Quantile Regression

Joint work with Prof. John Lafferty and Prof. David Pollard

Abstract

Quantile regression is a tool for learning conditional distributions. In this chapter we study quantile regression in the setting where a protected attribute is unavailable when fitting the model. This can lead to "unfair" quantile estimators for which the effective quantiles are very different for the subpopulations defined by the protected attribute. We propose a procedure for adjusting the estimator on a heldout sample where the protected attribute is available. The main result of the chapter is an empirical process analysis showing that the adjustment leads to a fair estimator for which the target quantiles are brought into balance, in a statistical sense that we call $\sqrt{n}$-fairness. We illustrate the ideas and adjustment procedure on a dataset of 200,000 live births, where the objective is to characterize the dependence of the birth weights of the babies on demographic attributes of the birth mother; the protected attribute is the mother's race.

3.1 Introduction

Recent research on fairness has formulated interesting new perspectives on machine learning methodologies and their deployment, through work on definitions, axiomatic characterizations, case studies, and algorithms (Hardt et al., 2016; Dwork et al., 2012; Kleinberg et al., 2017; Chouldechova, 2017; Woodworth et al., 2017).

Much of the work on fairness in machine learning has been focused on classification, although the influential paper of Hardt et al. (2016) considers general frameworks that include regression. Just as the mean gives a coarse summary of a distribution, the regression curve gives a rough summary of a family of conditional distributions (Mosteller and Tukey, 1977). Quantile regression targets a more complete understanding of the dependence between a response variable and a collection of explanatory variables.

Given a conditional distribution $F_X(y) = \mathbb{P}(Y \leq y \,|\, X)$, the quantile function $q_\tau(X)$ is characterized by $F_X(q_\tau(X)) = \tau$, or $q_\tau(X) = F_X^{-1}(\tau) = \inf\{y : F_X(y) \leq \tau\}$. We consider the setting where an estimate $\widehat{q_\tau}(X)$ is formed using a training set $\{(X_i, Y_i)\}$ for which a protected attribute A is unavailable. The estimate $\widehat{q_\tau}(X)$ will often give quantiles that are far from τ, when conditioned on the protected variable. We study methods that adjust the estimator using a heldout sample for which the protected attribute A is observed.

As example, to be developed at length below, consider forecasting the birth weight of a baby as a function of the demographics and personal history of the birth mother, including her prenatal care, smoking history, and educational background. As will be seen, when the race of mother is excluded, the quantile function may be very inaccurate, particularly at the lower quantiles $\tau < 0.2$ corresponding to low birth weights. If used as a basis for medical advice, such inaccurate forecasts could conceivably have health consequences for the mother and infant. It would be important to adjust the estimates if the race of the mother became available.

In this chapter we study the simple procedure that adjusts an initial estimate $\widehat{q}_\tau(X)$ by adding $\widehat{\mu}_\tau A + \widehat{\nu}_\tau$, by carrying out a quantile regression of $Y - \widehat{q}_\tau(X)$ onto A. We show that this leads to an estimate $\widetilde{q}_\tau(X, A) = \widehat{q}_\tau(X) + \widehat{\mu}_\tau A + \widehat{\nu}_\tau$ for which the conditional quantiles are close to the target level τ for both subpopulations $A = 1$ and $A = 0$. This result follows from an empirical process analysis that exploits the special dual structure of quantile regression as a linear program. Our main technical result is that our adjustment procedure is $\sqrt{n}$-fair at the population level. Roughly speaking, this means that the effective quantiles for the two subpopulations agree, up to a stochastic error that decays at a parametric $1/\sqrt{n}$ rate. We establish this result using empirical process techniques that generalize to more general types of attributes, not just binary.

In the following section we provide technical background on quantile regression, including its formulation in terms of linear programming, the dual program, and methods for inference. We also provide background on notions of fairness that are related to this work and give our definition of fairness. In Section 3.3 we formally state the methods and results. The proof is given in Section 3.5. We illustrate these results on synthetic data and birth weight data in Section 3.6. We finish with a discussion of the results and possible directions for future work.

3.2 Background

In this section we review the essentials of quantile regression that will be relevant to our analysis. We also briefly discuss definitions of fairness.

3.2.1 Linear programming formulation

The formulation of quantile estimates as solutions to linear programs starts with the "check" or "hockey stick" function $\rho_\tau(u)$ defined by $\rho_\tau(u) = (\tau - 1)u\mathbb{1}\{u \le 0\} + \tau u \mathbb{1}\{u > 0\}$.

For the median, $\rho_{1/2}(u) = \frac{1}{2}|u|$. If $Y \sim F$ is a random variable, define $\widehat{\alpha}(\tau)$ as the solution to the optimization $\widehat{\alpha}(\tau) = \arg\min_a \mathbb{E}\rho_\tau(Y - a)$. Then the stationary condition is seen to be

$$0 = (\tau - 1)\int_{-\infty}^{\alpha} dF(u) + \tau \int_{\alpha}^{\infty} dF(u) = (\tau - 1)F(\alpha) + \tau(1 - F(\alpha)),$$

from which we conclude $\widehat{\alpha}(\tau) = F^{-1}(\tau)$ is the τ-quantile of F. Similarly the conditional quantile of Y given random variable $X \in \mathbb{R}^p$ can be written as the solution to the optimization $q_\tau(x) = \arg\min_q \mathbb{E}\left(\rho_\tau(Y - q) \,|\, X = x\right)$. For a linear estimator $q_\tau(X) = X^T\widehat{\beta}_\tau$, minimizing the empirical check function loss leads to a convex optimization $\widehat{\beta}_\tau = \arg\min_\beta \sum_{i\leq n} \rho_\tau(Y_i - X_i^T\beta)$. Dividing the residual $Y_i - X_i^T\beta$ into positive part u_i and negative part v_i yields the linear program

$$\min_{u,v\in\mathbb{R}^n, \beta\in\mathbb{R}^p} \tau\mathbb{1}^T u + (1 - \tau)\mathbb{1}^T v, \quad \text{such that } Y = X\beta + u - v,\ u \geq 0,\ v \geq 0.$$

The dual linear program is then formulated as

$$\max_b \ Y^T b \quad \text{such that } X^T b = (1 - \tau)X^T\mathbb{1},\ b \in [0, 1]^n. \tag{3.1}$$

When $n > p$, the primal solution is obtained from a set of p observations $X_h \in \mathbb{R}^{p\times p}$ for which the residuals are exactly zero, through the correspondence $\widehat{\beta}_\tau = X_h^{-1}Y_h$. The dual variables $\widehat{b}_\tau \in [0, 1]^n$, also known as regression rank scores, play the role of ranks. In particular, the quantity $\int_0^1 \widehat{b}_{\tau,i} d\tau$ can be interpreted as the quantile at which Y_i lies for the conditional distribution of Y given X_i (Gutenbrunner and Jurečková, 1992). As seen below, the stochastic process $\widehat{b}_\tau$ plays an important role in fairness and inference for quantile regression.

3.2.2 Notions of fairness

Hardt et al. (2016) introduce the notion of *equalized odds* to assess fairness of classifiers.

Suppose a classifier $\widehat{Y}$ serves to estimate some unobserved binary outcome variable Y. Then the estimator is said to satisfy the equalized odds property with respect to a protected attribute A if

$$\widehat{Y} \perp\!\!\!\perp A \,|\, Y. \tag{3.2}$$

This fairness property requires that the true positive rates $\mathbb{P}\{\widehat{Y} = 1 \,|\, Y = 1, A\}$ and the false positive rates $\mathbb{P}\{\widehat{Y} = 1 \,|\, Y = 0, A\}$ are constant functions of A. In other words, $\widehat{Y}$ has the same proportion of type-I and type-II errors across the subpopulations determined by the different values of A.

This could be extended to a related notion of fairness for quantile regression estimators. Denote the true conditional quantiles for outcome Y given attributes X as $q_\tau(X)$. Analogous to the definition of equalized odds in (3.2), we would call a quantile estimator $\widehat{q}_\tau(X)$ fair if

$$\mathbb{1}\left\{Y > \widehat{q}_\tau(X)\right\} \perp\!\!\!\perp A \,|\, \mathbb{1}\left\{Y > q_\tau(X)\right\}. \tag{3.3}$$

Conditioned on the event $\{Y \le q_\tau(X)\}$, we say that $\{Y > \widehat{q}_\tau(X)\}$ is a false positive. Conditioned on the complementary event $\{Y > q_\tau(X)\}$, we say that $\{Y \le \widehat{q}_\tau(X)\}$ is a false negative. Thus, an estimator is fair if the false positive and false negative rates do not depend on the protected attribute A.

The notion of fairness that we focus on is a natural one. Considering binary A, we ask if the average quantiles conditional on the protected attribute agree for $A = 0$ and $A = 1$. More precisely, define the effective quantiles as

$$\widehat{\tau}_a = \mathbb{P}\left\{Y \le \widehat{q}_\tau(X) \,|\, A = a\right\}, \quad a = 0, 1. \tag{3.4}$$

We say that the estimator $\widehat{q}_\tau$ is fair if $\widehat{\tau}_0 = \widehat{\tau}_1$. Typically when $\widehat{q}_\tau$ is trained on a sample of size n, exact equality is too strong to ask for. If the estimators are accurate, each of the effective quantiles should be approximately τ, up to stochastic error that decays at rate

$1/\sqrt{n}$. We say $\widehat{q}_\tau$ is $\sqrt{n}$-fair if $\widehat{\tau}_0 = \widehat{\tau}_1 + O_p(1/\sqrt{n})$. As shall be seen, this fairness property follows from the linear programming formulation when A is included in the regression. As seen from the birth weight example in Section 3.6.2, if A is not available at training time, the quantiles can be severely under- or over-estimated for a subpopulation. This formulation of fairness is closely related to calibration by group, and demographic parity (Kleinberg et al., 2017; Hardt et al., 2016; Chouldechova, 2017). An advantage of this fairness definition is that it can be evaluated empirically, and does not require a correctly specified model.

3.3 Method and Results

With samples (A_i, X_i, Y_i) drawn i.i.d. from some joint distribution F on $\mathbb{R}\times\mathbb{R}^p\times\mathbb{R}$, consider the problem of estimating the conditional quantile $q_\tau(y \mid a, x)$. Let E_F denote the expected value operator under F, or $E_F f = \int f(a,x,y)\, dF(a,x,y)$. Similarly define the probability operator under F as P_F.

Evaluate the level of fairness of an estimator $\widehat{q}_\tau(a,x)$ with

$$\mathrm{Cov}_F\left(a, \mathbb{1}\left\{y > \widehat{q}_\tau(a,x)\right\}\right) = E_F(a - E_F a)\left(\mathbb{1}\left\{y > \widehat{q}_\tau(a,x)\right\} - P_F\left\{y > \widehat{q}_\tau(a,x)\right\}\right).$$

An estimator with a smaller $|\mathrm{Cov}_F(a, \mathbb{1}\{y > \widehat{q}_\tau\})|$ is considered more fair. This measurement of fairness generalizes the notion of balanced effective quantiles described in section 3.2.2. Note that when the protected attribute is binary, $\mathrm{Cov}_F(a, \mathbb{1}\{y > \widehat{q}_\tau\}) = 0$ is equivalent to $\widehat{\tau}_0 = \widehat{\tau}_1$ for $\widehat{\tau}$ defined in (3.4).

From an initial estimator $\widehat{q}_\tau$ that is potentially unfair, we propose the following correction procedure.

On a training set of size n, compute $R_i = Y_i - \widehat{q}_\tau(A_i, X_i)$ and run quantile regression of R on A at level τ. Obtain regression slope $\widehat{\mu}_\tau$ and intercept $\widehat{\nu}_\tau$. Define correction $\widetilde{q}_\tau(a,x) = \widehat{q}_\tau(a,x) + \widehat{\mu}_\tau a + \widehat{\nu}_\tau$.

We show that this estimator $\widetilde{q}_\tau$ will satisfy the following:

1. Faithful: $P_F\{y > \widetilde{q}_\tau\} \approx 1 - \tau$;
2. Fair: $\mathrm{Cov}_F(a, \mathbb{1}\{y > \widetilde{q}_\tau\}) \approx 0$;
3. Reduced risk: It almost always improves the fit of $\widehat{q}_\tau$.

Theorem 3.1 and Theorem 3.2 contain the precise statements of our claims.

Theorem 3.1 (Faithfulness and fairness)**.** *Suppose $(A_i, X_i, Y_i) \overset{i.i.d.}{\sim} F$, and $A_i - \mathbb{E}A_i$ has finite second moment. Then the corrected estimator $\widetilde{q}_\tau$ satisfies*

$$\sup_\tau \left| P_F\{y > \widetilde{q}_\tau(a, x)\} - (1 - \tau) \right| = O_p\left(1/\sqrt{n}\right), \qquad \text{and} \tag{3.5}$$

$$\sup_\tau \left| \mathrm{Cov}_F\left(a, \mathbb{1}\{y > \widetilde{q}_\tau(a, x)\}\right) \right| = O_p\left(1/\sqrt{n}\right). \tag{3.6}$$

Furthermore, there exist positive constants C, C_1, C_2 such that $\forall t > C_1/\sqrt{n}$,

$$\mathbb{P}\left\{ \sup_\tau \left| P_F\{y > \widetilde{q}_\tau(a, x)\} - (1 - \tau) \right| > t + p/n \right\} \le C \exp\left(-C_2 n t^2\right). \tag{3.7}$$

Under the stronger assumption that the distribution of $A_i - \mathbb{E}A_i$ is sub-Gaussian, there exist positive constants C, C_1, C_2, C_3, C_4 such that $\forall t > C_1/\sqrt{n}$,

$$\mathbb{P}\left\{ \sup_\tau \left| \mathrm{Cov}_F\left(a, \mathbb{1}\{y > \widetilde{q}_\tau(a, x)\}\right) \right| > t \right\} \le C\left(\exp\left(-C_2 n t^2\right) + \exp\left(-C_3\sqrt{n}\right) + n \exp\left(-C_4 n^2 t^2\right) \right). \tag{3.8}$$

The following corollary for binary protected attributes is an easy consequence of (3.5) and (3.6).

