DOES BENIGN OVERFITTING EXIST IN BAYESIAN LINEAR REGRESSION?

Ziheng Cheng, Hengzhi He, Puheng Li, Yang Tian (in α - β order)

ABSTRACT

Recently the phenomenon of benign overfitting has attracted much attention. Most previous works focus on the explanations for the benign overfitting in vanilla and ridge linear regression. In this paper, we study whether benign overfitting exists in Bayesian linear regression. To this end, after deriving the mathematical bound of generalisation error, we find surprisingly that benign overfitting will vanish if we take a suitable prior! That is, if we use the "best" way interpreted from our computations to learn our model in highly over-parametrised regimes , components of the data matrix in unimportant directions with small variance will not help. Inspired by our results, we design two algorithms, Hard Margin Algorithm and Soft Margin Algorithm, to help find the components in the most important directions. Experiments show that our algorithms are promising and may perform better than traditional estimators like minimum norm interpolation estimator and ridge estimator.

1 INTRODUCTION

Deep neural networks have played a vital role in scientific computing, computer vision, and natural language processing applications. One key factor that makes them successful is their abilities to achieve excellent generalisation even in highly over-parameterised situations. (See figure 1 from Zhang et al. (2017)) Inspired by this phenomenon, Bartlett et al. (2020) investigated the "benign overfitting phenomenon" in over-parameterised linear models and found that in linear models, over-parameterisation plays an important role in benign overfitting: a sufficient number of insignificant directions can play a certain kind of regularisation against model noise and reduce generalisation errors Bartlett et al. (2021);

More specifically,we consider the following setting: the linear model $y = x^T w + \epsilon$. We define the least square estimator with minimum norm as $\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\{\|\theta\| : \theta \in \underset{w}{\operatorname{argmin}} \Sigma_{i=1}^{n}(y_i - x_i^T w)^2\}$ and define the excess risk as $R_p(\theta) = \mathbb{E}_{x,y}[(y - x^T\theta)^2 - (y - x^T\theta^*)^2]$. If there exists a series of samples of size n(p) such that the dimension of parameters $p > n(p)$,and the excess risk $R_p(\hat{\theta}) \to 0$ as $p \to \infty$ (Here $\hat{\theta}$ is learnt through $n(p)$ samples), then we say there exists benign overfitting in linear regression.

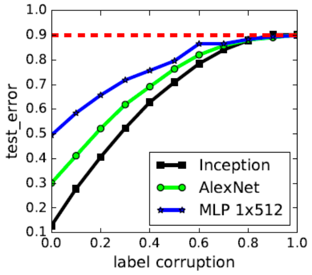

Figure 1: Deep network architectures using SGD give respectable predictions after fitting a standard image classification training set, even when significant levels of label noise are introduced.

Bartlett et al. (2020) establishes the following theorem:

Theorem 1. *For the covariance operator Σ, we define $\lambda_i = \mu_i(\Sigma)$ for $i = 1, 2, \ldots$. Then under some conditions and with probability at least $1-\delta$, we have*

$$R(\hat{\theta}) \leq c(||\theta^*||^2||\Sigma|| \max\{\sqrt{\frac{r_0(\sigma)}{n}, \frac{r_0(\sigma)}{n}, \frac{\log\frac{1}{\delta}}{n}}\}) + c\log(\frac{1}{\delta})\sigma_y^2(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)})$$

where $r_k(\Sigma) = \dfrac{\sum_{i>k}\lambda_i}{\lambda_{k+1}}$, $R_k(\Sigma) = \dfrac{\left(\sum_{i>k}\lambda_i\right)^2}{\sum_{i>k}\lambda_i^2}$, $k^* = \min\left\{k \geq 0 : r_k(\Sigma) \geq bn\right\}$.

The theorem shows that to achieve benign overfitting when $n \to \infty$ and $n(p) \to \infty$ with $n(p) \ll p$, the eigenvalues of $\Sigma = \mathbb{E}xx^T$ are expected to decay slowly enough, which means that there exists a lot of unimportant directions. These unimportant directions can help to make $R_k(\Sigma)$ small and achieve benign overfitting.

In this paper, inspired by Bartlett et al. (2020) and Bayesian linear regression, we study if benign overfitting exists in Bayesian linear regression, and find surprisingly that benign overfitting phenomenon will vanish if we take a suitable prior! That is, if we use the "best" way to learn our model in highly over-parametrised regime, more unimportant directions with small variance will not help. Motivated by our analysis, we further design two kinds of algorithms to help find the most important directions given the number of samples we can use. Numerical experiments show that our algorithms are promising and deserve studying further.

2 Background and Notations

2.1 Background

We consider **Bayesian linear regression** problems, with the regression model:

$$y|X, \theta, \sigma^2 \sim \mathcal{N}(X\theta, \sigma^2 I_n)$$

where y is a column vector of n observations for the outcome variable, X is an $n \times p$ matrix of observated predictors. We set a prior distribution $\theta \sim \mathcal{N}_{p+1}(\mu_0, \Lambda_0)$ and $\Lambda_0 = \text{diag}\{\tau_0^2, \tau_1^2, \cdots, \tau_p^2\}$. Therefore, we can derive the posterior distribution of θ has the following closed form :

$$\theta|X, y, \sigma^2 \sim \mathcal{N}(\mu_n, \Lambda_n)$$

$$\mu_n = (X_*^T\Sigma_*^{-1}X_*)^{-1}X_*^T\Sigma_*^{-1}y_*, \quad \Lambda_n = (X_*^T\Sigma_*^{-1}X_*)^{-1}$$

$$X_* = \begin{pmatrix} x \\ I_p \end{pmatrix}, \quad y_* = \begin{pmatrix} y \\ \mu_0 \end{pmatrix}, \quad \Sigma_* = \begin{pmatrix} \sigma^2 I_n & 0 \\ 0 & \Lambda_0 \end{pmatrix}$$

Note that $\mu_n = (X_*^T\Sigma_*^{-1}X_*)^{-1}X_*^T\Sigma_*^{-1}y_* = (X^TX + \sigma^2\Lambda_0^{-1})^{-1}X^Ty$.

2.2 Related works

This section will review the overparametrization phenomenon in deep learning and summarise the main works in benign overfitting.

In statistical theory Ruppert (2004), there is a trade-off between the fit to the training set and the complexity of prediction rules. However, it is the fact that some learning algorithms generalises well despite interpolating noisy data, and people call this fact as *Benign Overfitting* Bartlett et al. (2020). For instance, the experiment in Zhang et al. (2017) shows that standard deep network utilises stochastic gradient descent (SGD) to fit a training data set with significantly noised labels and demonstrates a respectable prediction performance.

Recent works about benign overfitting can be classified into two categories : (1) **Benign Overfitting in Regression problems**: This type of works concentrates on the mathematical properties of benign overfitting in regression. The most elementary setting that has been analysed theoretically is linear regression with Gaussian noise. Upper bounds and lower bounds for the generalisation error of the ordinary least square estimators have been obtained by Bartlett et al. (2020); Negrea et al. (2020). Besides, Wang et al. (2021) proves that in multi-class linear regression, good generalisation is possible for SVM solutions beyond the realm in which typical margin-based bounds apply. Some works

Lecu'e & Shang (2022); Zhou et al. (2021) extend the study of benign overfitting to ridge regression. Tsigler & Bartlett (2020) derives non-asymptotic upper bounds and lower bounds for the prediction mean squared error (MSE). It is stressed that our study is highly correlated to the Tsigler & Bartlett (2020), since our closed-form predictor can be regarded as implementing different regularisation terms in all directions, bringing us more insight beyond ridge regression. (2) **Questioning Benign Overfitting**: This type of works question the effect of benign overfitting in some certain settings. Shamir (2022) shows that benign overfitting can be easily broken when the underlying model is misspecified, leading us to rethink. Besides, Sanyal et al. (2021) demonstrates that overfitting noisy data can be severely detrimental to robustness, making the models vulnerable to adversarial attacks. Our study also shares some common insights with Sanyal et al. (2021) : we both summarise that some redundant noisy data can be overlooked with scarce sacrifice.