Corollary 3.1. *If A is binary, then the correction procedure gives balanced effective quantiles:*

$$\widehat{\tau}_0 = \tau + O_p(1/\sqrt{n}), \quad \widehat{\tau}_1 = \tau + O_p(1/\sqrt{n}).$$

Remark 3.1. *By modifying the proof of Theorem 3.1 slightly, Corollary 3.1 can be extended to the case where A is categorical with K categories. In this case the correction procedure needs to be adjusted accordingly. Instead of regressing R on A, regress R on the span of the indicators $\{A = k\}$ for $k = 1, ..., K - 1$, leaving one category out to avoid collinearity. The corrected estimators will satisfy $\widehat{\tau}_k = \tau + O_p(1/\sqrt{n})$ for all categories $k = 1, ..., K$.*

Define $\mathcal{R}(\cdot) = E_F \rho_\tau(y - \cdot)$ as the risk function, where $\rho_\tau(u) = \tau u \mathbb{1}\{u > 0\} + (1-\tau)u\mathbb{1}\{u \leq 0\}$.

Theorem 3.2 (Risk quantification)**.** *The adjustment procedure $\widetilde{q}_\tau(A, X)$ satisfies*

$$\mathcal{R}(\widetilde{q}_\tau) \leq \inf_{\mu,\nu\in\mathbb{R}} \mathcal{R}(\widehat{q}_\tau + \mu A + \nu) + O_p(1/\sqrt{n}).$$

We note that in the different setting where A is a treatment rather than an observational variable, it is of interest to obtain an unbiased estimate of the treatment effect μ_τ. In this case a simple alternative approach is the so-called "double machine learning" procedure by Chernozhukov et al. (2016); in the quantile regression setting this would regress the residual onto the transformed attribute $A - \widehat{A}$ where $\widehat{A} = \widehat{A}(X)$ is a predictive model of A in terms of X.

3.4 Fairness on the training set

When a set of regression coefficients $\widehat{\beta_\tau}$ is obtained by running quantile regression of $Y \in \mathbb{R}^n$ on a design matrix $X \in \mathbb{R}^{n\times p}$, the estimated conditional quantiles on the training set $\widehat{q}_\tau(X) = X^T\widehat{\beta_\tau}$ are always "fair" with respect to any binary covariate that enters the regression. Namely, if a binary attribute X_j is included in the quantile regression, then no matter what other attributes are regressed upon, on the training set the outcome Y will lie above the estimated conditional quantile for approximately a proportion $1 - \tau$, for each of the two subpopulations $X_j = 0$ and $X_j = 1$. This phenomenon naturally arises from the mathematics behind quantile regression. This section explains this property, and lays some groundwork for the out-of-training-set analysis of the following section.

We claim that for any binary attribute X_j, the empirical effective quantiles are balanced:

$$\mathbb{P}_n\{Y > \widehat{q}_\tau(X) \,|\, X_j = 0\} \approx \mathbb{P}_n\{Y > \widehat{q}_\tau(X) \,|\, X_j = 1\} \approx 1 - \tau, \tag{3.9}$$

where $\mathbb{P}_n$ denotes the empirical probability measure on the training set.

To see why (3.9) holds, consider the dual of the quantile regression LP (3.1). This optimization has Lagrangian

$$\mathcal{L}(b, \beta) = -Y^T b + \beta^T (X^T b - (1-\tau) X^T \mathbb{1}) = -\sum_{i \le n} (Y_i - X_i^T \beta) b_i - (1-\tau) \sum_{i \le n} X_i^T \beta.$$

For fixed β, the vector $b \in [0,1]^n$ minimizing the Lagrangian tends to lie on the "corners" of the n-dimensional cube, with many of its coordinates taking value either 0 or 1 depending on the sign of $Y_i - X_i^T \beta$. We thus arrive at a characterization for $\widehat{b_\tau}$, the solution to the dual program. For i such that $Y_i \neq X_i^T \widehat{\beta_\tau}$, $\widehat{b_{\tau,i}} = \mathbb{1}\{Y_i > X_i^T \widehat{\beta_\tau}\}$. For i such that $Y_i = X_i^T \widehat{\beta_\tau}$, the values $\widehat{b_{\tau,i}}$ are solutions to the linear system that makes (3.1) hold. But with p covariates and $n > p$, such equality will typically only occur at most p out of n terms. For large n, these only enter the analysis as lower order terms. Excluding these points, the equality constraint in (3.1) translates to

$$\sum_i X_{ij} \mathbb{1}\{Y_i > X_i^T \widehat{\beta_\tau}\} = (1-\tau) \sum_i X_{ij} \quad \text{for all } j. \tag{3.10}$$

Assuming that the intercept is included as one of the regressors, the above implies that

$$\frac{1}{n} \sum_i \mathbb{1}\{Y_i > X_i^T \widehat{\beta_\tau}\} = 1 - \tau,$$

which together with (3.10), implies balanced effective quantiles for binary $X_{\cdot j}$. In particular, if the protected binary variable A is included in the regression, the resulting model will be fair on the training data, in the sense that the quantiles for the subpopulations $A = 0$ and $A = 1$ will be approximately equal, and at the targeted level τ.

This insight gives reason to believe that the quantile regression coefficients, when evaluated on an independent heldout set, should still produce conditional quantile estimates that are,what we are calling $\sqrt{n}$-fair. In the following section we establish $\sqrt{n}$-fairness for our proposed adjustment procedure. This requires us to again exploit the connection between the regression coefficients and the fairness measurements formed by the duality of the two linear programs.

3.5 Proofs

We first establish some necessary notation. From the construction of $\widetilde{q}_\tau$, the event $\{y > \widetilde{q}_\tau\}$ is equivalent to $\{r > \widehat{\mu}_\tau a + \widehat{\nu}_\tau\}$ for $r = y - \widehat{q}_\tau(a, x)$, which calls for analysis of stochastic processes of the following form. For $d \in \mathbb{R}^n$, let

$$W_d(\mu, \nu) = \frac{1}{n} \sum_{i \leq n} d_i \{R_i > \mu A_i + \nu\}.$$

Let $\overline{W}_d(\mu, \nu) = \mathbb{E} W_d(\mu, \nu)$. It is easy to check that

$$E_F(\{y > \widetilde{q}_\tau\}) = \overline{W}_{\mathbb{1}}\left(\widehat{\mu}_\tau, \widehat{\nu}_\tau\right), \qquad \mathrm{Cov}_F(a, \{y > \widetilde{q}_\tau\}) = \overline{W}_{A - \mathbb{E}A}\left(\widehat{\mu}_\tau, \widehat{\nu}_\tau\right).$$

The following lemma is essential for establishing concentration results of the W processes around $\overline{W}$, on which the proofs of Theorem 3.1 and Theorem 3.2 heavily rely.

Lemma 3.1. *Suppose $\mathcal{F}$ is a countable family of real functions on $\mathcal{X}$ and P is some probability measure on $\mathcal{X}$. Let $X_1, ..., X_n \overset{i.i.d.}{\sim} P$. If*

1. *there exists $F : \mathcal{X} \to \mathbb{R}$ such that $|f(x)| \leq F(x)$ for all x and $C^2 := \int F^2 dP < \infty$;*

2. *the collection* Subgraph$(\mathcal{F}) = \{\{(x,t) \in \mathcal{X} \times \mathbb{R} : f(x) \leq t\} : f \in \mathcal{F}\}$ *is a Vapnik-Chervonenkis(VC) class of sets,*

then there exist positive constant C_1, C_2 for which

$$\begin{aligned}&\mathbb{P}\left\{\sqrt{n}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i\leq n} f(X_i) - \int f dP\right| > t\right\}\\ \leq &4\mathbb{E}\exp\left(-\left(\frac{t}{C_2\|F\|_n} - 1\right)^2\right) + 4\mathbb{P}\Big\{2\|F\|_n > t/C_2\Big\}, \quad \forall t \geq C_1,\end{aligned} \tag{3.11}$$

where $\|F\|_n = n^{-1/2}\sqrt{\sum F(X_i)^2}$. *In particular,* $\sqrt{n}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i\leq n} f(X_i) - \int f dP\right| = O_p(1)$.

We note that more standard results could be used for concentration of measure over VC classes of Boolean functions, or over bounded classes of real functions. We use the lemma above because of its generality and to make our analysis self-contained. The proof of this result is given below.

Recall that $\mathrm{Cov}_F(a, \{y > \widetilde{q}_\tau\}) = \overline{W}_{A-\mathbb{E}A}(\widehat{\mu}_\tau, \widehat{\nu}_\tau)$. We have

$$\sup_\tau |\mathrm{Cov}_F\left(a, \{y > \widetilde{q}_\tau(a,x)\}\right)| \leq \sup_{\mu,\nu\in\mathbb{R}}\left|\left(W_{A-\mathbb{E}A}(\mu,\nu) - \overline{W}_{A-\mathbb{E}A}(\mu,\nu)\right)\right| + \sup_\tau\left|W_{A-\mathbb{E}A}\left(\widehat{\mu}_\tau, \widehat{\nu}_\tau\right)\right|.$$

We use Lemma 3.1 to control the tail of the first term using the VC class Subgraph$(\mathcal{F})$ where $\mathcal{F} = \{f : (a,r) \mapsto (a - \mathbb{E}a)\mathbb{1}\{r > \mu a + \nu\} : \mu, \nu \in \mathbb{Q}\}$. For the second term we have

$$W_{A-\mathbb{E}A}\left(\widehat{\mu}_\tau, \widehat{\nu}_\tau\right) = \frac{1}{n}\sum_{i\leq n}(A_i - \mathbb{E}A_i)\left\{R_i > \widehat{\mu}_\tau A_i + \widehat{\nu}_\tau\right\}.$$

and we exploit the dual form of quantile regression in terms of rank scores together with large deviation bounds for sub-Gaussian random variables.

The proof of Theorem 3.2 similarly exploits Lemma 3.1, but using the VC class Subgraph$(\mathcal{F})$ where $\mathcal{F} = \{f : (a,r) \mapsto \rho_\tau(r - \mu a - \nu) : \mu, \nu \in \mathbb{Q}\}$.

Proof of Lemma 3.1. To prove the lemma we first transform the problem into bounding the tail of a Rademacher process via a symmetrization technique. Let ϵ_i be distributed *i.i.d.* Rademacher ($\mathbb{P}\{\epsilon_i = 1\} = \mathbb{P}\{\epsilon_i = -1\} = 1/2$). Write P_n for the empirical (probability) measure that puts mass n^{-1} at each X_i. We claim that for all $t > 2\sqrt{2}C =: C_1$,

$$\mathbb{P}\left\{\sqrt{n}\sup_{f\in\mathcal{F}}\left|\int fdP_n - \int fdP\right| > t\right\} \le 4\mathbb{P}\left\{\sqrt{n}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i\le n}\epsilon_i f(X_i)\right| > \frac{t}{4}\right\}. \qquad (3.12)$$

Proof of (3.12): Let $\widetilde{X}_1, ..., \widetilde{X}_n$ be independent copies of $X_1, ..., X_n$ and let $\widetilde{P}_n$ be the corresponding empirical measure. Define events

$$\mathcal{A}_f = \left\{\sqrt{n}\left|\int fdP_n - \int fdP\right| > t\right\}; \quad \text{and } \mathcal{B}_f = \left\{\sqrt{n}\left|\int fd\widetilde{P}_n - \int fdP\right| \le \frac{t}{2}\right\}.$$

For all $t > C_1$,

$$\mathbb{P}\mathcal{B}_f = 1 - \mathbb{P}\left\{\sqrt{n}\left|\int fd\widetilde{P}_n - \int fdP\right| > \frac{t}{2}\right\} \ge 1 - \frac{\text{Var}f(X_1)}{(t/2)^2} \ge 1 - \frac{\int F^2dP}{(t/2)^2} \ge \frac{1}{2}.$$

On the other hand, because $\mathcal{F}$ is countable, we can always find mutually exclusive events $\mathcal{D}_f$ for which

$$\mathbb{P}\cup_{f\in\mathcal{F}}\mathcal{A}_f = \mathbb{P}\cup_{f\in\mathcal{F}}\mathcal{D}_f = \sum_{f\in\mathcal{F}}\mathbb{P}\mathcal{D}_f.$$

Since $2\mathbb{P}\mathcal{B}_f \ge 1$ for all f, the above is upper bounded by $2\sum_{f\in\mathcal{F}}\mathbb{P}\mathcal{D}_f\mathbb{P}\mathcal{B}_f$. From independence of X and $\widetilde{X}$, it can be rewritten as

$$2\sum_{f\in\mathcal{F}}\mathbb{P}(\mathcal{D}_f\cap\mathcal{B}_f) = 2\mathbb{P}\cup_{f\in\mathcal{F}}(\mathcal{D}_f\cap\mathcal{B}_f) \le 2\mathbb{P}\cup_{f\in\mathcal{F}}(\mathcal{A}_f\cap\mathcal{B}_f),$$

which is no greater than $2\mathbb{P}\{\sqrt{n}\sup_f|\int fdP_n - \int fd\widetilde{P}_n| > t/2\}$ since

$$\mathcal{A}_f\cap\mathcal{B}_f \subset \left\{\sqrt{n}\left|\int fdP_n - \int fd\widetilde{P}_n\right| > t/2\right\}.$$

Because $\widetilde{X}_i$ is an independent copy of X_i, by symmetry $f(X_i)-f(\widetilde{X}_i)$ and $\epsilon_i(f(X_i)-f(\widetilde{X}_i))$ are equal in distribution. Therefore

$$\begin{aligned}&\mathbb{P}\left\{\sqrt{n}\sup_{f\in\mathcal{F}}\left|\int fdP_n-\int fdP\right|>t\right\}\\ =&\mathbb{P}\cup_{f\in\mathcal{F}}\mathcal{A}_f\leq 2\mathbb{P}\left\{\sqrt{n}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i\leq n}\epsilon_i(f(X_i)-f(\widetilde{X}_i))\right|>\frac{t}{2}\right\}\\ \leq&2\mathbb{P}\left\{\sqrt{n}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i\leq n}\epsilon_i f(X_i)\right|>\frac{t}{4}\right\}+2\mathbb{P}\left\{\sqrt{n}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i\leq n}\epsilon_i f(\widetilde{X}_i)\right|>\frac{t}{4}\right\}\\ =&4\mathbb{P}\left\{\sqrt{n}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i\leq n}\epsilon_i f(X_i)\right|>\frac{t}{4}\right\}.\end{aligned}$$

That concludes the proof of (3.12).

Denote as $Z_n(f)$ the Rademacher process $n^{-1/2}\sum\epsilon_i f(X_i)$. Let $\mathbb{P}_X$ be the probability measure of ϵ conditioning on X. By independence of ϵ and X, ϵ_i is still Rademacher under $\mathbb{P}_X$, and it is sub-Gaussian with parameter 1. This implies that for all $f,g\in\mathcal{F}$, $Z_n(f)-Z_n(g)\sim\text{subG}\left(\sqrt{\int(f-g)^2dP_n}\right)$ under $\mathbb{P}_X$. In other words,

$$\mathbb{P}_X\left\{|Z_n(f)-Z_n(g)|>2\sqrt{\int(f-g)^2dP_n}\sqrt{u}\right\}\leq 2e^{-u},\quad\forall u>0.$$

We have shown that conditioning on X, $Z_n(f)$ is a process with sub-Gaussian increments controlled by the $\mathcal{L}^2$ norm with respect to P_n. For brevity write $\|f\|$ for $\sqrt{\int f^2dP_n}$. Apply Theorem 3.5 in Dirksen et al. (2015) to deduce that there exists positive constant C_3, such that for all $f_0\in\mathcal{F}$,

$$\mathbb{P}_X\left\{\sup_{f\in\mathcal{F}}|Z_n(f)-Z_n(f_0)|\geq C_3\left(\Delta(\mathcal{F},\|\cdot\|)\sqrt{u}+\gamma_2(\mathcal{F},\|\cdot\|)\right)\right\}\leq e^{-u}\quad\forall u\geq 1,\qquad(3.13)$$

where $\Delta(\mathcal{F},\|\cdot\|)$ is the diameter of $\mathcal{F}$ under the metric $\|\cdot\|$, and γ_2 is the generic chaining functional that satisfies

$$\gamma_2(\mathcal{F}, \|\cdot\|) \leq C_4 \int_0^{\Delta(\mathcal{F},\|\cdot\|)} \sqrt{\log N(\mathcal{F}, \|\cdot\|, \delta)} d\delta$$

for some constant C_4. Here $N(\mathcal{F}, \|\cdot\|, \delta)$ stands for the δ-covering number of $\mathcal{F}$ under the metric $\|\cdot\|$. We should comment that the generic chaining technique by Dirksen et al. (2015) is a vast overkill for our purpose. With some effort the large deviation bounds we need can be derived using the classical chaining technique.

Because $|f| \leq F$ for all $f \in \mathcal{F}$, we have $\Delta(\mathcal{F}, \|\cdot\|) \leq 2\|F\|$, so that

$$\begin{aligned}
&\int_0^{\Delta(\mathcal{F},\|\cdot\|)} \sqrt{\log N(\mathcal{F}, \|\cdot\|, \delta)} d\delta \\
\leq &\int_0^{2\|F\|} \sqrt{\log N(\mathcal{F}, \|\cdot\|, \delta)} d\delta = 2\|F\| \int_0^1 \sqrt{\log N(\mathcal{F}, \|\cdot\|, 2\delta\|F\|)} d\delta \qquad (3.14)
\end{aligned}$$

via change of variables. To bound the covering number, invoke the assumption that Subgraph($\mathcal{F}$) is a VC class of sets. Suppose the VC dimension of Subgraph($\mathcal{F}$) is V. By Lemma 19 in Nolan et al. (1987), there exists positive constant C_5 for which the $\mathcal{L}^1(Q)$ covering numbers satisfy

$$N\left(\mathcal{F}, \mathcal{L}^1(Q), \delta \int F dQ\right) \leq (C_5/\delta)^V$$

for all $0 < \delta \leq 1$ and any Q that is a finite measure with finite support on $\mathcal{X}$. Choose Q by $dQ/dP_n = F$. Choose $f_1, ..., f_N \in \mathcal{F}$ with $N = N(\mathcal{F}, \mathcal{L}^1(Q), \delta \int F dQ)$ and $\min_i \int |f - f_i| dQ \leq \delta \int F dQ$ for each $f \in \mathcal{F}$. Suppose f_i achieves the minimum. Since F is an envelope function for both f and f_i,

$$\int |f - f_i|^2 \, dP_n \leq \int 2F \, |f - f_i| \, dP_n,$$

which by definition of Q, is equal to

$$2\int |f - f_i|\, dQ \leq 2\delta \int F dQ = 2\delta \int F^2 dP_n.$$

Take square roots on both sides to deduce that

$$N(\mathcal{F}, ||\cdot||, 2\delta||F||) \leq (C_5/\delta^2)^V.$$

Plug into (3.14) this upper bound on the covering number to deduce that the integral in (3.14) converges, and $\gamma_2(\mathcal{F}, ||\cdot||)$ is no greater than a constant multiple of $||F||$. Recall that we also have $\Delta(\mathcal{F}, ||\cdot||) \leq 2||F||$. From (3.13), there exists positive constant C_6 for which

$$\mathbb{P}_X\left\{\sup_{f\in\mathcal{F}} |Z_n(f) - Z_n(f_0)| \geq C_6||F||\left(\sqrt{u}+1\right)\right\} \leq e^{-u} \quad \forall u \geq 1.$$

Take $f_0 = 0$ so we have $Z_n(f_0) = 0$. If the zero function does not belong in $\mathcal{F}$, including it in $\mathcal{F}$ does not disrupt the VC set property, and all previous analysis remains valid for $\mathcal{F} \cup \{0\}$. Letting $u = (t/4C_6||F|| - 1)^2$ yields

$$\mathbb{P}_X\left\{\sup_{f\in\mathcal{F}} |Z_n(f)| > \frac{t}{4}\right\} \leq \exp\left(-\left(\frac{t}{4C_6||F||} - 1\right)^2\right), \quad \forall t \geq 8C_6||F||.$$

Under $\mathbb{P}$, $||F||$ is no longer deterministic. Divide the probability space according to the event $\{t \geq 8C_6||F||\}$:

$$\begin{aligned}\mathbb{P}\left\{\sup_{f\in\mathcal{F}} |Z_n(f)| > \frac{t}{4}\right\} \leq& \mathbb{E}\mathbb{1}\{t \geq 8C_6||F||\}\mathbb{P}_X\left\{\sup_{f\in\mathcal{F}} |Z_n(f)| > \frac{t}{4}\right\} + \mathbb{P}\{t < 8C_6||F||\} \\ \leq& \mathbb{E}\mathbb{1}\{t \geq 8C_6||F||\}\exp\left(-\left(\frac{t}{4C_6||F||} - 1\right)^2\right) + \mathbb{P}\{t < 8C_6||F||\}.\end{aligned}$$

Choose $C_2 = 4C_6$ and (3.11) follows. □

Proof of Theorem 3.1. Recall that $\mathrm{Cov}_F(a, \{y > \widetilde{q}_\tau\}) = \overline{W}_{A-\mathbb{E}A}(\widehat{\mu}_\tau, \widehat{\nu}_\tau)$. Therefore

$$\begin{aligned}
&\sup_\tau |\mathrm{Cov}_F\,(a, \{y > \widetilde{q}_\tau(a,x)\})| \\
&= \sup_\tau \left|\overline{W}_{A-\mathbb{E}A}(\widehat{\mu}_\tau, \widehat{\nu}_\tau)\right| \\
&\leq \sup_{\mu,\nu\in\mathbb{R}} \left|\left(W_{A-\mathbb{E}A}(\mu,\nu) - \overline{W}_{A-\mathbb{E}A}(\mu,\nu)\right)\right| + \sup_\tau \left|W_{A-\mathbb{E}A}\,(\widehat{\mu}_\tau, \widehat{\nu}_\tau)\right|. \qquad (3.15)
\end{aligned}$$

Use Lemma 3.1 to control the tail of the first term. Apply Lemma 3.1 with

$$\mathcal{F} = \{f : (a,r) \mapsto (a - \mathbb{E}a)\mathbb{1}\{r > \mu a + \nu\} : \mu, \nu \in \mathbb{Q}\}.$$

Note that we are only allowing μ, ν to take rational values because Lemma 3.1 only applies to countable sets of functions. This restriction will not hurt us because the supremum of the W processes over all $\mu, \nu \in \mathbb{R}$ equals the supremum over all $\mu, \nu \in \mathbb{Q}$. Let $F(a,r) = |a - \mathbb{E}a|$ be the envelope function. We need to check that Subgraph($\mathcal{F}$) is a VC class of sets.

$$\begin{aligned}
\mathrm{Subgraph}(\mathcal{F}) &= \{\{(a,r,t) : (a - \mathbb{E}a)\mathbb{1}\{r > \mu a + \nu\} \leq t\} : \mu, \nu \in \mathbb{R}\} \\
&= \{\{(a,r,t) : (\{r > \mu a + \nu\} \cap \{a - \mathbb{E}a \leq t\}) \cup (\{r \leq \mu a + \nu\} \cap \{t \geq 0\})\} : \mu, \nu \in \mathbb{Q}\}. \qquad (3.16)
\end{aligned}$$

Since half spaces in $\mathbb{R}^2$ are of VC dimension 3 (Alon and Spencer, 2004, p 221), the set $\{\{r \leq \mu a + \nu\} : \mu, \nu \in \mathbb{Q}\}$ forms a VC class. By the same arguments all four events in (3.16) form VC classes. Deduce that Subgraph($\mathcal{F}$) is also a VC class because the VC property is stable under any finitely many union/intersection operations. The assumptions of Lemma 3.1 are satisfied, which gives that for all $t \geq 2C_1/\sqrt{n}$,

$$
\begin{aligned}
&\mathbb{P}\left\{\sup_{\mu,\nu\in\mathbb{R}} \left|W_{A-\mathbb{E}A}(\mu,\nu) - \overline{W}_{A-\mathbb{E}A}(\mu,\nu)\right| > \frac{t}{2}\right\} \\
\leq & 4\mathbb{E}\exp\left(-\left(\frac{nt}{2C_2|A-\mathbb{E}A|} - 1\right)^2\right) + 4\mathbb{P}\left\{2|A-\mathbb{E}A| > nt/2C_2\right\} \\
\leq & 4\exp\left(-\left(\frac{\sqrt{n}t}{2C_2 u} - 1\right)^2\right) + 4\mathbb{P}\left\{|A-\mathbb{E}A| > \sqrt{n}u\right\} + 4\mathbb{P}\left\{2|A-\mathbb{E}A| > nt/2C_2\right\}, \quad \forall u > 0.
\end{aligned}
$$

Here $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^n$. Under the assumption that A_i has finite second moment, we could pick u to be a large enough constant, and pick t to be a large enough constant multiple of $1/\sqrt{n}$ to make the above arbitrarily small. In other words,

$$
\sup_{\mu,\nu\in\mathbb{R}} \left|W_{A-\mathbb{E}A}(\mu,\nu) - \overline{W}_{A-\mathbb{E}A}(\mu,\nu)\right| = O_p\left(\frac{1}{\sqrt{n}}\right).
$$

Under the stronger assumption that $A_i - \mathbb{E}A_i$ is sub-Gaussian, we have that $(A_i - \mathbb{E}A_i)^2 - \mathrm{Var}(A_i)$ is sub-exponential. Choose u to be a large enough constant and we have

$$
\mathbb{P}\left\{|A-\mathbb{E}A| > \sqrt{n}u\right\} = \mathbb{P}\left\{\frac{1}{\sqrt{n}}\sum_{i\leq n}(A_i - \mathbb{E}A_i)^2 > \sqrt{n}u^2\right\} \leq \exp(-C_4(\sqrt{n}u^2 - 1)).
$$

Similarly if $t > C_1/\sqrt{n}$ for some large enough constant C_1, there exists $C_5 > 0$ such that

$$
\mathbb{P}\left\{2|A-\mathbb{E}A| > nt/2C_2\right\} \leq \exp\left(-C_5(nt^2 - 1)\right).
$$

Organizing all the terms yields for some positive constants C, C_1, C_2, C_3 whose values may have changed from previous lines,

$$
\mathbb{P}\left\{\sup_{\mu,\nu\in\mathbb{R}} \left|W_{A-\mathbb{E}A}(\mu,\nu) - \overline{W}_{A-\mathbb{E}A}(\mu,\nu)\right| > \frac{t}{2}\right\} \leq C\left(\exp\left(-C_2 nt^2\right) + \exp\left(-C_3\sqrt{n}\right)\right), \quad \forall t > C_1/\sqrt{n}.
$$

For the second term of (3.15), write

$$W_{A-\mathbb{E}A}\left(\widehat{\mu_\tau},\widehat{\nu_\tau}\right) = \frac{1}{n}\sum_{i\le n}(A_i-\mathbb{E}A_i)\left\{R_i > \widehat{\mu_\tau}A_i + \widehat{\nu_\tau}\right\}.$$

By the dual form of quantile regression (Gutenbrunner and Jurečková, 1992, p 308), there exists regression rank scores $b_\tau \in [0,1]^n$ such that

$$A^T b = (1-\tau)A^T\mathbb{1}, \quad \mathbb{1}^T b = (1-\tau)n, \quad \text{and}$$

$$b_{\tau,i} = \mathbb{1}\{R_i > \widehat{\mu_\tau}A_i + \widehat{\nu_\tau}\}, \quad \forall i \notin M_\tau,$$

for some $M_\tau \subset [n]$ of size at most p. As a result,

$$\begin{aligned}
&\sup_\tau \left|W_{A-\mathbb{E}A}\left(\widehat{\mu_\tau},\widehat{\nu_\tau}\right)\right| \\
\le&\frac{1}{n}\sup_\tau\left|A^Tb_\tau - \mathbb{E}A_1\frac{1}{n}\mathbb{1}^Tb_\tau\right| + \frac{1}{n}\sup_\tau\left|\sum_{i\in M_\tau}(A_i-\mathbb{E}A_i)\left(b_{\tau,i} - \{R_i > \widehat{\mu_\tau}A_i + \widehat{\nu_\tau}\}\right)\right| \\
=&\frac{1}{n}\sup_\tau\left|(1-\tau)A^T\mathbb{1} - \mathbb{E}A_1(1-\tau)n\right| + \frac{1}{n}\sup_\tau\left|\sum_{i\in M_\tau}(A_i-\mathbb{E}A_i)\left(b_{\tau,i} - \{R_i > \widehat{\mu_\tau}A_i + \widehat{\nu_\tau}\}\right)\right| \\
\le&\left|\frac{1}{n}\sum_{i\le n}(A_i-\mathbb{E}A_i)\right| + \frac{p}{n}\max_{i\le n}|A_i-\mathbb{E}A_i|.
\end{aligned}$$