2.3 Notations

We are interested in whether benign overfitting really happens in Bayesian Linear Regression, so throughout the whole paper we make the following assumptions:

- $X \in \mathbb{R}^{n\times p}$ —— a random matrix with i.i.d. centered rows and $p >> n$.
- Denote a row of X is x^T, $x \in \mathbb{R}^p$. The true model is $y = x^T\theta^* + \epsilon$, where θ^* is the unknown ground truth and ϵ is noise that is independent of X and has a sub-gaussian norm bounded by σ_ϵ.
- Assume a row of X is x and the covariance matrix of x is $\mathbb{E}xx^T = \Sigma =$ diag $\{\mu_1, \cdots, \mu_p\}$ with the order $\mu_1 \geq \cdots \geq \mu_p$ (We can assume the covariance matrix is diagonal with out loss of generality since we can achieve this by a change of the basis).

Recall that a random variable Z is sub-gaussian if it has a finite norm $\|Z\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}\exp(Z^2/t^2) \leq 2\}$ and that the sub-gaussian norm of a random vector Z is $\|Z\|_{\psi_2} := \sup\limits_{s\neq 0} \|\langle s, Z\rangle/\|s\|\|_{\psi_2}$.

We further introduce the following notations:

- $a \lesssim b$ if there exists an absolute constant c such that $a \leq cb$.
- For any positive semidefinite (PSD) matrix $M \in \mathbb{R}^{m\times m}$ and any $x \in \mathbb{R}^m$ we denote $\|x\|_M := \sqrt{x^\top Mx}$.
- For any matrix $M \in \mathbb{R}^{n\times p}$ we denote $M_{0:k}$ to be the matrix which is comprised of the first k columns of M , and $M_{k:\infty}$ to be the matrix comprised of the rest of the columns of M.
- For any vector $\eta \in \mathbb{R}^p$ we denote $\eta_{0:k}$ to be the vector comprised of the first k components of η , and $\eta_{k:\infty}$ to be the vector comprised of the rest of the coordinates of η.
- Denote $\Sigma_{0:k}$ =diag$\{\mu_1, \cdots, \mu_k\}$, $\Sigma_{k:\infty} =$ diag$\{\mu_{k+1}, \cdots, \mu_p\}$.
- Denote $\sigma^2\Lambda_0^{-1} =$ diag $\{\lambda_1, \cdots, \lambda_p\}$ and $A_k = I_n + X_{k:\infty}$ diag $\left\{\frac{1}{\lambda_{k+1}}, \cdots, \frac{1}{\lambda_p}\right\} X_{k:\infty}^T$.
- Denote $A = I + X$ diag $\{\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_p}\}X^T$ = $A_k + X_{0:k}$ diag $\left\{\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_k}\right\} X_{0:k}^T$
- Denote $\mu_i(A)$ the i-th largest eigenvalue of a positive definite matrix A.

Denote the Bayesian estimator as $\hat{\theta}(X, \zeta) = (X^TX+$ diag $\{\lambda_1, \cdots, \lambda_p\})^{-1}X^T\zeta$, which is also the minimiser of $\min\limits_\theta \|X\theta - \zeta\|^2 + \theta^T \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix} \theta$

We aim to evaluate the MSE of the Bayesian estimator. So for any x who has the same distribution as the row of X (independent of X and ϵ), we define the prediction MSE :

$$\begin{aligned}\mathbb{E}(x^T(\theta-\theta^*))^2|X,y &= \|\theta-(X^TX+\sigma^2\Lambda_0^{-1})^{-1}X^Ty\|_\Sigma^2\\ &\lesssim \|(I_p-(X^TX+\sigma^2\Lambda_0^{-1}))^{-1}X^TX)\theta\|_\Sigma^2\\ &+\|(X^TX+\sigma^2\Lambda_0^{-1})^{-1}X^T\epsilon\|_\Sigma^2\end{aligned}$$

We further denote :

$$\begin{aligned}B &:= \|(I_p-(X^TX+\sigma^2\Lambda_0^{-1}))^{-1}X^TX)\theta\|_\Sigma^2 = \|\theta^*-\hat\theta(X,X\theta^*)\|_\Sigma^2\\ V &:= \|(X^TX+\sigma^2\Lambda_0^{-1})^{-1}X^T\epsilon\|_\Sigma^2 = \|\hat\theta(X,\epsilon)\|_\Sigma^2\end{aligned}$$

as bias and variance terms.

3 Main Results

For simplicity, we regard the high-probability-index as a constant in the following results.

Theorem 2. *There exists an absolute constant $c > 1$, s.t. $\forall\ k < \frac{n}{c}$, with high probability, the following inequalities hold:*

$$\begin{aligned}B \lesssim_{\sigma_x} &\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + L^2\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}}\frac{n^2}{(\mu_1(A_k)\min_{i\le k}\frac{\lambda_i}{\mu_i}+n)^2}\\ &+\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}}(n^2\mu_{k+1}^2+n\sum_{i>k}\mu_i^2)\frac{L^2}{(1+\sum_{i>k}\frac{\mu_i}{\lambda_i})^2(\min_{i>k}\lambda_i)^2}\\ &+\|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{k:\infty}}\frac{L^2(\max_{i\le k}\lambda_i)^2(1+\sum_{i>k}\frac{\mu_i}{\lambda_i})^2}{(\mu_1(A_k)\min_{i\le k}\frac{\lambda_i}{\mu_i}+n)^2}+\|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{k:\infty}}L^2\frac{(\max_{i\le k}\lambda_i)^2}{(\min_{i>k}\lambda_i)^2}(\mu_{k+1}^2+\frac{1}{n}\sum_{i>k}\mu_i^2)\end{aligned} \tag{1}$$

$$\frac{V}{\sigma_\epsilon^2}\lesssim_{\sigma_x} L^2\frac{kn}{(\mu_1(A_k)\min_{i\le k}\frac{\lambda_i}{\mu_i}+n)^2}+L^2\frac{n}{(1+\sum_{i>k}\frac{\mu_i}{\lambda_i})^2}\sum_{i>k}\frac{\mu_i^2}{\lambda_i^2} \tag{2}$$

Proof. By Lemma 2 and Lemma 3 in Appendix B, we derive a bound for B and V that can be applied concentration inequalities to. The concentration part is given by Appendix C. □