If A_i has finite second moment, the above is clearly of order $O_p(1/\sqrt{n})$. If we have in addition that $A_i - \mathbb{E}A_i \sim \mathrm{SubG}(\sigma)$, then $|\sum_i(A_i-\mathbb{E}A_i)/n| \sim \mathrm{SubG}(\sigma/\sqrt{n})$. For all $t>0$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i\le n}(A_i-\mathbb{E}A_i)\right| > \frac{t}{4}\right\} \le 2\exp\left(-\frac{nt^2}{32\sigma^2}\right).$$

We also have

$$\mathbb{P}\left\{\frac{p}{n}\max_{i\le n}|A_i-\mathbb{E}A_i| > \frac{t}{4}\right\} \le n\mathbb{P}\left\{|A_1-\mathbb{E}A_1| > \frac{tn}{4p}\right\} \le 2\exp\left(-\frac{n^2t^2}{32\sigma^2p^2} + \log n\right).$$

Hence

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\tau}|\mathrm{Cov}_F\left(a,\{y>\widetilde{q}_\tau(a,x)\}\right)|>t\right\}\\
\leq&\mathbb{P}\left\{\sup_{\mu,\nu\in\mathbb{R}}\left|\left(W_{A-\mathbb{E}A}(\mu,\nu)-\overline{W}_{A-\mathbb{E}A}(\mu,\nu)\right)\right|>\frac{t}{2}\right\}\\
&+\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i\leq n}(A_i-\mathbb{E}A_i)\right|>\frac{t}{4}\right\}+\mathbb{P}\left\{\frac{p}{n}\max_{i\leq n}|A_i-\mathbb{E}A_i|>\frac{t}{4}\right\}\\
\leq&C\left(\exp\left(-C_2nt^2\right)+\exp\left(-C_3\sqrt{n}\right)\right)+2\exp\left(-\frac{nt^2}{32\sigma^2}\right)+2\exp\left(-\frac{n^2t^2}{32\sigma^2p^2}+\log n\right)\\
\leq&C\left(\exp\left(-C_2'nt^2\right)+\exp\left(-C_3\sqrt{n}\right)+n\exp\left(-C_4n^2t^2\right)\right).
\end{aligned}$$

That concludes the proof of (3.8). The proof of (3.7) is similar. Simply note that

$$\begin{aligned}
&\sup_{\tau}|E_F\{y>\widetilde{q}_\tau(a,x)\}-(1-\tau)|\\
=&\sup_{\tau}\left|\overline{W}_{\mathbb{1}}(\widehat{\mu}_\tau,\widehat{\nu}_\tau)-(1-\tau)\right|\\
\leq&\sup_{\mu,\nu}\left|W_{\mathbb{1}}(\mu,\nu)-\overline{W}_{\mathbb{1}}(\mu,\nu)\right|+\sup_{\tau}\left|\frac{1}{n}\sum_{i\in M_\tau}\left(b_{\tau,i}-\mathbb{1}\{R_i>\widehat{\mu}_\tau A_i+\widehat{\nu}_\tau\}\right)\right|\\
\leq&\sup_{\mu,\nu}\left|W_{\mathbb{1}}(\mu,\nu)-\overline{W}_{\mathbb{1}}(\mu,\nu)\right|+\frac{p}{n}
\end{aligned}$$

because $|M_\tau|\leq p$. Apply Lemma 3.1 with

$$\mathcal{F}=\{f:(a,r)\mapsto\mathbb{1}\{r>\mu a+\nu\}:\mu,\nu\in\mathbb{Q}\},\quad\text{and } F\equiv 1.$$

The subgraph of $\mathcal{F}$ also forms a VC set via similar analysis. Lemma 3.1 implies that if $t\geq C_1/\sqrt{n}$ for large enough C_1

$$\mathbb{P}\left\{\sup_{\mu,\nu}\left|W_{\mathbb{1}}(\mu,\nu)-\overline{W}_{\mathbb{1}}(\mu,\nu)\right|>t\right\}\leq 4\exp\left(-\left(\frac{\sqrt{n}t}{C_2}-1\right)^2\right)+\mathbb{P}\left\{2>\frac{C_1}{C_2}\right\}.$$

The second term is 0 for $C_1>2C_2$, and the desired inequality (3.7) immediately follows. □

Proof of Theorem 3.2. Suppose $\mu^*_\tau, \nu^*_\tau \in \arg\min_{\mu,\nu\in\mathbb{R}} \mathcal{R}(\widehat{q_\tau} + \mu A + \nu)$. There exists some finite constant K for which

$$(\mu^*_\tau, \nu^*_\tau) \in B_K = \{(\mu, \nu) : \max(|\mu|, |\nu|) \leq K\}.$$

Invoke Lemma 3.1 with

$$\mathcal{F} = \{f : (a, r) \mapsto \rho_\tau(r - \mu a - \nu) : \mu, \nu \in \mathbb{Q}\}.$$

The subgraph of $\mathcal{F}$ forms a VC class of sets, and on the compact set B_K, we have $|f| \leq F$ where $F(a, r) = |r| + K|a| + K$ has bounded second moment. By Lemma 3.1,

$$\sup_{(\mu,\nu)\in B_{2K}} \left| \frac{1}{n} \sum_{i\leq n} (\rho_\tau(R_i - \mu A_i - \nu) - \mathbb{E}\rho_\tau(R_i - \mu A_i - \nu)) \right| = O_p(1/\sqrt{n}). \tag{3.17}$$

Use continuity of ρ_τ to deduce existence of some $\delta > 0$ for which

$$\mathbb{E}\rho_\tau(R_1 - \mu A_1 - \nu) > \mathbb{E}\rho_\tau(R_1 - \mu^*_\tau A_1 - \nu^*_\tau) + 2\delta \quad \forall(\mu, \nu) \in \partial B_{2K}.$$

Use (3.17) to deduce that with probability $1 - o(1)$,

$$\min_{(\mu,\nu)\in\partial B_{2K}} \frac{1}{n} \sum_{i\leq n} \rho_\tau(R_i - \mu A_i - \nu) > \mathbb{E}\rho_\tau(R_1 - \mu^* A_1 - \nu^*) + \delta > \frac{1}{n} \sum_{i\leq n} \rho_\tau(R_i - \mu^*_\tau A_i - \nu^*_\tau).$$

By convexity of ρ_τ, the minimizers $\widehat{\mu}_\tau, \widehat{\nu}_\tau$ must appear with B_{2K}. Recall that $\widehat{\mu}_\tau$ and $\widehat{\nu}_\tau$ are obtained by running quantile regression of R against A on the training set, so we have

$$\frac{1}{n} \sum_{i\leq n} \rho_\tau(R_i - \widehat{\mu}_\tau A_i - \widehat{\nu}_\tau) \leq \frac{1}{n} \sum_{i\leq n} \rho_\tau(R_i - \mu^*_\tau A_i - \nu^*_\tau). \tag{3.18}$$

A few applications of the triangle inequality yields

$$\begin{aligned}
&R(\widetilde{q}_\tau) - R(\widetilde{q}^*_\tau) \\
\leq& \frac{1}{n}\sum_{i\leq n} \rho_\tau(R_i - \widehat{\mu}_\tau A_i - \widehat{\nu}_\tau) - \frac{1}{n}\sum_{i\leq n} \rho_\tau(R_i - \mu^*_\tau A_i - \nu^*_\tau) \\
&+ 2 \sup_{(\mu,\nu)\in B_{2K}} \left| \frac{1}{n}\sum_{i\leq n} (\rho_\tau(R_i - \mu A_i - \nu) - \mathbb{E}\rho_\tau(R_i - \mu A_i - \nu)) \right| \\
\leq& 0 + O_p(1/\sqrt{n})
\end{aligned}$$

by (3.18) and (3.17) . □

3.6 Experiments

3.6.1 Experiments on synthetic data

In this section we show experiments on synthetic data that verify our theoretical claims. [1] The experiment is carried out in $N = 10{,}000$ independent repeated trials. In each trial, $n = 1{,}000$ data points $(X, A, Y) \in \mathbb{R}^p \times \{0, 1\} \times \mathbb{R}$ are generated independently as follows:

- Let $p = 20$. Generate X from the multivariate distribution with correlated attributes: $X \sim \mathcal{N}(0, \Sigma)$, where the the covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ takes value 1 for diagonal entries and 0.3 for off-diagonal entries.

- The protected attribute A depends on X through a logistic model: $A \mid X \sim \text{Bernoulli}(b)$ with
$$b = \exp\left(X^T\gamma\right) / \left(1 + \exp\left(X^T\gamma\right)\right).$$

- Given A, X, generate Y from a heteroscedastic model: $Y \mid A, X \sim \mathcal{N}\left(X^T\beta + \mu A, (X^T\eta)^2\right)$.

[1] Code and data for all experiments are available online at `https://drive.google.com/file/d/1Ibaq5VWaAE4539hec4-UdIOgPsNv0x_t/view?usp=sharing`

The parameters β, γ, η are all generated independently from $\mathcal{N}(0, I_p)$ and stay fixed throughout all trials. The coefficient μ is set to be 3.

In each of the N trials, conditional quantile estimators are trained on a training set of size $n/2$ and evaluated on the remaining size $n/2$ held out set. We train three sets of quantile estimators at $\tau = 0.5$:

1. Full quantile regression of Y on A and X.

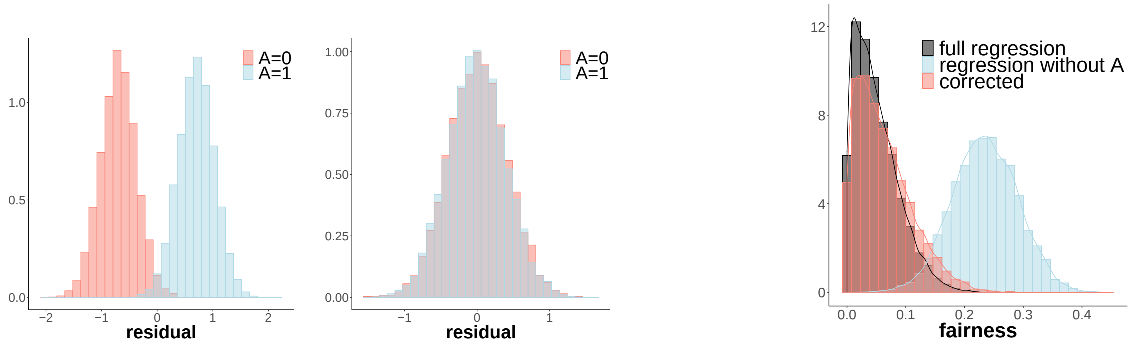
2. Quantile regression of Y on X only.

3. Take the estimator from procedure 2 and correct it with the method described in Section 3.3.

The average residuals $Y - \widehat{q}_\tau(X, A)$ are then evaluated on the test set for the $A = 0$ and $A = 1$ subpopulations. In Figure 3.1 we display the histograms of these average residuals across all N trials for the quantile regression estimator on X (3.1a) and the corrected estimator (3.1b). In the simulation we are running, A is positively correlated with the response Y. Therefore when A is excluded from the regression, the quantile estimator underestimates when $A = 1$ and overestimates when $A = 0$. That is why we observe different residual distributions for the two subpopulations. This effect is removed once we apply the correction procedure, as shown in Figure 3.1c.

We also test whether our correction procedure corrects the unbalanced effective quantiles of an unfair initializer. In each trial we measure the fairness level of an estimator $\widehat{q}_\tau$ by the absolute difference $|\widehat{\tau}_1 - \widehat{\tau}_0|$ between the effective quantiles of the two subpopulations on a heldout set S, where $\widehat{\tau}_a$ is defined as in (3.4).

We established in previous sections that quantile regression excluding attribute A is in general not fair with respect to A. A histogram of the fairness measure obtained from this procedure is shown in Figure 3.1c (salmon). Plotted together are the fairness measures after the correction procedure (light blue). For comparison we also include the histogram



(a) (b) (c)

Figure 3.1: From left to right: (a): histograms of average residuals for quantile regression of Y on X only; (b): histograms of average residuals for the corrected quantile estimators; (c): histograms and density estimates of the fairness measures obtained by running quantile regression on X, before (salmon) and after (light blue) the adjustment procedure. The histogram from the full regression (black) serves as a benchmark for comparison.

obtained from the full regression (black). Note that the full regression has the “unfair” advantage of having access to all observations of A. Figure 3.1c shows that the correction procedure pulls the fairness measure to a level comparable to that of a full regression, which as we argued in Section 3.4, produces $\sqrt{n}$-fair estimators.

3.6.2 Birthweight data analysis

The birth weight dataset from Abrevaya (2001), which is analyzed by Koenker and Hallock (2001), includes the weights of 198,377 newborn babies, and other attributes of the babies and their mothers, such as the baby’s gender, whether or not the mother is married, and the mother’s age. One of the attributes includes information about the race of the mother, which we treat as the protected attribute A. The variable A is binary—black ($A = 1$) or not black ($A = 0$). The birth weight is reported in grams. The other attributes include education of the mother, prenatal medical care, an indicator of whether the mother smoked during pregnancy, and the mother’s reported weight gain during pregnancy.

Figure 3.2 shows the coefficients $\widehat{\beta}_\tau$ obtained by fitting a linear quantile regression model, regressing birth weight on all other attributes. The model is fit two ways, either

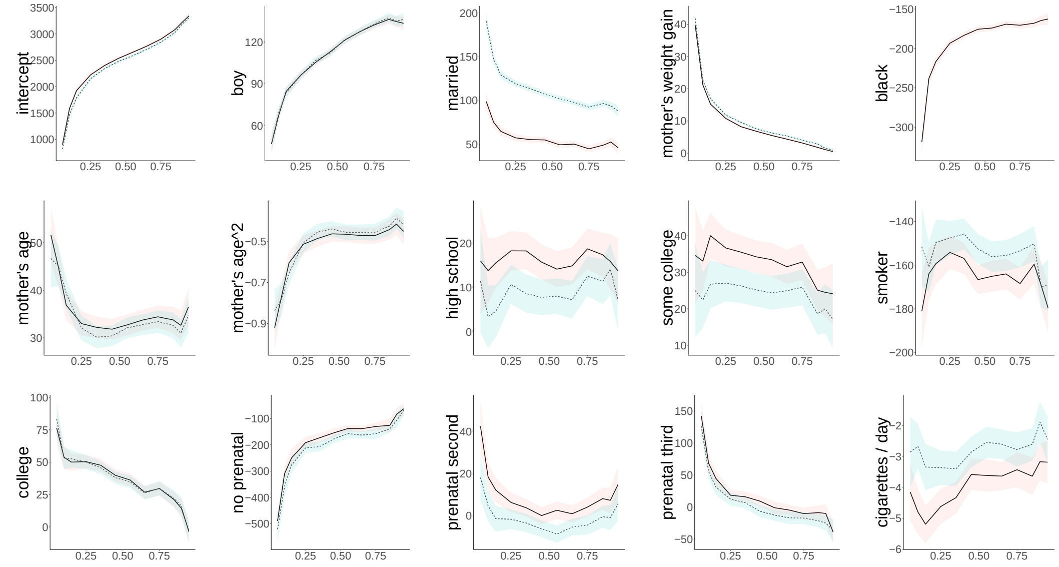


Figure 3.2: Quantile regression coefficients for birth data. The quantile τ runs along horizontal axis; curves are the coefficients $\widehat{\beta_\tau}$; unit is grams. Solid/salmon: race is included in the model; dashed/blue: race excluded. When race is excluded, the strongly correlated variable "married" can be seen as serving as a kind of proxy.

including the protected race variable A (solid, salmon confidence bands), or excluding A (long dashed, light blue confidence bands). The top-right figure shows that babies of black mothers weigh less on average, especially near the lower quantiles where they weigh nearly 300 grams less compared to babies of nonblack mothers. A description of other aspects of this linear model is given by Koenker and Hallock (2001). A striking aspect of the plots is the disparity between birth weights of infants born to black and nonblack mothers, especially at the left tail of the distribution. In particular, at the 5th percentile of the conditional distribution, the difference is more than 300 grams. Just as striking is the observation that when the race attribute A is excluded from the model, the variable "married," with which it has a strong negative correlation, effectively serves as a proxy, as seen by the upward shift in its regression coefficients. However, this and the other variables do not completely account for race, and as a result the model overestimates the weights of infants born to black mothers, particularly at the lower quantiles.

To correct for the unfairness of $\widehat{q}_\tau$, we apply the correction procedure described in Section 3.3. For the target quantile $\tau = 20\%$, the corrected estimator $\widetilde{q}_\tau$ achieves effective quantiles 20.4% for the black population and 20.1% for the nonblack population. Table 3.1 (left) shows the effective quantiles at a variety of quantile levels. We see that the correction procedure consistently pulls the effective quantiles for both subpopulations closer to the target quantiles.

target	5	25	50	75	95
$\widehat{\tau}_0$(before)	4.42	23.14	47.87	73.12	94.50
$\widehat{\tau}_1$(before)	7.91	33.80	60.02	82.46	96.90
$\widehat{\tau}_0$(after)	5.03	25.01	49.77	74.61	94.79
$\widehat{\tau}_1$(after)	5.02	24.02	49.95	74.62	94.99

Table 3.1: The effective quantiles before and after correction.

For 2000 randomly selected individuals from the test set, Figure 3.3 shows their observed birth weights plotted against the conditional quantile estimation at $\tau = 20\%$ before (left) and after (right) the correction. The dashed line is the identity. When A is not included in the quantile regression, the conditional quantiles for the black subpopulation are overestimated. Our procedure achieves fairness correction by shifting the estimates for the $A_i = 1$ data points smaller (to the left) and shifting the $A_i = 0$ data points larger (to the right). After the correction, the proportion of data points that satisfy $Y \leq \widetilde{q}_\tau$ are close to the target 20% for both subpopulations.

3.7 Discussion

In this chapter we have studied the effects of excluding a distinguished attribute from quantile regression estimates, together with procedures to adjust for the bias in these estimates through post-processing. The linear programming basis for quantile regression leads to properties and analyses that complement what has appeared previously in the

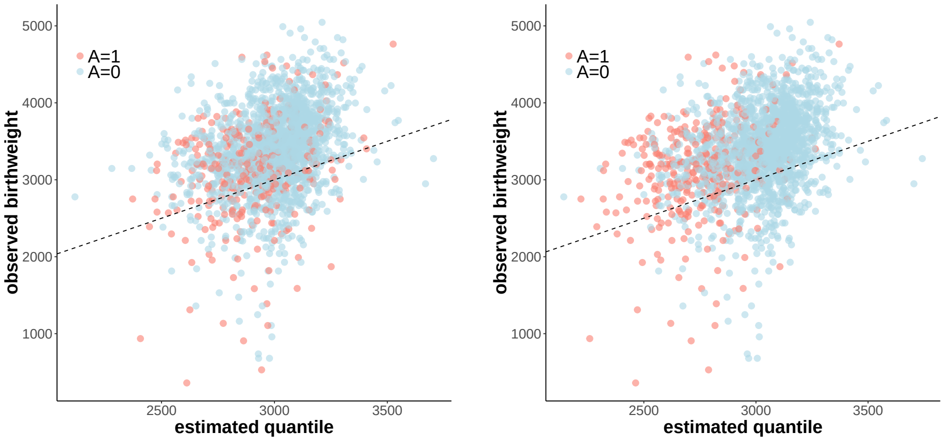


Figure 3.3: scatter plots of observed birth weights against estimated 20% conditional quantiles over the test set, before and after adjustment.

fairness literature. Several extensions of the work presented here should be addressed in future work. For example, the generality of the concentration result of Lemma 3.1 could allow the extension of our results to multiple attributes of different types. In the fairness analysis in Section 3.5 we used a linear quantile regression in the adjustment step, which allows us to more easily leverage previous statistical analyses on quantile rank scores (Gutenbrunner and Jurečková (1992)). Nonparametric methods would be another interesting direction to explore.

The birth data studied here has been instrumental in developing our thinking on fairness for quantile regression. It will be interesting to investigate the ideas introduced here for other data sets. If the tail behaviors, including outliers, of the conditional distributions for a set of subpopulations are very different, and the identification of those subpopulations is subject to privacy restrictions or other constraints that do not reveal them in the data, the issue of bias in estimation and decision making will come into play.

Bibliography

Abrevaya, J. (2001). The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics 26*(1), 247–257. 68

Alon, N. and J. H. Spencer (2004). *The probabilistic method.* John Wiley & Sons. 61

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. K. Newey (2016). Double machine learning for treatment and causal parameters. Technical report, CEMMAP working paper, Centre for Microdata Methods and Practice. 53

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data 5*(2), 153–163. 47, 51

Dirksen, S. et al. (2015). Tail bounds via generic chaining. *Electronic Journal of Probability 20*. 58, 59

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, New York, NY, USA, pp. 214–226. ACM. 47

Gutenbrunner, C. and J. Jurečková (1992). Regression rank scores and regression quantiles. *The Annals of Statistics 20*(1), 305–330. 49, 63, 71

Hardt, M., E. Price, N. Srebro, et al. (2016). Equality of opportunity in supervised learning.

In *Advances in Neural Information Processing Systems*, pp. 3315–3323. Red Hook, NY: Curran Associates, Inc. 47, 49, 51

Kleinberg, J., S. Mullainathan, and M. Raghavan (2017). Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS)*, Volume 67 of *LIPIcs*. Schloss Dagstuhl: Leibniz-Zentrum fuer Informatik. 47, 51

Koenker, R. and K. Hallock (2001). Quantile regression: An introduction. *Journal of Economic Perspectives 15*(4), 43–56. 68, 69

Mosteller, F. and J. W. Tukey (1977). *Data Analysis and Regression: A Second Course in Statistics*. Reading, Massachusetts: Addison-Wesley. 47

Nolan, D., D. Pollard, et al. (1987). *u*-processes: Rates of convergence. *The Annals of Statistics 15*(2), 780–799. 59

Woodworth, B., S. Gunasekar, M. I. Ohannessian, and N. Srebro (2017). Learning non-discriminatory predictors. arXiv preprint arXiv:1702.06081. 47

Chapter 4

Consistent recovery threshold of hidden nearest neighbor graphs

Joint work with Prof. Jian Ding, Prof. Yihong Wu and Prof. Jiaming Xu

Abstract

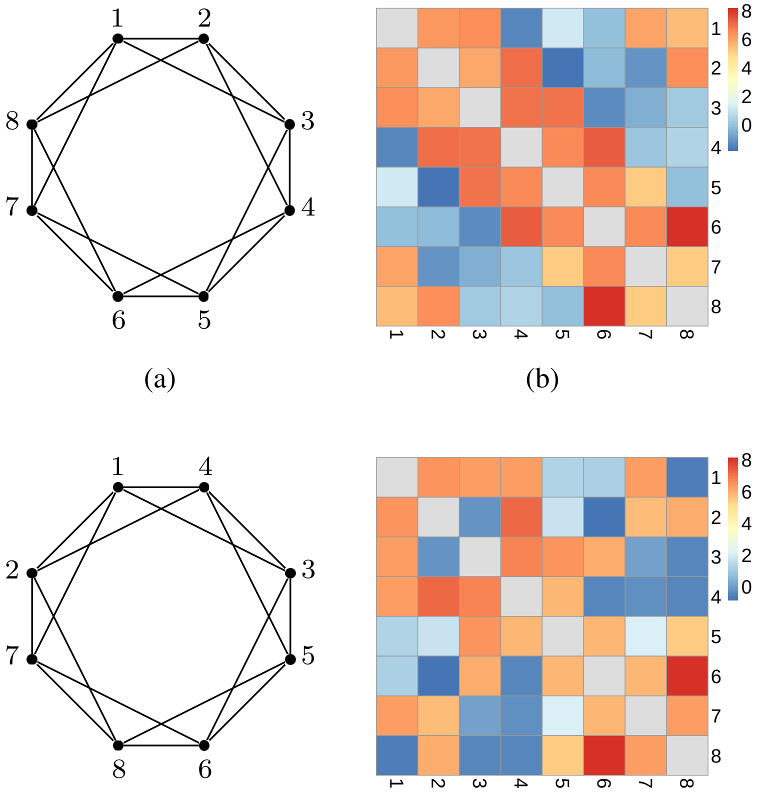
In this chapter we study the problem of recovering a hidden $2k$-NN graph from its noisy observation. We give sufficient conditions under which exact recovery is possible and is achieved by the maximum likelihood estimator. We also give a matching information-theoretic lower bound. The exact recovery threshold we obtain is sharp with provable matching constants.

4.1 Introduction

Many datasets call for network models that demonstrate both strong local links and weak global links. One abstraction of such real life models is to combine weak signals on a complete graph with strong signals on a nearest neighbor (NN) graph.

Definition 4.1 (2k-NN graph)**.** *A simple undirected graph on the vertex set* $[n]$ *is called a* 2k*-NN graph if there exists a permutation* σ *on* $[n]$*, such that* $i \sim j$ *if and only if* $\min\{|\sigma(i) - \sigma(j)|, n - |\sigma(i) - \sigma(j)|\} \leq k$*.*

In other words, each 2k-NN graph is isomorphic to the usual circulant graph where n vertices are equally spaced on a circle and each pair of vertices within distance k are connected (see Fig. 4.1a, 4.1c). A 2k-NN graph can be constructed as follows: first, construct a Hamiltonian cycle $(\sigma(1), \sigma(2), \ldots, \sigma(n), \sigma(1))$, then connect pairs of vertices that are at distance at most k.



(a) (b)

(c) (d)

Figure 4.1: Examples of 2k-NN graphs for $n = 8$ and $k = 2$. (a): the 2k-NN graph characterized by the Hamiltonian cycle $(1, 2, 3, 4, 5, 6, 7, 8, 1)$; (c): the 2$k$-NN graph characterized by the Hamiltonian cycle $(1, 4, 3, 5, 6, 8, 7, 2, 1)$; (b), (d): heatmap of one realization of w under the Hidden 2k-NN graph model with underlying 2k-NN graphs represented by (a), (c) respectively.

In this chapter, we consider the following statistical model for weighted graphs which can be viewed as noisy observation of a hidden $2k$-NN graph (ground truth).

Definition 4.2 (Hidden $2k$-NN graph model)**.** *Let x^* denote the adjacency vector of a $2k$-NN graph on n vertices. Given probability distributions P_n and Q_n, let w denote the weighted adjacency vector of a complete weighted graph, where w_e's are independently drawn from P_n if $x^*_e = 1$ and Q_n if $x^*_e = 0$, respectively.*

Given the weighted graph w (*e.g.* see Fig. 4.1b, 4.1d), the goal is to infer the underlying $2k$-NN graph.

When $k = 1$, the model simplifies to the model of hidden Hamiltonian cycle, which was studied extensively by Bagaria et al. (2018). The application they considered arises from *Genome Scaffolding*, that is, extending genome subsequences (so-called contigs) to the whole genome by ordering them according to their positions on the genome. The data available for genome scaffolding is the linkage strength between contigs measured by randomly sampled Hi-C reads Lieberman-Aiden et al. (2009); Putnam et al. (2016) , where one expects to see a larger count of Hi-C reads when two contigs are close on the genome. In order to infer the ordering of the contigs, the authors of Bagaria et al. (2018) studied the hidden Hamiltonian cycle model, where each node of the random graph is a contig, the hidden Hamiltonian cycle corresponds to the underlying true ordering of contigs, and the edge weights represent the counts of the Hi-C reads linking the contigs. This hidden Hamiltonian cycle model is limited by the assumption that only contigs that are adjacent on the genome demonstrate strong signal – an elevated mean number of Hi-C reads. The general $2k$-NN graph model is a closer approximation to the real data, capturing the large Hi-C counts observed between contigs that are nearby on the genome.

By restricting both P_n and Q_n to Bernoulli distributions with corresponding success probabilities $p_n > q_n$, we arrive at a variant of the "small-world" network model Watts and Strogatz (1998), Newman and Watts (1999). The "small-world" model can be interpreted as an interpolation between a ring lattice and an Erdös-Rényi random graph. It is often used to model the social networks that exhibit some local “neighborhood” structure as well as small graph diameters. In particular, under the “small-world” model, if two people in the network share a mutual friend, then they are likely to be also connected in the random graph. In addition, if two people are uniformly drawn at random, then it is very unlikely that there is an edge connecting them; however, they can typically reach each other through a short path on the graph, hence the name “small-world”. The hidden graphs may correspond to the geographical locations of people in the physical space, or to their demographic characteristics in the embedded space. Beyond social networks, the hidden graphs my unveil patterns of viral infection, or help generate word embeddings from word co-occurrence networks. It is of interest to discover these hidden graph structures from the “small-world” graphs.

We study the following two recovery problems.