Theorem 3. *The bias term B has a tighter bound w.r.t. the second and the fourth term in 1.*

$$\begin{aligned}B \lesssim_{\sigma_x} &\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}}L^2(1+\sum_{i>k}\frac{\mu_i}{\lambda_i})^2[\frac{(n+p)^2(\max_{i>k}\frac{\mu_i}{\lambda_i})^2}{(1+(n+p)\max_{i>k}\frac{\mu_i}{\lambda_i})^2}+\frac{k\ln k}{n}]\\ &+\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}}(n^2\mu_{k+1}^2+n\sum_{i>k}\mu_i^2)\frac{L^2}{(1+\sum_{i>k}\frac{\mu_i}{\lambda_i})^2(\min_{i>k}\lambda_i)^2}\\ &+\frac{L^2\sum_{i\le k}\frac{\lambda_i^2}{\mu_i}\theta_i^{*2}}{(\mu_1(A_k)\min_{i\le k}\frac{\lambda_i}{\mu_i}+n)^2}+\|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{k:\infty}}L^2\frac{(\max_{i\le k}\lambda_i)^2}{(\min_{i>k}\lambda_i)^2}(\mu_{k+1}^2+\frac{1}{n}\sum_{i>k}\mu_i^2)\end{aligned} \tag{3}$$

Proof. We bound the bias term B like in 3 through 15 in Appendix B, and then the concentration part is given by Appendix D. □

Remark 1. *Theorem 2 is derived by similar methodologies in Tsigler & Bartlett (2020). However, we notice that there are some technical details neglected by in Tsigler & Bartlett (2020). Therefore, we manage some technical contribution and derive a tighter bound in Theorem 3.*

4 Discussion of the main results

We present the MSE bound 4 and 5 in Ridge regression Tsigler & Bartlett (2020) and ours 6 and 7. Our technical contribution is shown as 8 and 9.

$$
\begin{aligned}
B \lesssim_{\sigma_x} & \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + L^2\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}}(n^2\mu_{k+1}^2 + n\sum_{i>k}\mu_i^2)\frac{L^2}{(\lambda+\sum_{i>k}\mu_i)^2} \\
& + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{k:\infty}}\frac{L^2(\lambda+\sum_{i>k}\mu_i)^2}{n^2} + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{k:\infty}}L^2(\mu_{k+1}^2+\frac{1}{n}\sum_{i>k}\mu_i^2)
\end{aligned} \tag{4}
$$

$$
\frac{V}{\sigma_\epsilon^2} \lesssim_{\sigma_x} L^2\frac{k}{n} + L^2\frac{n}{(\lambda+\sum_{i>k}\mu_i)^2}\sum_{i>k}\mu_i^2 \tag{5}
$$

$$
\begin{aligned}
B \lesssim_{\sigma_x} & \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + L^2\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}}\frac{n^2}{(\mu_1(A_k)\min_{i\le k}\frac{\lambda_i}{\mu_i}+n)^2} \\
& + \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}}(n^2\mu_{k+1}^2 + n\sum_{i>k}\mu_i^2)\frac{L^2}{(1+\sum_{i>k}\frac{\mu_i}{\lambda_i})^2(\min_{i>k}\lambda_i)^2} \\
& + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{k:\infty}}\frac{L^2(\max_{i\le k}\lambda_i)^2(1+\sum_{i>k}\frac{\mu_i}{\lambda_i})^2}{(\mu_1(A_k)\min_{i\le k}\frac{\lambda_i}{\mu_i}+n)^2} + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{k:\infty}}L^2\frac{(\max_{i\le k}\lambda_i)^2}{(\min_{i>k}\lambda_i)^2}(\mu_{k+1}^2+\frac{1}{n}\sum_{i>k}\mu_i^2)
\end{aligned} \tag{6}
$$

$$
\frac{V}{\sigma_\epsilon^2} \lesssim_{\sigma_x} L^2\frac{kn}{(\mu_1(A_k)\min_{i\le k}\frac{\lambda_i}{\mu_i}+n)^2} + L^2\frac{n}{(1+\sum_{i>k}\frac{\mu_i}{\lambda_i})^2}\sum_{i>k}\frac{\mu_i^2}{\lambda_i^2} \tag{7}
$$

$$
blue \rightarrow \frac{L^2\sum_{i\le k}\frac{\lambda_i^2}{\mu_i}\theta_i^{*2}}{(\mu_1(A_k)\min_{i\le k}\frac{\lambda_i}{\mu_i}+n)^2} \tag{8}
$$

$$
Melon \rightarrow \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}}L^2(1+\sum_{i>k}\frac{\mu_i}{\lambda_i})^2[\frac{(n+p)^2(\max_{i>k}\frac{\mu_i}{\lambda_i})^2}{(1+(n+p)\max_{i>k}\frac{\mu_i}{\lambda_i})^2}+\frac{k\ln k}{n}] \tag{9}
$$

Comparing our results with that in the case of Ridge regression, we can find that **benign overfitting may not exist in Bayesian linear regression**, thanks to the anisotropic regular coefficients. The $magenta$ term in 6 would vanish if we set $\lambda_{k+1},\cdots,\lambda_p\rightarrow\infty$; meanwhile, $blue, SeaGreen$ in 6 would also vanish if $\lambda_1,\cdots,\lambda_k\rightarrow 0$. Similarly, the $Cerulean$ term in variance would also vanish if $\lambda_{k+1},\cdots,\lambda_p\rightarrow\infty$. In fact, the bound does reach the optimal value when $\lambda_1,\cdots,\lambda_k\rightarrow 0,\lambda_{k+1},\cdots,\lambda_p\rightarrow\infty$ .However, this can never occur in Ridge regression.

Note that the upper bound is tight up to some constants. In Ridge regression, the regular coefficient λ can not be too large considering the $blue$ term in 4. Therefore, much more noise directions, which corrsponds to singular values $\mu_{k:\infty}$ are necessary in order to let $magenta$ in 4 and $Goldenrod$ in 5 vanish as $n,p\rightarrow\infty$. Besides, $\mu_{k:\infty}$ must decay slowly. Therefore, benign overfitting exists in the setting of Ridge regression.

However, things would beome totally different in Bayesian linear regression. As we have mentioned before, those terms would all vanish as $\lambda_1,\cdots,\lambda_k\rightarrow 0,\lambda_{k+1},\cdots,\lambda_p\rightarrow\infty$, which illustrates that we may only do regressions with the most important k factors. In this way, there is no need to let those noise directions with singular values $\mu_{k:\infty}$ remain in our model, i.e. benign overfitting does not exist. Also note that the anisotropic regular coefficients in Bayesian linear regression are induced by

$$
\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix} = \sigma_0^2\Lambda_0^{-1},
$$

which serves as prior information in Bayesian framework. If somehow we are blessed with sufficiently informative prior knowledge, then just removing those noise factors and doing regressions with those important factors only would lead to the optimal results. In other words, **benign overfitting may not exist in Bayesian linear regression** given informative prior. This conclusion is also coincident with our intutions.

As for our technical contributions, which are shown in 8 and 9, it is easy to notice that the new bound is strictly tighter. In 9, the $Melon$ term would converge to 0 as $\lambda_{k+1}, \cdots, \lambda_p \to \infty$, which remains as constant in the original bound 6.

Finally, as $\lambda_1, \cdots, \lambda_k \to 0, \lambda_{k+1}, \cdots, \lambda_p \to \infty$, the MSE bound degrades to

$$B \lesssim_{\sigma_x} \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} = \sum_{i>k} \mu_i {\theta_i^*}^2, \tag{10}$$

$$\frac{V}{\sigma_\epsilon^2} \lesssim_{\sigma_x} \frac{k}{n}. \tag{11}$$

5 Adaptive Algorithms

From the analysis given above, if somehow we can choose a suitable k^* (given sufficient prior information), then the optimal choice of λ is to let $\lambda_{k^*:\infty} \to \infty$ and $\lambda_{0:k^*} \to 0$, i.e. we completely ignore the noises from those unimportant directions. Enlightend by this insight, we put forward an adaptive algorithm.