Definition 4.3 (Exact recovery)**.** *An estimator $\widehat{x}$ achieves exact recovery if, as $n \to \infty$,*

$$\inf_{x^*} \mathbb{P}\left\{\widehat{x} \neq x^*\right\} = o(1),$$

where the infimum is over all adjacency vectors of $2k$-NN graphs on $[n]$, and the expected value is taken under the $2k$-NN graph model where the underlying $2k$-NN graph corresponds to x^*.

We aim to find the weakest conditions on P_n and Q_n for exact recovery to be possible in the $k = n^{o(1)}$ regime. In the context of the “small-world” network, the problem of weak recovery was previously addressed in Cai et al. (2017). They specify the model with

$$P_n = \text{Bernoulli}\left(1 - \beta\left(1 - \beta\frac{2k}{n-1}\right)\right); \quad Q_n = \text{Bernoulli}\left(\beta\frac{2k}{n-1}\right),$$

which can be interpreted as a rewiring model where one starts from a $2k$-NN graph. Each edges is removed independently with probability β. Each edge on the complete graph is then connected independently with probability $\beta \frac{k}{n-1}$.

It is shown in Cai et al. (2017) that a necessary condition to achieve both reliable testing and exact recovery is $(1-\beta) = o\left(\sqrt{\frac{\log n}{n}} \vee \frac{\log n}{k} \frac{1}{\log \frac{n \log n}{k^2}}\right)$. When $k = n^{o(1)}$, this condition translates to $1-\beta = o(1/k)$. Compare with our exact recovery threshold $\beta = o(1/\sqrt{n})$. We also show that this condition is sufficient for exact recovery.

We investigate the necessary and the sufficient conditions for exact recovery under the more general hidden $2k$-NN graph model. The conditions we obtain are sharp, in the sense that the boundary between the possibility and impossibility regions is precisely characterized, with exact constant coefficients. The proof techniques we use are considerably different from Cai et al. (2017). We show that, roughly speaking, the necessary and sufficient condition for exact recovery to be possible is

$$\liminf_{n\to\infty} \frac{2\alpha_n}{\log n} > 1$$

, where α_n is the Rényi divergence between P_n and Q_n or order $1/2$:

$$\alpha_n = -2\log \int \sqrt{\mathrm{d}P_n \mathrm{d}Q_n}.$$

The following section contains the exact statements of proofs of our result.

4.2 Exact recovery

The maximum likelihood estimator (MLE) for the hidden $2k$-NN graph problem is equivalent to finding the max-weighted $2k$-NN graph with weights given by the log likelihood ratios. Specifically, assuming that dP_n/dQ_n is well-defined, for each edge $e \in [\binom{n}{2}]$, let $L_e = \log \frac{\mathrm{d}P_n}{\mathrm{d}Q_n}(w_e)$. Then the MLE is the solution to the following combinatorial optimization problem:

$$\widehat{x}_{\mathrm{ML}} = \arg\max_{x \in \mathcal{X}} \ \langle L, x \rangle \tag{4.1}$$

where the feasible set $\mathcal{X} \subset \{0,1\}^{\binom{n}{[2]}}$ is the collection of adjacency vectors of all $2k$-NN graphs on $[n]$. Note that in the Poisson or Gaussian model where the log likelihood ratio is an affine function of the edge weight, we can simply take L to be w itself.

Assumption 4.1. *[c.f. (Bagaria et al., 2018, Lemma 1)]*

$$\sup_{\tau \in \mathbb{R}} (\log \mathbb{P}\{Y_1 \geq \tau\} + \log \mathbb{P}\{X_1 \leq \tau\}) \geq -(1 + o(1))\alpha_n + o(\log n).$$

The authors of Bagaria et al. (2018) remarked that 4.1 is very general and is fulfilled when the weight distributions are either Poisson, Gaussian or Bernoulli.

Theorem 4.1. *Let $k \geq 2$. Suppose $\alpha_n - \frac{1}{2}(\log n + 17 \log k) \to \infty$. Then $\mathbb{P}\{\widehat{x}_{\mathrm{ML}} \neq x^*\} \to 0$. In particular, if $k = n^{o(1)}$ and*

$$\liminf_{n \to \infty} \frac{2\alpha_n}{\log n} > 1,$$

the MLE achieves exact recovery.

Conversely, assume that $k < n/12$ and 4.1 holds. If exact recovery is possible, then

$$\liminf_{n \to \infty} \frac{2\alpha_n}{\log n} \geq 1.$$

4.2.1 Suboptimality of two algorithms

So far no polynomial-time algorithm is known to achieve the sharp threshold $\liminf_{n\to\infty} \frac{2\alpha_n}{\log n} = 1$. We investigated these two natural relaxations of maximum likelihood to find an efficient algorithm. However to achieve exact recovery both algorithms require that $\liminf_{n\to\infty} \frac{\alpha_n}{\log n} > 1$, a condition stronger than in Theorem 4.1.

1. The $2k$-factor ILP amounts to the following optimization problem:

$$\widehat{x}_{2kF} = \arg\max_{x} \langle L, x \rangle$$
$$\text{s.t.} \sum_{v \sim u} x_{(u,v)} = 2k, \quad \forall u,$$
$$x_e \in \{0, 1\}, \quad \forall e.$$

   To see why the $2k$-factor ILP does not achieve the sharp threshold, consider $P_n = \mathcal{N}(\mu_n, 1)$ and $Q_n = \mathcal{N}(0, 1)$. In this case $L_e = \mu_n w_e - \mu_n^2/2$, hence we could equivalently maximize $\langle w, x \rangle$. The condition $\liminf \frac{2\alpha_n}{\log n} > 1$ translates to $\liminf \frac{\mu_n}{\sqrt{2 \log n}} > 1$. Assume that x^* is characterized by the identity permutation $\sigma^*(i) \equiv i$ and consider alternative solutions of this form: fix two vertices i, j for which $d_{x^*}(i, j) > k$, define $x^{(i,j)}$ to be the solution that removes the edges $(i, i+1)$ and $(j, j+1)$ and adds the edges (i, j) and $(i+1, j+1)$. That is, $x^{(i,j)}(i, i+1) = x^{(i,j)}(j, j+1) = 0$, $x^{(i,j)}(i, j) = x^{(i,j)}(i+1, j+1) = 1$ and $x^{(i,j)}$ agrees with x^* in all other positions. There are $O(n^2)$ such alternative solutions and they are close to being mutually independent. For each pair (i, j), we have $\langle w, x^{(i,j)} - x^* \rangle \sim \mathcal{N}(-2\mu_n, 4)$. Unless $\liminf \frac{\mu_n}{\sqrt{4 \log n}} > 1$, with high probability at least one of the feasible solutions of the $2k$-factor ILP is such that $\langle w, x \rangle > \langle w, x^* \rangle$. From the analysis in (Bagaria et al., 2018, Section 4.2), $\liminf \frac{\mu_n}{\sqrt{4 \log n}} > 1$ is sufficient for the $2k$-factor ILP to achieve exact recovery.

2. The LP relaxation yields from relaxing the integer constraint in the $2k$-factor ILP to $x_e \in [0, 1]$. By the same argument as the ILP, the LP relaxation cannot consistently recover x^* if $\liminf \frac{\mu_n}{\sqrt{4 \log n}} < 1$. From similar analysis in (Bagaria et al., 2018, Section 5), the LP relaxation achieves exact recovery when $\liminf \frac{\mu_n}{\sqrt{4 \log n}} > 1$.

4.2.2 Proof of correctness for MLE

To analyze the MLE, we first introduce the notion of *difference graph*, which encodes the difference between a proposed $2k$-NN graph and the ground truth. Given $x, x^* \in \{0,1\}^{\binom{n}{2}}$, let $G = G(x)$ be a bi-colored simple graph on $[n]$ whose adjacency vector is $x - x^* \in \{0, \pm 1\}^{\binom{n}{2}}$, in the sense that each pair (i, j) is connected by a blue (resp. red) edge if $x_{ij} - x^*_{ij} = 1$ (resp. -1). See Figure 4.2 on page 81 for an example.

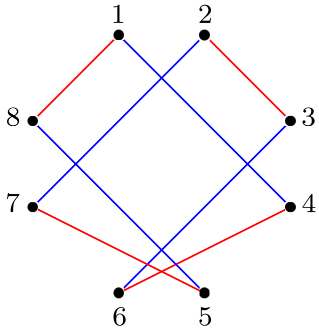


Figure 4.2: An example for a difference graph G. Here G is obtained by letting x^* (resp. x) be the $2k$-NN graph in Fig. 4.1a (resp. 4.1c), and then taking the difference $x - x^*$.

Therefore, red edges in $G(x)$ are true edges in x^* that are missed by the proposed solution x, and blue edges correspond to spurious edges that are absent in the ground truth.

A key property of difference graphs is the following: Since $2k$-NN graphs are $2k$-regular, the difference graph G is *balanced*, in the sense that for each vertex, its red degree (the number of incident red edges) coincides with its blue degree. Consequently, G has equal number of red edges and blue edges, and the number of red (or blue) edges measures the closeness of x to the truth x^*. Denote

$$\mathcal{X}_\Delta = \{x \in \mathcal{X} : d(x, x^*) = 2\Delta\} = \{x \in \mathcal{X} : G(x) \text{ contains exactly } \Delta \text{ red edges}\}. \qquad (4.2)$$

In particular, $\{\mathcal{X}_\Delta : \Delta \geq 0\}$ partitions the feasible set $\mathcal{X}$. The crux of the proof lies on the following combinatorial lemma (proved in Section 4.2.4) bounding the cardinality of $\mathcal{X}_\Delta$:

Lemma 4.1. *There exist an absolute constant C such that for any $\Delta \geq 0$ and any $1 \leq k \leq n$,*

$$|\mathcal{X}_\Delta| \leq 2\left(Ck^{17}n\right)^{\Delta/2} \tag{4.3}$$

Assuming 4.1, the proof of the correctness of MLE follows from the union bound. First of all,

$$\mathbb{P}\{\exists x \in \mathcal{X} : \langle L, x - x^*\rangle > 0\} \leq \sum_{\Delta \geq 1} \mathbb{P}\{\exists x \in \mathcal{X}_\Delta : \langle L, x - x^*\rangle > 0\}. \tag{4.4}$$

Recall that $L_e = \log(dP_n/dQ_n)(A_e)$. Hence for each $x \in \mathcal{X}_\Delta$, the law of $\langle L, x - x^*\rangle$ only depends on Δ, which can be represented as follows:

$$\langle L, x - x^*\rangle \overset{\mathcal{D}}{=} \sum_{i \leq \Delta} Y_i - \sum_{i \leq \Delta} X_i,$$

where $X_1, \ldots, X_\Delta$ are i.i.d. copies of $\log(\mathrm{d}P_n/\mathrm{d}Q_n)$ under P_n; $Y_1, \ldots, Y_\Delta$ are i.i.d. copies of $\log(\mathrm{d}P_n/\mathrm{d}Q_n)$ under Q_n, and $\overset{\mathcal{D}}{=}$ denotes equality in distribution. Applying Chernoff's bound yields

$$\mathbb{P}\left\{\sum_{i \leq \Delta} Y_i - \sum_{i \leq \Delta} X_i > 0\right\} \leq \inf_{\lambda > 0} \exp\left(\Delta\left(\psi_Q(\lambda) + \psi_P(-\lambda)\right)\right),$$

where ψ_P, ψ_Q are the log-moment generating functions of $\log(\mathrm{d}P_n/\mathrm{d}Q_n)$ under P_n and Q_n respectively. Note that

$$\psi_P(-\lambda) = \log \int \left(\frac{\mathrm{d}P_n}{\mathrm{d}Q_n}(x)\right)^{-\lambda} P_n(dx) = \log \int \left(\frac{\mathrm{d}P_n}{\mathrm{d}Q_n}(x)\right)^{1-\lambda} Q_n(dx) = \psi_Q(1 - \lambda).$$

Choosing $\lambda = 1/2$ yields:

$$\mathbb{P}\{\langle L, x - x^*\rangle > 0\} = \mathbb{P}\left\{\sum_{i \leq \Delta} Y_i - \sum_{i \leq \Delta} X_i > 0\right\} \leq \exp\left(2\Delta\psi_Q\left(\frac{1}{2}\right)\right) = \exp\left(-\alpha_n\Delta\right). \tag{4.5}$$

Combining (4.3) and (4.5) and applying the union bound, we obtain

$$\sum_{x \in \mathcal{X}_\Delta} \mathbb{P}\{\langle L, x - x^* \rangle > 0\} \leq 2 \exp(-\Delta \kappa_n),$$

where $\kappa_n \triangleq \alpha_n - \frac{\log(Ck^{17}n)}{2} \to \infty$ by assumption. Finally, from (4.4) we get

$$\mathbb{P}\{\exists x \in \mathcal{X} : \langle L, x - x^* \rangle > 0\} \leq \sum_{\Delta \geq 1} 2 \exp(-\Delta \kappa_n) = \frac{2\exp(-2\kappa_n)}{1 - \exp(-\kappa_n)} \xrightarrow{n \to \infty} 0,$$

4.2.3 Information-theoretic lower bound

For the purpose of lower bound, consider the Bayesian setting where x^* is drawn uniformly at random from the set $\mathcal{X}$ of all $2k$-NN graphs. Then MLE maximizes the probability of success, which, by definition, can be written as follows:

$$\mathbb{P}\{\widehat{x}_{\mathrm{ML}} = x^*\} = \mathbb{P}\{\langle L, x - x^* \rangle < 0, \quad \forall x \neq x^*\}.$$

It is difficult to work with the intersection of dependent events. The proof strategy is to select a subset of feasible solutions for which the events $\langle L, x - x^* \rangle < 0$ are mutually independent. We select x that are of the following form.

For the ground truth x^*, assume WLOG that σ^* is the identity permutation. Define $x^{(i)}$ to be solution that corresponds to σ, with $\sigma(i) = i + 1$, $\sigma(i + 1) = i$, and $\sigma = \sigma^*$ everywhere else. It is easy to see that the difference graph $G_{x^{(i)}}$ contains four edges: (see Figure 4.3 on page 84)

$$\text{red edges:} \quad (i - k, i), \quad (i + 1, i + k + 1);$$
$$\text{blue edges:} \quad (i - k, i + 1), \quad (i, i + k + 1).$$

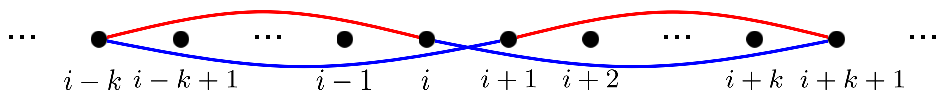


Figure 4.3: The difference graph $G(x^{(i)})$.

For two feasible solutions $x^{(i)}$ and $x^{(j)}$ wth $k+1 \leq i \leq j \leq n-k$, the edges sets $E\left(G_{x^{(i)}}\right)$ and $E\left(G_{x^{(j)}}\right)$ intersect if and only if $j-i \in \{0, k, k+1\}$. We can avoid proposing such pairs of (i, j) by dividing the x^* cycle into blocks of $3k$, each divided evenly into sections of length k, and only propose $x^{(i)}$ whose index i lies in the middle section of a block. Define

$$D = \{k+1, k+2, \ldots, 2k, 4k+1, \ldots, 5k, \ldots, 3k(\lfloor n/3k \rfloor - 1) + k + 1, \ldots, 3k(\lfloor n/3k \rfloor - 1) + 2k\}.$$

Each element in D corresponds to a solution x with 4 edges in the difference graph and none of these edges would appear in the edge set of the difference graph of another solution in D. That means all elements of $\left\{\langle L, x^{(i)} - x^* \rangle : i \in D\right\}$ are mutually independent.

For each $i \in D$, we have

$$\begin{aligned}
&\mathbb{P}\{\langle L, x^{(i)} - x^* \rangle < 0\} \\
=&\mathbb{P}\left\{L(i-k, i+1) + L(i, i+k+1) - L(i-k, i) - L(i+1, i+k+1) < 0\right\} \\
=&\mathbb{P}\left\{Y_1 + Y_2 - X_1 - X_2 < 0\right\},
\end{aligned}$$

where X_1, X_2 are independent copies of $\log(\mathrm{d}P_n/\mathrm{d}Q_n)$ under P_n, and Y_1, Y_2 are independent copies of $\log(\mathrm{d}P_n/\mathrm{d}Q_n)$ under Q_n. Therefore

$$\begin{aligned}
&\mathbb{P}\left\{\langle L, x - x^* \rangle < 0, \quad \forall x \neq x^*\right\} \\
\leq&\mathbb{P}\left\{\langle L, x^{(i)} - x^* \rangle < 0 \quad \forall i \in D\right\} \\
=&\left(\mathbb{P}\left\{Y_1 + Y_2 - X_1 - X_2 < 0\right\}\right)^{|D|} \\
\leq&\exp\left(-|D|\mathbb{P}\left\{Y_1 + Y_2 - X_1 - X_2 \geq 0\right\}\right).
\end{aligned} \tag{4.6}$$

From the mutual independence of X_1, X_2, Y_1, Y_2,

$$\log \mathbb{P}\{Y_1 + Y_2 - X_1 - X_2 \geq 0\} \geq 2 \sup_{\tau \in \mathbb{R}} \left(\log \mathbb{P}\{Y_1 \geq \tau\} + \log \mathbb{P}\{X_1 \leq \tau\}\right).$$

Recall that (4.6) is an upper bound for $\mathbb{P}\{\widehat{x}_{\mathrm{ML}} = x^*\}$. The success of MLE must require that

$$\log |D| + 2 \sup_{\tau \in \mathbb{R}} \left(\log \mathbb{P}\{Y_1 \geq \tau\} + \log \mathbb{P}\{X_1 \leq \tau\}\right) \to -\infty.$$

On one hand, by 4.1,

$$\sup_{\tau \in \mathbb{R}} \left(\log \mathbb{P}\{Y_1 \geq \tau\} + \log \mathbb{P}\{X_1 \leq \tau\}\right) \geq -(1 + o(1))\alpha_n + o(\log n).$$

On the other hand, from the way $|D|$ was constructed, we also have $|D| \geq n/3 - k \geq n/4$ under the assumption $k < n/12$. Hence

$$\log n - \log 4 - 2\left((1 + o(1))\alpha_n + o(\log n)\right) = (1 + o(1))(\log n - 2\alpha_n) - \log 4 \to -\infty.$$

Conclude that $\liminf_{n\to\infty} \frac{2\alpha_n}{\log n} \geq 1$ is necessary for $\mathbb{P}\{\widehat{x}_{\mathrm{ML}} = x^*\} \to 1$.

4.2.4 Counting difference graphs

To prove 4.1, we begin with some notations. For a $2k$-NN graph x, let $E_{\mathrm{red}}(x)$ (resp. $E_{\mathrm{blue}}(x)$) be the set of red (resp. blue) edges in $G(x)$. The proof strategy is to first enumerate

$$\mathcal{E}_{\mathrm{red}}(\Delta) = \{E_{\mathrm{red}}(x) : x \in \mathcal{X}_\Delta\}. \qquad \text{(Lemma 4.3)}$$

Then for each $E_{\mathrm{red}} \in \mathcal{E}_{\mathrm{red}}(\Delta)$, enumerate

$$\mathcal{X}_\Delta(E_{\mathrm{red}}) = \{x \in \mathcal{X}_\Delta : E_{\mathrm{red}}(x) = E_{\mathrm{red}}\}.$$

which contains all sets of blue edges that are compatible with E_{red} (4.4). This completely specifies the difference graph $G(x)$, and hence the $2k$-NN graph x.

For a given $2k$-NN graph x associated with the permutation σ, let $\mathcal{N}_x(i)$ denote the set of neighbors of i in x. Let $d_x(i, j) = \min\{|\sigma(i) - \sigma(j)|, n - |\sigma(i) - \sigma(j)|\}$, which is the distance between i and j on the Hamiltonian cycle defined by σ. It is easy to check that d_x is a well-defined metric on $[n]$. For the hidden $2k$-graph x^*, define $\mathcal{N}_{x^*}(\cdot)$ and $d_{x^*}(\cdot, \cdot)$ accordingly.

Definition 4.4. *In the $2k$-NN graph x^*, define the distance between two edges $e = (i, \widetilde{i})$ and $f = (j, \widetilde{j})$ as*

$$d(e, f) = \min\{d_{x^*}(i, j), d_{x^*}(i, \widetilde{j}), d_{x^*}(\widetilde{i}, j), d_{x^*}(\widetilde{i}, \widetilde{j})\}.$$

We say e and f are nearby *if $d(e, f) \leq 2k$.*

Since a $2k$-NN graph has a total of kn edges, the cardinality of $\mathcal{E}_{\mathrm{red}}(\Delta)$ is at most $\binom{kn}{\Delta}$. The following lemma provides additional structural information for elements of $\mathcal{E}_{\mathrm{red}}(\Delta)$ that allows us to improve this trivial bound.

Lemma 4.2. *For each red edge e in the difference graph G, there exists a nearby red edge f distinct from e in G.*

Proof. We divide the proof into two cases according to the degree of one of the endpoints of $e = (i, \widetilde{i})$, say i, in the difference graph.

1. The degree of i is strictly larger than 2. Then by balancedness the number of red edges attached to i is at least 2. Other than $(i, \widetilde{i})$, there must exist at least one other red edge (i, i'). By definition

$$d((i, \widetilde{i}), (i, i')) \leq d_{x^*}(i, i) = 0 < 2k.$$

That is, (i, i') and $(i, \widetilde{i})$ are nearby. See Figure 4.4a on page 87.

2. The degree of i is equal to 2. Then i is only attached to one red edge and one blue edge in G. Denote the blue edge as (i, j). Since the only red edge attached to i is $(i, \widetilde{i})$, we have that in the proposed solution x, the vertex i is connected to all its old neighbors in x^* except $\widetilde{i}$. Deduce that $\mathcal{N}_x(i) = \mathcal{N}_{x^*}(i) \cup \{j\} \backslash \{\widetilde{i}\}$. As a result, out of the two vertices j_1, j_2 that are right next to j in the x cycle ($d_x(j, j_1) = d_x(j, j_2) = 1$), at least one of them is an old neighbor of i. WLOG say $j_1 \in \mathcal{N}_{x^*}(i)$. Consider these cases:

   (a) $d_{x^*}(j, j_1) \leq k$. By triangle inequality $d_{x^*}(j, i) \leq d_{x^*}(j, j_1) + d_{x^*}(i, j_1) \leq 2k$. Because G is a balanced graph, there is at least one red edge $(j, \widetilde{j})$ attached to j, and

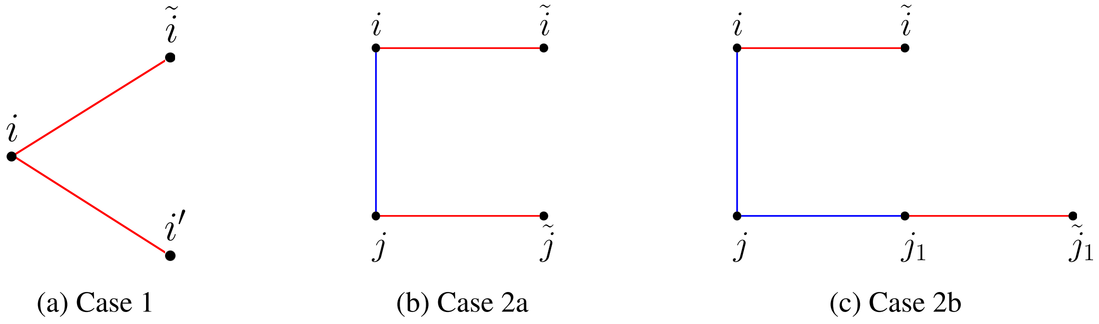
   $$d((i, \widetilde{i}), (j, \widetilde{j})) \leq d_{x^*}(j, i) \leq 2k.$$

   In other words, $(j, \widetilde{j})$ and $(i, \widetilde{i})$ are nearby. See Figure 4.4b on page 87.

   (b) $d_{x^*}(j, j_1) > k$. In this case (j, j_1) appears in the difference graph as a blue edge. Therefore j_1 is one of the vertices in G and attached to at least one red edge $(j_1, \widetilde{j_1})$. Recall that $j_1 \in \mathcal{N}_{x^*}(i)$. Therefore

   $$d((i, \widetilde{i}), (j_1, \widetilde{j_1})) \leq d_{x^*}(i, j_1) \leq k.$$

   In other words, $(j_1, \widetilde{j_1})$ and $(i, \widetilde{i})$ are nearby. See Figure 4.4c on page 87.



(a) Case 1 (b) Case 2a (c) Case 2b

Figure 4.4: Three cases considered in the proof of 4.2.

□

The following lemma gives an upper bound for the size of $\mathcal{E}_{\text{red}}(\Delta)$ and it is a direct consequence of 4.2.

Lemma 4.3.

$$|\mathcal{E}_{\text{red}}(\Delta)| \leq (96k^2)^{\Delta}\binom{kn}{\lfloor \Delta/2 \rfloor}.$$

Proof. To each member E_{red} of $\mathcal{E}_{\text{red}}(\Delta)$, we associate an undirected graph $\widetilde{G}(E_{\text{red}})$ with vertex set E_{red} and edge set $\mathcal{E}(E_{\text{red}})$, such that for $e, f \in E_{\text{red}}$, $(e, f) \in \mathcal{E}(E_{\text{red}})$ if e and f are nearby per 4.4. It suffices to enumerate all E_{red} for which $\widetilde{G}(E_{\text{red}})$ is compliant with the structural property enforced by 4.2. Our enumeration scheme is as follows:

1. Fix $m \in [\Delta]$ to be the number of connected components of $\widetilde{G}(E_{\text{red}})$. Select $\{e_1, \ldots, e_m\}$ from the edge set of x^*. Since x^* is a $2k$-NN graph with kn edges, there are $\binom{kn}{m}$ ways to select this set.

2. Let $\Delta_1, \ldots, \Delta_m$ be the sizes of the connected components $C_1, \ldots, C_m$ of $\widetilde{G}(E_{\text{red}})$. Since $\Delta_i \geq 1$ and $\sum \Delta_i = \Delta$, the total number of such (Δ_i) sequences is $\binom{\Delta-1}{m-1}$, as each sequence can be viewed as the result of replacing $m - 1$ of the "+" symbols with "," in the expression $\Delta = 1 + 1 + \ldots + 1 + 1$.

3. For each C_i, there is at least one spanning tree T_i. Since C_i and T_i share the same vertex set, it suffices to enumerate T_i. First enumerate the isomorphism class of T_i, that is, count the total number of unlabeled rooted trees with of Δ_i vertices. From Otter (1948), there are at most 3^{Δ_i} such unlabeled trees.

4. For $i = 1, \ldots, m$, let e_i be the root of T_i. Enumerate the ways to label the rest of tree T_i. To start, label the vertices on the first layer of T_i, that is, the children of e_i. A red edge f being a child of e_i on T_i means f and e_i are nearby, limiting the number of labels to at most $16k^2$. To see why, note that at least one endpoint of f is of d_{x^*}

distance at most $2k$ from one of the endpoints of e_i. No more than $8k$ vertices fit this description. The other endpoint of f can then only choose from $2k$ vertices because f is in the edge set of x^*.

The remaining layers of T_i can be labeled similarly, with at most $16k^2$ possibilities to label each vertex. In total there are at most $(16k^2)^{\Delta_i - 1}$ to label T_i.

This enumeration scheme accounts for all members of $\mathcal{E}_{\text{red}}(\Delta)$. By 4.2, $\widetilde{G}$ does not contain singletons, *i.e.* $\Delta_i \geq 2$ for all i. Thus $m \leq \lfloor \Delta/2 \rfloor$, and

$$\begin{aligned}|\mathcal{E}_{\text{red}}(\Delta)| \leq &\sum_{m \leq \lfloor \Delta/2 \rfloor} \binom{kn}{m}\binom{\Delta - 1}{m - 1} \prod_{i \leq m} 3^{\Delta_i}(16k^2)^{\Delta_i - 1} \\ \leq &\binom{kn}{\lfloor \Delta/2 \rfloor} 2^{\Delta-1} 3^{\Delta} (16k^2)^{\Delta} \leq (96k^2)^{\Delta} \binom{kn}{\lfloor \Delta/2 \rfloor}.\end{aligned}$$

□

Lemma 4.4. *For each $E_{\text{red}} \in \mathcal{E}_{\text{red}}(\Delta)$,*

$$|\mathcal{X}_{\Delta}(E_{\text{red}})| \leq 2(32k^3)^{2\Delta}\Delta^{\Delta/k}.$$

Proof. For a given permutation σ, let $x(\sigma)$ denote the corresponding $2k$-NN graph. Hereafter the dependence on σ is suppressed when it is clear from the context.