Algorithm 1: Hard margin algorithm

```
Approximate $\mu$ by empirical covariance and $\theta^*$ by vanilla linear regression.
Search an appropriate $k$ and the most important $k$ variables.
Discard the rest $p - k$ variables and run vanilla linear regression on the selected $k$ variables.
```

Based on the above algorithm, we propose a modified algorithm for Bayesian linear regression which does not require a hard margin k.

Algorithm 2: Soft margin algorithm

```
Approximate $\mu$ by empirical covariance and $\theta^*$ by vanilla linear regression.
Set $\lambda_i = g(\mu_i {\theta_i^*}^2)$, where $g$ is a non-increasing function. (Inspired by LeCun et al. (1989))
Run Bayesian linear regression with $\lambda$.
```

In the soft margin algorithm, the Bayesian linear regression is of the form

$$\hat{\theta} = (X^T X + \sigma^2 \Lambda_0^{-1})^{-1} (X^T y + \sigma^2 \Lambda_0^{-1} \mu_0),$$

where $\sigma^2 \Lambda_0^{-1} = \lambda$ and μ_0 is the prior of $\hat{\theta}$ (In our algorithm, it's the estimate of $\hat{\theta}$ in the last batch).

6 Experiment

We conducted several numerical experiments to empirically show the effectiveness of our algorithms.

For the hard margin algorithm, we first did algorithm under setting n = 1000, p = 20000, decaying rate of eigenvalues of X's covariance matrix $\propto i^{2.5}$. Under this setting, our algorithm didn't perform very well compared to ridge regression. However, we find that our theoretical bound (using the true ranking instead of the extimated one) beat ridge regression remarkably (see figure 2).

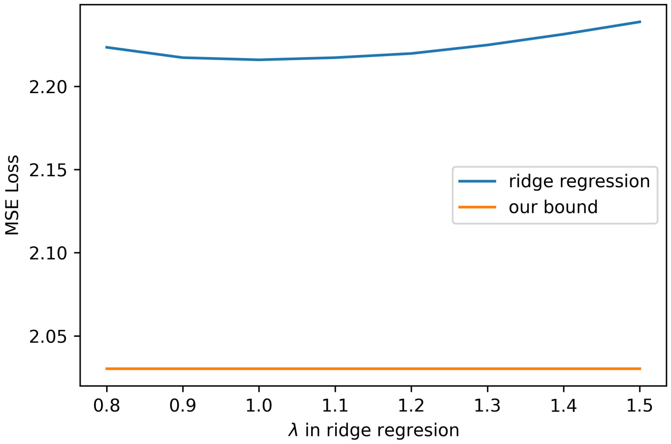

Figure 2: Our bound v.s. Ridge

On the other hand, it is theoretically true that when n increases, central limit theorem will promise that our algorithm will turn closer to our bound, since we use the diagonal elements of the covariance matrix to estimate its eigenvalues.

Then we did experiments when n is larger and it proved our theory empirically.

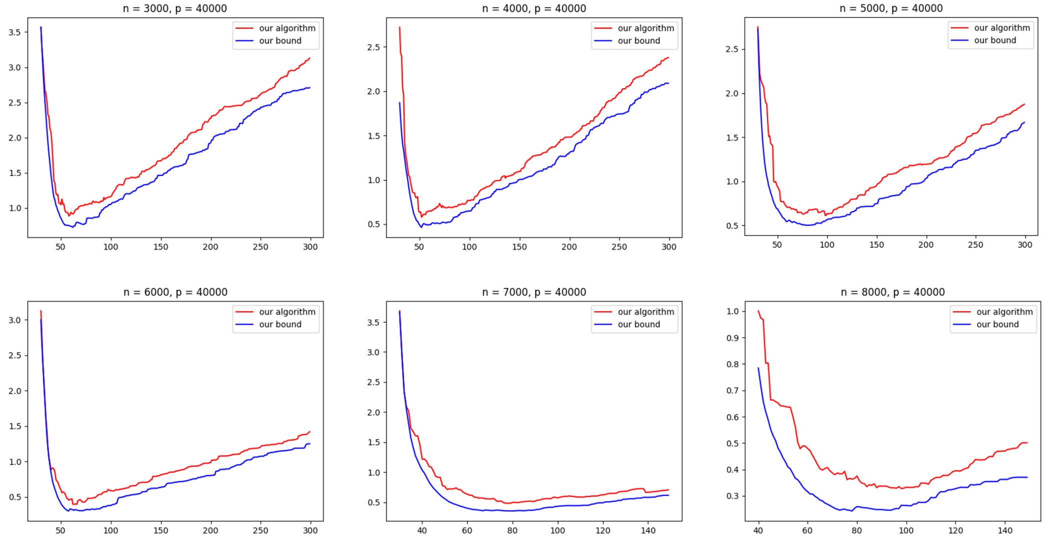

Figure 3: Our algorithm and our bound under different n, p = 40000, decaying rate of eigenvalues of X's covariance matrix $\propto i^{2.5}$

From the figures above, we can see that our algorithm estimates our bound well when n and p are both large, which indicates that our algorithm has its potential under large samples and dimension.

The above experiments imply that hard margin algorithm has its advantage when n is large, but it fails under small n. For the setting when n is not big enough, we adopt soft margin algorithm and it works effectively.

For the soft margin algorithm, we did experiment under setting where n = 1000, p = 20000, batchsize = 250, decaying rate of eigenvalues of X's covariance matrix $\propto i^{0.9}$.

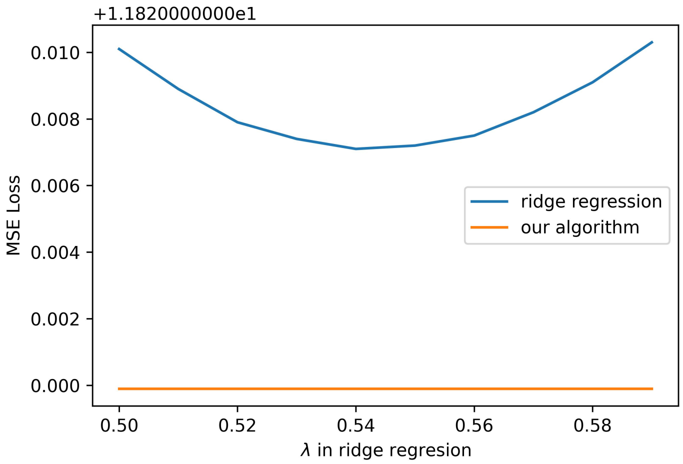

Figure 4: Soft margin v.s. Ridge

From figure 4, we find that our soft margin algorithm beats ridge regression under this setting.

From the result of our experiments, we have the following observations:

- When n is small and the decaying rate of eigenvalues of X's covariance matrix is small, our soft margin algorithm is more effective and it beats ridge regression.
- When n is large (Central Limit Theorem) and the decaying rate of eigenvalues of X's covariance matrix is big (the variables discarded shouldn't be important), our hard margin algorithm may be more effective and it may beat ridge regression remarkably (promised by the impressive lower bound).

7 Conclusions and Future work

In this work, we study whether benign overfitting will exist in Bayesian linear regression. By setting proper priors and imposing different regularisations in diverse directions, we ultimately draw the conclusion that benign overfitting will vanish. We further propose two adaptive algorithms and numerical experiments show that our algorithms may beat ridge regression.

Admittedly, our experiment is not state-of-the-art, but it does shed light on the methods to leverage the prior information or set regular coefficients. We wish to design more efficient and stable algorithms. Also, there are several natural future directions. First, we assume that $\mathbb{E}(y|x)$ is a linear function of x, and it is important to understand whether Bayesian methods would deny benign overfitting in more complex or even implicit forms. Besides, we only consider square-error as the loss function. We are wondering if our conclusions are still valid in the case of other loss function is like cross entropy loss in classification problem. What's more, we'd like to derive the lower bound similarly as the upper bound and design an explicit way to find the optimal k^* to show clearly our bound is indeed tight. Moreover, we would like to apply the insights of our conclusions to algorithms in dimensionality reduction, leading to a better performance.

8 Author Contribution

(in $\alpha - \beta$ order)
Ziheng Cheng: Preprocessing in Theorem 2, Technical improvements in Theorem 3, Notations
Hengzhi He: Concentration in Theorem 2, Technical improvements in Theorem 3, Algorithm
Puheng Li: Concentration in Theorem 2, Algorithm, Experiments
Yang Tian: Preprocessing in Theorem 2, Background, Experiments
Four authors contribute equally to the writing.

REFERENCES

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Guillaume Lecu'e and Zongyuan Shang. A geometrical viewpoint on the benign overfitting property of the minimum l_2-norm interpolant estimator. 2022.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.

Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel M. Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *ICML*, 2020.

David Ruppert. The elements of statistical learning: Data mining, inference, and prediction. *Journal of the American Statistical Association*, 99:567 – 567, 2004.

Amartya Sanyal, Puneet Kumar Dokania, Varun Kanade, and Philip H. S. Torr. How benign is benign overfitting? *ArXiv*, abs/2007.04028, 2021.

Ohad Shamir. The implicit bias of benign overfitting. *arXiv preprint arXiv:2201.11489*, 2022.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. In *NeurIPS*, 2021.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *ArXiv*, abs/1611.03530, 2017.

Lijia Zhou, Frederic Koehler, Danica J. Sutherland, and Nathan Srebro. Optimistic rates: A unifying theory for interpolation learning and regularization in linear regression. *ArXiv*, abs/2112.04470, 2021.

Appendices

Appendix A Algebraic Identity

Lemma 1 (Sherman–Morrison–Woodbury formula)**.** *For matrices A, B, C, D with suitable sizes,*

$$(A + BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1}.$$

Appendix B Preprocessing in Theorem 2

Lemma 2 (Bias term)**.** *The bias term $B = \|\theta^* - \hat{\theta}(X, X\theta^*)\|_{\Sigma}^2$ is bounded as follows,*

$$\begin{aligned} B \lesssim & \frac{\|\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T A_k^{-1} X_{k:\infty}\theta_{k:\infty}^*\|^2 + \|diag\{\lambda_1, \cdots, \lambda_k\}\|^2 \|\theta_{0:k}^*\|^2_{\Sigma_{0:k}^{-\frac{1}{2}}}}{(\min_{i \le k} \frac{\lambda_i}{\mu_i} + \mu_n(A_k^{-1})\mu_k(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\Sigma_{0:k}))^2} \\ & + \|\theta_{k:\infty}\|^2_{\Sigma_{k:\infty}} \\ & + \|diag\left\{\frac{1}{\lambda_{k+1}}, \cdots, \frac{1}{\lambda_p}\right\}\|^2 \|X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^T\| \mu_1(A^{-1})^2 \|X_{k:\infty}\theta_{k:\infty}^*\|^2 \\ & + \frac{\|diag\{\frac{1}{\lambda_{k+1}}, \cdots, \frac{1}{\lambda_p}\}\|^2}{\mu_k(diag\{\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_k}\})^2} L^2 \|X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^T\| \frac{\mu_1(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\Sigma_{0:k}^{-\frac{1}{2}})}{\mu_k(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\Sigma_{0:k}^{\frac{1}{2}})} \cdot \|\Sigma_{0:k}^{-\frac{1}{2}}\theta_{0:k}^*\|^2. \end{aligned}$$

Further in our technical contribution, the first term above is replaced by

$$\frac{\|\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{k:\infty}\theta_{k:\infty}^*\|^2 + \|\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T (I - A_k^{-1} X_{k:\infty}\theta_{k:\infty}^*\|^2 + \|diag\{\lambda_1, \cdots, \lambda_k\}\|^2 \|\theta_{0:k}^*\|^2_{\Sigma_{0:k}^{-\frac{1}{2}}}}{(\min_{i \le k} \frac{\lambda_i}{\mu_i} + \mu_n(A_k^{-1})\mu_k(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\Sigma_{0:k}))^2} \tag{12}$$

Proof. For the first k components, we have

$$\hat{\theta}_{0:k}(X, X\theta^*) + diag\left\{\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_k}\right\} X_{0:k}^T A_k^{-1} X_{0:k}\hat{\theta}_{0:k}(X, X\theta^*) = diag\left\{\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_k}\right\} X_{0:k}^T A_k^{-1} X\theta^*. \tag{13}$$

Let $\zeta = \hat{\theta} - \theta^*$ and left-multiply $\zeta_{0:k}^T \text{diag}\{\lambda_1, \cdots, \lambda_k\}$ in the above identity, and then by some trivial calculation, we derive

$$\begin{aligned} & \underset{i \le k}{min} \frac{\lambda_i}{\mu_i} \|\zeta_{0:k}\|^2_{\Sigma_{0:k}} + \|\zeta_{0:k}\|^2_{\Sigma_{0:k}} \mu_n(A_k^{-1})\mu_k(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\Sigma_{0:k}) \\ \le & \|\zeta_{0:k}\|_{\Sigma_{0:k}} \|\text{diag}\{\lambda_1, \cdots, \lambda_k\}\| \|\theta_{0:k}^*\|_{\Sigma_{0:k}^{-\frac{1}{2}}} \\ + & \|\zeta_{0:k}^T \Sigma_{0:k}^{\frac{1}{2}} \Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T A_k^{-1} X_{k:\infty}\theta_{k:\infty}^*\|. \end{aligned}$$

Thus

$$\begin{aligned} \|\zeta_{0:k}\|_{\Sigma_{0:k}} \lesssim & \frac{\|\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T A_k^{-1} X_{k:\infty}\theta_{k:\infty}^*\|}{min_{i \le k} \frac{\lambda_i}{\mu_i} + \mu_n(A_k^{-1})\mu_k(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\Sigma_{0:k})} \\ & + \frac{\|diag\{\lambda_1, \cdots, \lambda_k\}\| \|\theta_{0:k}^*\|_{\Sigma_{0:k}^{-\frac{1}{2}}}}{min_{i \le k} \frac{\lambda_i}{\mu_i} + \mu_n(A_k^{-1})\mu_k(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\Sigma_{0:k})}. \end{aligned} \tag{14}$$

or further

$$\|\zeta_{0:k}\|_{\Sigma_{0:k}} \lesssim \frac{\|\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{k:\infty}\theta_{0:k}^*\| + \|\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T (I_n - A_k^{-1}) X_{k:\infty}\theta_{k:\infty}^*\|}{min_{i\le k}\frac{\lambda_i}{\mu_i} + \mu_n(A_k^{-1})\mu_k(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\Sigma_{0:k})} + \frac{\|diag\left\{\lambda_1,\cdots,\lambda_k\right\}\|\|\theta_{0:k}^*\|_{\Sigma_{0:k}^{-\frac{1}{2}}}}{min_{i\le k}\frac{\lambda_i}{\mu_i} + \mu_n(A_k^{-1})\mu_k(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\Sigma_{0:k})}. \tag{15}$$

The inequality 14 is inspired by the methodologies in Tsigler & Bartlett (2020), while 15 refers to our technical contribution, which would lead to a tighter bound after concentration.