It suffices to enumerate all σ such that $E_{\text{red}}(x(\sigma)) = E_{\text{red}}$. WLOG assume that $\sigma(1) = 1$ and for the ground truth x^*, $\sigma^*(i) \equiv i$. The following is the outline of our enumeration scheme:

1. Enumerate all possibilities for the set $\mathcal{N}_x(1) = \{\sigma(n-k+1), \ldots, \sigma(n), \sigma(2), \ldots, \sigma(k+1)\}$.

2. With $\mathcal{N}_x(1)$ determined, enumerate all possibilities for the values of $(\sigma(n - k + 1), \ldots, \sigma(n), \sigma(2), \ldots, \sigma(k + 1))$.

3. For i from 1 to $n-2k-1$, enumerate $\sigma(i+k+1)$ sequentially, assuming at step i that σ were determined from $\sigma(n-k+1)$ up to $\sigma(i+k)$.

Now we give the details on how cardinality bounds are obtained for each step of the enumeration scheme.

Step 1: Decompose $\mathcal{N}_x(1)$ according to the set of true neighbors and false neighbors. The set of true neighbors $\mathcal{N}_x(1) \cap \mathcal{N}_{x^*}(1)$ is determined by the set of red edges in G. Indeed, this set consists of all members $i \in \mathcal{N}_{x^*}(1)$ for which $(1, i)$ is *not* a red edge.

The set $\mathcal{N}_x(1) \backslash \mathcal{N}_{x^*}(1)$ cannot be read directly from the set of red edges. However we know all members of this set must be connected to 1 via a blue edge. Hence $\mathcal{N}_x(1) \backslash \mathcal{N}_{x^*}(1)$ is a subset of $\mathcal{V}(G)$, the vertex set of G. In addition $\mathcal{V}(G)$ is known and $|\mathcal{V}(G)| \leq 2\Delta$. In fact $\mathcal{V}(G)$ is the vertex set of the subgraph of G induced by the Δ red edges, from balancedness of G. The number of possibilities for $\mathcal{N}_x(1) \backslash \mathcal{N}_{x^*}(1)$ does not exceed the number of subsets of $\mathcal{V}(G)$, which is at most $2^{2\Delta}$.

Step 2: With the set $\mathcal{N}_x(1)$ determined, we next enumerate all ways to place the elements in $\mathcal{N}_x(1)$ are on the cycle represented by σ. That is, we specify the sequence $(\sigma(n-k+1), \ldots, \sigma(n), \sigma(2), \ldots, \sigma(k+1))$, or equivalently, specify $\sigma^{-1}(j)$ for all $j \in \mathcal{N}_x(1)$.

Start with $\mathcal{N}_x(1) \cap \mathcal{V}(G)^C$. We claim that for all j, j_1, j_2 in this set, $d_x(1, j)$ and $d_x(j_1, j_2)$ are completely determined by $\mathcal{N}_x(1)$. Therefore the sequence $(\sigma^{-1}(j) : j \in \mathcal{N}_x(1) \cap \mathcal{V}(G)^C)$ is determined up to a symmetric flip around 1, contributing a factor of 2.

To see why $d_x(1, j)$ is determined from $\mathcal{N}_x(1)$, note that for all $j \in \mathcal{N}_x(1)$,

$$d_x(1, j) = 2k - 1 - |\mathcal{N}_x(1) \cap \mathcal{N}_x(j)|\,.$$

Both $\mathcal{N}_x(1)$ and $\mathcal{N}_x(j)$ are determined. The latter is because $j \in \mathcal{V}(G)^C$, thus $\mathcal{N}_x(j) = \mathcal{N}_{x^*}(j)$.

Similarly $\mathcal{N}_x(j_1) = \mathcal{N}_{x^*}(j_1)$ and $\mathcal{N}_x(j_2) = \mathcal{N}_{x^*}(j_2)$, which allows to determine $d_x(j_1, j_2)$:

$$d_x(j_1, j_2) = \begin{cases} 2k - 1 - |\mathcal{N}_x(j_1) \cap \mathcal{N}_x(j_2)| & \text{if } j_2 \in \mathcal{N}_x(j_1)\,; \\ 2k + 1 - |\mathcal{N}_x(j_1) \cap \mathcal{N}_x(j_2)| & \text{otherwise.} \end{cases}$$

Next handle all $j \in \mathcal{N}_x(1) \cap \mathcal{V}(G)$. Note that $\sigma^{-1}(j) \in \{n-k+1, \ldots, n, 2, \ldots, k+1\}$ because $j \in \mathcal{N}_x(1)$. Among those $2k$ possible values, some are already taken by $\{\sigma^{-1}(j) : j \in \mathcal{N}_x(1) \cap \mathcal{V}(G)^C\}$, leaving $|\mathcal{N}_x(1) \cap \mathcal{V}(G)|$ values to which all $j \in \mathcal{N}_x(1) \cap \mathcal{V}(G)$ are to be assigned. The number of possible assignments is bounded by $|\mathcal{N}_x(1) \cap \mathcal{V}(G)|! \leq (2k)^{2\Delta}$. The inequality is because $\mathcal{N}_x(1) = 2k$ and $\mathcal{V}(G) \leq 2\Delta$.

Overall the cardinality bound induced by step 2 is

$$2 \cdot 2^{2\Delta} \cdot (2k)^{2\Delta} = 2(4k)^{2\Delta}.$$

Step 3: In the previous two steps the values of $(\sigma(n-k+1), \ldots, \sigma(k+1))$ have been determined. Determined with that are the blue edges between members of $\{\sigma(n-k+1), \ldots, \sigma(k+1)\}$. That is because $(\sigma(i), \sigma(j))$ is a blue edge if and only if $d_{x^*}(i, j) \leq k$ and $d_{x^*}(\sigma(i), \sigma(j)) > k$. Denote this set of blue edges as $E_{\text{blue}}^{(0)}$, which can be empty. Recall that the total number of blue edges in G is Δ. If $|E_{\text{blue}}^{(0)}|$ is already Δ, then the enumeration scheme is complete because x is completely specified by the difference graph. Otherwise we determine the value of $\sigma(i+k+1)$ sequentially, starting from $i = 1$. At step i we assign a value for $\sigma(i+k+1)$ and update the set of determined blue edges accordingly. We repeat this process until the quota of Δ blue edges is met.

At step i, all of $\sigma(n-k+1), \ldots, \sigma(i+k)$ have been determined. Denote the set of blue edges between members of $\{\sigma(n-k+1), \ldots, \sigma(i+k)\}$ as $E_{\text{blue}}^{(i-1)}$. Unless $|E_{\text{blue}}^{(i-1)}| = \Delta$, specify $\sigma(i+k+1)$ as follows.

Discuss three cases split according to the red degree of $\sigma(i+1)$, *i.e.* the number of red edges attached to $\sigma(i + 1)$. Note that one of these cases must occur or else balancedness of G would be violated.

1. (Figure 4.5 on page 92) The red degree of $\sigma(i+1)$ is zero, meaning that $\mathcal{N}_x(\sigma(i+1)) = \mathcal{N}_{x^*}(\sigma(i + 1))$. At step i, all but one members of $\mathcal{N}_x(\sigma(i + 1))$ are determined, and $\sigma(i+k+1)$ has to be the true neighbor of $\sigma(i+1)$ that has not appeared in the previous steps. Thus there is only a single choices for $\sigma(i + k + 1)$.

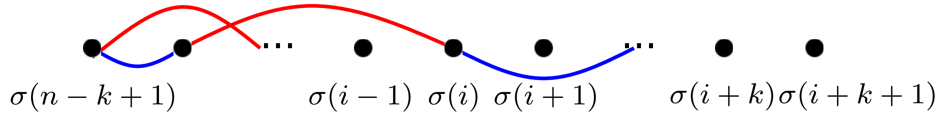


Figure 4.5: Vertices arranged by their order on the cycle corresponding to σ. At step i of the procedure, the values of $\sigma(n - k + 1)$ to $\sigma(i + k)$ are determined. The figure shows an example of case 1: the vertex $\sigma(i + 1)$ is not attached to any red edges.

2. (Figure 4.6 on page 92) The red degree of $\sigma(i + 1)$ is nonzero and equals the number of blue edges in $E^{(i-1)}_{\text{blue}}$ attached to $\sigma(i+1)$. In this case by balancedness all blue edges attached to $\sigma(i+1)$ are contained in $E^{(i-1)}_{\text{blue}}$ and therefore the edge $(\sigma(i+1), \sigma(i+k+1))$ does not appear in the difference graph G. That implies $\sigma(i + k + 1)$ is connected to $\sigma(i + 1)$ in x^*, limiting the number of choices for $\sigma(i + k + 1)$ to at most $2k$.

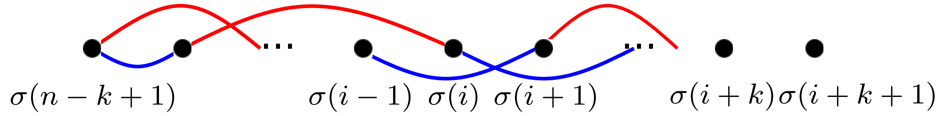


Figure 4.6: Case 2: $\sigma(i+1)$ is attached to some red edge(s) and is already balanced at step i. In the figure the red degree and blue degree of $\sigma(i+1)$ are both 1, thus $(\sigma(i+1), \sigma(i+k+1))$ cannot be a blue edge.

3. (Figure 4.7 on page 93) The red degree of $\sigma(i + 1)$ is nonzero and is one plus the number of blue edges in $E^{(i-1)}_{\text{blue}}$ attached to $\sigma(i + 1)$. By balancedness $(\sigma(i + 1), \sigma(i + k + 1))$ has to appear in G as a blue edge. In this case, either at least one of

$\{(\sigma(i+j), \sigma(i+k+1))\}_{2\le j\le k}$ is not a blue edge in G, or all of $\{(\sigma(i+j), \sigma(i+k+1))\}_{2\le j\le k}$ are blue edges in G. Suppose this is the t'th time case 3 happens. Let ξ_t encode which of the two possibilities occurs and specify the value $\sigma(i+k+1)$ as follows:

(a) Let $\xi_t = 0$ and specify $\sigma(i+k+1)$ such that at least one of $\{(\sigma(i+j), \sigma(i+k+1))\}_{2\le j\le k}$ is not a blue edge in G. In this case $\sigma(i+k+1)$ is a true neighbor of at least one of $\{\sigma(i+2), \ldots, \sigma(i+k)\}$, *i.e.*, choose $\sigma(i+k+1) \in \cup_{2\le j\le k}\mathcal{N}_{x^*}(i+j)$. The number of choices is at most $2k(k-1)$.

(b) Let $\xi_t = 1$ and specify $\sigma(i+k+1)$ such that all of $\{(\sigma(i+j), \sigma(i+k+1))\}_{2\le j\le k}$ are blue edges in G. Combined with $(\sigma(i+1), \sigma(i+k+1))$, in this case the value of $\sigma(i+k+1)$ determines k blue edges that have not appeared in $E^{(i-1)}_{\text{blue}}$. Here $\sigma(i+k+1)$ can choose from at most $|\mathcal{V}(G)| \le 2\Delta$ vertices.

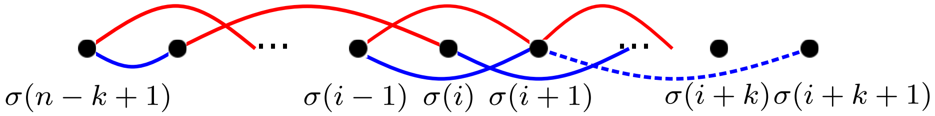


Figure 4.7: Case 3: $\sigma(i+1)$ is attached to some red edge(s) and is not already balanced at step i. In the figure $\sigma(i+1)$ has red degree 2 and blue degree 1. Therefore $(\sigma(i+1), \sigma(i+k+1))$ must appear G as a blue edge.

Repeat this process until $|E^{(i)}_{\text{blue}}| = \Delta$. In total we would encounter case 3b) at most $\lfloor \Delta/k \rfloor$ times because k blue edges get included in the updated blue edge set each time. Also, case 2) and case 3a) combined can occur at most 2Δ times, because they only occur when $\sigma(i+1) \in \mathcal{V}(G)$ (this also guarantees that the ξ sequence of of length at most 2Δ). Overall after step 2, the total number of ways to specify the difference graph is at most

$$\sum_{\xi\in\{0,1\}^{2\Delta}} (2k(k-1))^{2\Delta}(2\Delta)^{\Delta/k}$$
$$\le \left(8k^2\right)^{2\Delta} \Delta^{\Delta/k}.$$

Combined with the cardinality bounds from step 1 and step 2, we have

$$|\mathcal{X}_\Delta(E_{\mathrm{red}})| \leq 2(4k)^{2\Delta} \cdot (8k^2)^{2\Delta}\Delta^{\Delta/k} = 2(32k^3)^{2\Delta}\Delta^{\Delta/k}.$$

□

Finally, Lemma 4.1 follows from combining 4.3 and 4.4:

$$|\mathcal{X}_\Delta| \leq (96k^2)^\Delta \binom{kn}{\lfloor \Delta/2 \rfloor} \cdot 2(32k^3)^{2\Delta}\Delta^{\Delta/k} \leq 2\left(Ck^{17}n\right)^{\Delta/2}$$

for a universal constant $C > 0$.

Bibliography

Bagaria, V., J. Ding, D. Tse, Y. Wu, and J. Xu (2018). Hidden hamiltonian cycle recovery via linear programming. *arXiv preprint arXiv:1804.05436*. 76, 79, 80

Cai, T. T., T. Liang, and A. Rakhlin (2017). On detection and structural reconstruction of small-world random networks. *IEEE Transactions on Network Science and Engineering 4*(3), 165–176. 77, 78

Lieberman-Aiden, E., N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science 326*(5950), 289–293. 76

Newman, M. E. and D. J. Watts (1999). Scaling and percolation in the small-world network model. *Physical review E 60*(6), 7332. 76

Otter, R. (1948). The number of trees. *Annals of Mathematics 49*(3), 583–599. 88

Putnam, N. H., B. L. O'Connell, J. C. Stites, B. J. Rice, M. Blanchette, R. Calef, C. J. Troll, A. Fields, P. D. Hartley, and C. W. Sugnet (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome research 26*(3), 342–350. 76

Watts, D. J. and S. H. Strogatz (1998). Collective dynamics of 'small-world' networks. *nature 393*(6684), 440. 76