For the rest components,

$$\begin{aligned}&\|\theta_{k:\infty}^* - \text{diag}\left\{\frac{1}{\lambda_{k+1}},\cdots,\frac{1}{\lambda_p}\right\} X_{k:\infty}^T A^{-1} X\theta^*\|_{\Sigma_{k:\infty}}^2 \\ \lesssim&\|\theta_{k:\infty}^*\|_{\Sigma_{k:\infty}}^2 + \|\text{diag}\left\{\frac{1}{\lambda_{k+1}},\cdots,\frac{1}{\lambda_p}\right\} X_{k:\infty}^T A^{-1} X_{k:\infty}\theta_{k:\infty}^*\|_{\Sigma_{k:\infty}}^2 \\ &+ \|\text{diag}\left\{\frac{1}{\lambda_{k+1}},\cdots,\frac{1}{\lambda_p}\right\} X_{k:\infty}^T A^{-1} X_{0:k}\theta_{0:k}^*\|_{\Sigma_{k:\infty}}^2.\end{aligned}$$

Let's deal with the second term:

$$\begin{aligned}&\|\text{diag}\left\{\frac{1}{\lambda_{k+1}},\cdots,\frac{1}{\lambda_p}\right\} X_{k:\infty}^T A^{-1} X_{k:\infty}\theta_{k:\infty}^*\|_{\Sigma_{k:\infty}}^2 \\ \le&\ \|\text{diag}\left\{\frac{1}{\lambda_{k+1}},\cdots,\frac{1}{\lambda_p}\right\}\|^2\|X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^T\|\|\mu_1(A^{-1})\|^2 + \|X_{k:\infty}\theta_{k:\infty}^*\|^2.\end{aligned}$$

As for the last term, note that $A = A_k + X_{0:k}\ \text{diag}\left\{\frac{1}{\lambda_1},\cdots,\frac{1}{\lambda_k}\right\}X_{0:k}^T$. By Lemma 1,

$$\begin{aligned}&\|\text{diag}\left\{\frac{1}{\lambda_{k+1}},\cdots,\frac{1}{\lambda_p}\right\} X_{k:\infty}^T A^{-1} X_{0:k}\theta_{0:k}^*\|_{\Sigma_{k:\infty}}^2 \\ \le&\|\text{diag}\left\{\frac{1}{\lambda_{k+1}},\cdots,\frac{1}{\lambda_p}\right\}\|^2\|X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^T\|\|\mu_1(A_k^{-1})\|^2\frac{1}{\mu_k(diag\left\{\frac{1}{\lambda_1},\cdots,\frac{1}{\lambda_k}\right\})^2}\|\Sigma_{0:k}^{-\frac{1}{2}}\theta_{0:k}^*\|^2 \\ &\cdot\frac{\mu_1(\Sigma_{0:k}^{\frac{1}{2}}X_{0:k}^T X_{0:k}\Sigma_{0:k}^{\frac{1}{2}})}{\mu_k(\Sigma_{0:k}^{\frac{1}{2}}X_{0:k}^T A_k^{-1}X_{0:k}\Sigma_{0:k}^{\frac{1}{2}})^2}.\end{aligned}$$

Combine all these inequalities and we get Lemma 2. □

Lemma 3 (Variance term)**.** *The variance term $V = \|\hat{\theta}(X,\epsilon)\|_\Sigma^2$ is bounded as follows*

$$\begin{aligned}V \lesssim&\frac{\epsilon^T A_k^{-1} X_{0:k}\Sigma_{0:k}^{-1} X_{0:k}^T A_k^{-1}\epsilon}{(\min_{i\le k}\frac{\lambda_i}{\mu_i} + \mu_n(A_k^{-1})\mu_k(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\Sigma_{0:k}^{-\frac{1}{2}}))^2} \\ &+ \epsilon^T A^{-1} X_{k:\infty} diag\left\{\frac{1}{\lambda_{k+1}},\cdots,\frac{1}{\lambda_p}\right\}\Sigma_{k:\infty} diag\left\{\frac{1}{\lambda_{k+1}},\cdots,\frac{1}{\lambda_p}\right\} X_{k:\infty}^T A^{-1}\epsilon.\end{aligned}$$

Proof. For the first k components, similarly we have

$$\begin{aligned}\hat{\eta}_{0:k} &= diag\{\frac{1}{\lambda_1},\cdots,\frac{1}{\lambda_k}\}X_{0:k}^T A^{-1}\epsilon \\ &= diag\{\frac{1}{\lambda_1},\cdots,\frac{1}{\lambda_k}\}X_{0:k}^T(A_k + X_{0:k}diag\{\frac{1}{\lambda_1},\cdots,\frac{1}{\lambda_k}\}X_{0:k}^T)^{-1}\epsilon.\end{aligned}$$

A key observation by Lemma 1 is that

$$\hat{\eta}_{0:k} + diag\{\frac{1}{\lambda_1},\cdots,\frac{1}{\lambda_k}\}X_{0:k}^T A_k^{-1}X_{0:k}\hat{\eta}_{0:k} = diag\{\frac{1}{\lambda_1},\cdots,\frac{1}{\lambda_k}\}X_{0:k}^T A_k^{-1}\epsilon.$$

Left-multiplying $\hat{\eta}_{0:k}^T \text{diag}\{\lambda_1, \cdots, \lambda_k\}$ induces

$$\hat{\eta}_{0:k}^T diag\{\lambda_1, \cdots, \lambda_k\}\hat{\eta}_{0:k} + \hat{\eta}_{0:k}^T X_{0:k}^T A_k^{-1} X_{0:k}\hat{\eta}_{0:k} = \hat{\eta}_{0:k}^T X_{0:k}^T A_k^{-1}\epsilon, \tag{16}$$

where

$$LHS \geq \min_{i\leq k} \frac{\lambda_i}{\mu_i}\|\hat{\eta}_{0:k}\|_{\Sigma_{0:k}}^2 + \mu_n(A_k^{-1})\|\hat{\eta}_{0:k}\|_{\Sigma_{0:k}}^2 \mu_k(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\Sigma_{0:k}^{-\frac{1}{2}})$$
$$RHS \leq \|\hat{\eta}_{0:k}\|_{\Sigma_{0:k}}\sqrt{\epsilon^T A_k^{-1} X_{0:k}\Sigma_{0:k}^{-1} X_{0:k}^T A_k^{-1}\epsilon}.$$

Thus

$$\|\hat{\eta}_{0:k}\|_{\Sigma_{0:k}}^2 \leq \frac{\epsilon^T A_k^{-1} X_{0:k}\Sigma_{0:k}^{-1} X_{0:k}^T A_k^{-1}\epsilon}{(\min_{i\leq k}\frac{\lambda_i}{\mu_i} + \mu_n(A_k^{-1})\mu_k(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\sigma_{0:k}^{-\frac{1}{2}}))^2}. \tag{17}$$

For the rest components, we can rewrite it as

$$\begin{aligned}&\|\Sigma_{k:\infty}^{\frac{1}{2}}\text{diag}\left\{\frac{1}{\lambda_{k+1}}, \cdots, \frac{1}{\lambda_p}\right\} X_{k:\infty}^T A^{-1}\epsilon\|^2 \\ =&\epsilon^T A^{-1} X_{k:\infty}\text{diag}\left\{\frac{1}{\lambda_{k+1}}, \cdots, \frac{1}{\lambda_p}\right\}\Sigma_{k:\infty}\text{diag}\left\{\frac{1}{\lambda_{k+1}}, \cdots, \frac{1}{\lambda_p}\right\} X_{k:\infty}^T A^{-1}\epsilon.\end{aligned} \tag{18}$$

Add up 17 with 18 and we get the desired results Lemma 3. □

Appendix C Concentration in Theorem 2

Next, we will use several concentration inequalities and several lemmas from random matrix to bound our generalisation errors. In the following arguments we assume $\sqrt{n} \geq c_x^{'}\sqrt{k} + \sqrt{t}$ for some absolute constant $c_x^{'}$.

First we bound two items: $u_k(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\Sigma_{0:k}^{-\frac{1}{2}})$, $u_1(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\Sigma_{0:k}^{-\frac{1}{2}})$ in our generalisation errors bound, the following lemma shows that under our conditions,they are both $O(n)$.

Lemma 4. *Assume $X_{0:k}\Sigma_{0:k}^{-\frac{1}{2}}$ has n i.i.d. rows with isotropic sub-gaussian distribution in $\mathbb{R}^k$.Then with prob $1 - 2\exp(-c_x^{'}t)$, $u_k(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\Sigma_{0:k}^{-\frac{1}{2}}) \geq (\sqrt{n} - C_x^{'}\sqrt{k} - \sqrt{t})^2$, $u_1(\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{0:k}\Sigma_{0:k}^{-\frac{1}{2}}) \leq (\sqrt{n} + C_x^{'}\sqrt{k} + \sqrt{t})^2$, here $C_x^{'}, c_x^{'}$ only depends on σ_x*

The proof of this lemma can be found in Theorem 5.39 in Vershynin (2010)
Next we bound $\epsilon^T A_k^{-1} X_{0:k}\Sigma_{0:k}^{-1} X_{0:k}^T A_k^{-1}\epsilon$,which needs the following lemma:

Lemma 5. *Suppose $A \in \mathbb{R}^{n\times n}$ is a random psd matrix and $\epsilon \in \mathbb{R}^n$ is a centered vector whose components $\{\epsilon_i\}_{i=1}^n$ are independent and have sub-gaussian norm at most σ. Then for some absolute constant c, C and any $t > 1$ with prob at least $1 - 2e^{-\frac{t}{c}}$, $\epsilon^T A\epsilon \leq Ct\sigma^2 tr(A)$*

Proof. From Hanson-Wright inequality, for some absolute constant c_1 for any $t > 0$

$$\mathbb{P}_A\left\{\left|\varepsilon^\top A\varepsilon - \mathbb{E}\varepsilon^\top A\varepsilon\right| \geq t\right\} \leq 2\exp\left(-c_1\min\left\{\frac{t^2}{\|A\|_F^2\sigma^4}, \frac{t}{\sigma^2\|A\|}\right\}\right),$$

where $\mathbb{P}_A$ denotes conditional probability given A . Since for any i, $\mathbb{E}\varepsilon_i = 0$, and $\text{Var}(\varepsilon_i) \lesssim \sigma^2$, and A is psd we have

$$\mathbb{E}\varepsilon^\top A\varepsilon \leq c_2\sigma^2 \operatorname{tr}(A).$$

Moreover, since $\|A\|_F^2 \leq \operatorname{tr}(A)^2$ and $\|A\| \leq \operatorname{tr}(A)$, we obtain

$$\mathbb{P}_A\left\{\varepsilon^\top A\varepsilon > \sigma^2(c_2 + t)\operatorname{tr}(A)\right\} \leq 2\exp\left\{-c_1\min\left(t, t^2\right\}\right).$$

Restricting to $t > 1$ and adjusting the constants gives the result (note that since the RHS doesn't depend on A ,we can substitute $\mathbb{P}_A$ with $\mathbb{P}$) □

Lemma 5 is a weakened Hanson-Wright inequality. Using lemma 5,we get:

$$\begin{aligned}
&\epsilon^T A_k^{-1} X_{0:k} \Sigma_{0:k}^{-1} X_{0:k}^T A_k^{-1} \epsilon \\
\leq &c_1 \sigma_\epsilon^2 tr(X_{0:k} \Sigma_{0:k}^{-1} X_{0:k}^T A_k^{-1}) t \\
\leq &c_1 \sigma_\epsilon^2 t \mu_1 (A_k^{-1}) tr(X_{0:k} \Sigma_{0:k}^{-1} X_{0:K}^T)
\end{aligned} \tag{19}$$

Then we bound $tr(X_{0:k}\Sigma_{0:k}^{-1}X_{0:K}^T)$ and $tr(X_{k:\infty} diag\{\frac{1}{\lambda_{k+1}}, \ldots \frac{1}{\lambda_\infty}\}\Sigma_{k:\infty}\{\frac{1}{\lambda_{k+1}}, \ldots \frac{1}{\lambda_\infty}\}X_{k:\infty}^T)$,this need the following lemma:

Lemma 6. *(Concentration of the sum of squared norms) Suppose $z \in \mathbb{R}^{n\times p}$ is a matrix with independent isotropic sub-gaussian rows with $\|z_i\|_{\phi_2} \leq \sigma$. Consider $\Sigma = diag\{\lambda_1, \cdots, \lambda_p\}$ for some positive sequence $\{\lambda_i\}_{i=1}^p$. Then for some absolute constant c and any $t \in [0, n)$ with prob at least $1 - 2\exp(-ct)$, $(n - t\sigma^2)\sum_{i>k} \lambda_i \leq \sum_{i=1}^n \|\Sigma_{k:\infty}^{\frac{1}{2}} z_{ik:\infty}\|^2 \leq (n + t\sigma^2)\sum_{i>k} \lambda_i$*

Proof. Since $\{Z_{i,k:\infty}\}_{i=1}^n$ are independent, isotropic and sub-gaussian, $\left\|\Sigma_{k:\infty}^{1/2} Z_{i,k:\infty}\right\|^2$ are independent subexponential r.v's with expectation $\sum_{i>k} \lambda_i$ and sub-exponential norms bounded by $c_1\sigma^2 \sum_{i>k} \lambda_i$. Applying Bernstein's inequality gives

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \left\|\Sigma_{k:\infty}^{1/2} Z_{i,k:\infty}\right\|^2 - \sum_{i>k}\lambda_i\right| \geq t\sigma^2 \sum_{i>k}\lambda_i\right) \leq 2\exp\left(-c_2 \min\left(t, t^2\right) n\right)$$

Changing $t \to t/n$ gives the result. □

Applying lemma 6 to $X_{0:k}\Sigma_{0:k}^{-1}X_{0:K}^T$ and $X_{k:\infty} diag\{\frac{1}{\lambda_{k+1}}, \ldots \frac{1}{\lambda_\infty}\}\Sigma_{k:\infty}\{\frac{1}{\lambda_{k+1}}, \ldots \frac{1}{\lambda_\infty}\}X_{k:\infty}^T$,we get with probability at least $1 - 4e^{-c_2 t}$,

$$\begin{aligned}
&tr(X_{0:k}\Sigma_{0:k}^{-1}X_{0:K}^T) \leq (n + t\sigma_x^2)k \\
&tr(X_{k:\infty} diag\{\frac{1}{\lambda_{k+1}}, \ldots \frac{1}{\lambda_\infty}\}\Sigma_{k:\infty}\{\frac{1}{\lambda_{k+1}}, \ldots \frac{1}{\lambda_\infty}\}X_{k:\infty}^T) \leq (n + t\sigma_x^2)\Sigma_{i>k}\frac{u_i^2}{\lambda_i^2}
\end{aligned} \tag{20}$$

Next, we bound $\|X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^T\|$,this need the following lemma:

Lemma 7. *Suppose $\{z_i\}_{i=1}^n$ is a sequence of independent sub-gaussian vectors in $\mathbb{R}^p$ with $\|z_i\|_{\phi_2} \leq \sigma$. Consider $\Sigma = diag\{\lambda_1, \cdot, \lambda_p\}$ for some positive $\{\lambda_i\}_{i=1}^p$. Denote X to be the matrix with rows $\{z_i^T \Sigma^{\frac{1}{2}}\}_{i=1}^n$ and $A = XX^T$. Then with prob at least $1 - \sigma e^{-\frac{t}{c}}$, $\|A\| \leq c\sigma^2 (max_i \lambda_i (t+n) + \sum_i \lambda_i)$*

For the proof of this lemma, please see lemma 20 in Tsigler & Bartlett (2020). Using it,we have:

$$\|X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^T\| \leq c_3\sigma_x^2(\max_{i>k}\mu_i^2(t+n) + \Sigma_{i>k}\mu_i^2)$$

The vector $\frac{X_{k:\infty}\theta_{k:\infty}^*}{\|\theta_{k:\infty}^*\|_{\Sigma_{k:\infty}}}$ has n i.i.d centered components with unit variance and sub-gaussian norms at most σ_x. Then by lemma 6, we get

$$\|X_{k:\infty}\theta_{k:\infty}^*\|^2 \leq (n + t\sigma_x^2)\|\theta_{k:\infty}^*\|_{\Sigma_{k:\infty}}^2 \tag{21}$$

Using the results above, we get that with high probability,

$$
\begin{aligned}
B \lesssim_{\sigma_x} & \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + L^2\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} \frac{n^2}{(\mu_1(A_k)\min_{i\le k}\frac{\lambda_i}{\mu_i}+n)^2} \\
& + \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} (n^2\mu^2_{k+1} + n\sum_{i>k}\mu_i^2) \frac{L^2}{(1+\sum_{i>k}\frac{\mu_i}{\lambda_i})^2(\min_{i>k}\lambda_i)^2} \\
& + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{k:\infty}} \frac{L^2(\max_{i\le k}\lambda_i)^2(1+\sum_{i>k}\frac{\mu_i}{\lambda_i})^2}{(\mu_1(A_k)\min_{i\le k}\frac{\lambda_i}{\mu_i}+n)^2} + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{k:\infty}} L^2 \frac{(\max_{i\le k}\lambda_i)^2}{(\min_{i>k}\lambda_i)^2}(\mu^2_{k+1} + \frac{1}{n}\sum_{i>k}\mu_i^2)
\end{aligned}
\tag{22}
$$

$$
\frac{V}{\sigma_\epsilon^2} \lesssim_{\sigma_x} L^2 \frac{kn}{(\mu_1(A_k)\min_{i\le k}\frac{\lambda_i}{\mu_i}+n)^2} + L^2\frac{n}{(1+\sum_{i>k}\frac{\mu_i}{\lambda_i})^2}\sum_{i>k}\frac{\mu_i^2}{\lambda_i^2} \tag{23}
$$

Next we control $\mu_1(A_k)$ and $\mu_n(A_k)$,using the following lemma:

Lemma 8. *Suppose the condition number of the matrix A_k is at most L,then we have:*

$$
\frac{n-t\sigma^2}{nL}(1+\Sigma_{i>k}\frac{\mu_i}{\lambda_i}) \le \mu_n(A_k) \le \mu_1(A_k) \le \frac{n+t\sigma^2}{n}L(1+\Sigma_{i>k}\frac{\mu_i}{\lambda_i}) \tag{24}
$$

The proof of this lemma is analogous to the proof of Lemma 21 in Tsigler & Bartlett (2020) Plugging in,ignoring all constants and high probability characteristic $t, \sigma_x, \sigma_\epsilon$,we get total generalisation bound:

$$
\begin{aligned}
B \lesssim & \frac{n^2\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}}\frac{L^2}{(1+\Sigma_{i>k}\frac{\mu_i}{\lambda_i})^2} + \max_{i\le k}\lambda_i^2\|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{0:k}}}{(\min_{i\le k}\frac{\lambda_i}{\mu_i} + \frac{n}{L^2(1+\Sigma_{i>k}\frac{\mu_i}{\lambda_i})^2})^2} \\
& + \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} \\
& + \max_{i>k}\frac{1}{\lambda_i^2}(\max_{i>k}\mu_i^2 n + \Sigma_{i>k}\mu_i^2)L^2(1+\Sigma_{i>k}\frac{\mu_i}{\lambda_i})^2 n\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} \\
& + \frac{\max_{i>k}\frac{1}{\lambda_i^2}}{\min_{i<=k}\frac{1}{\lambda_i^2}} \frac{(\max_{i>k}\mu_i^2 n + \Sigma_{i>k}\mu_i^2)L^2\ \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{0:k}}}{n} \\
V \lesssim & \frac{L^2}{(1+\Sigma_{i>k}\frac{\mu_i}{\lambda_i})^2}\frac{nk}{(\min_{i<=k}\frac{\lambda_i}{\mu_i} + \frac{n}{L^2(1+\Sigma_{i>k}\frac{\mu_i}{\lambda_i})^2})^2} \\
& + L^2(1+\Sigma_{i>k}\frac{\mu_i}{\lambda_i})^2(n+t\sigma_x^2)\Sigma_{i>k}\frac{\mu_i^2}{\lambda_i^2}
\end{aligned}
\tag{25}
$$

Appendix D Technical improvement in Theorem 3

In the bounds above, 25 refers to $\|\theta^*_{0:k} - \hat\theta_{0:k}(X, X\theta^*)\|^2_{\Sigma_{0:k}}$,where $\hat\theta_{0:k}(X, X\theta^*)$ satisfies 13,25 is derived based on 14,through our technique improvement(namely 15),we can give a sharper bound of 25.We now bound the red part in 15 First we estimate $\|\Sigma_{0:k}^{-\frac{1}{2}}X^T_{0:k}X_{k:\infty}\theta^*_{k:\infty}\|$, namely

$$
\|\Sigma_i \begin{pmatrix} \frac{x_{i1}}{\sqrt{\mu_1}} \\ \dots \\ \frac{x_{ik}}{\sqrt{\mu_k}} \end{pmatrix} \Sigma_{j>k}x_{ij}\theta^*_j\|.
$$

To use Matrix Bernstein's inequality, we need to further assume

$$
\| \begin{pmatrix} \frac{x_{i1}}{\sqrt{\mu_1}} \\ \dots \\ \frac{x_{ik}}{\sqrt{\mu_k}} \end{pmatrix} \Sigma_{j>k}x_{ij}\theta^*_j\| \lesssim \sqrt{k}\sqrt{\Sigma_{j>k}\mu_j{\theta^*_j}^2},
$$

which is a mild condition. Then we can use the following lemma to bound it:

Lemma 9. *(Matrix Bernstein's inequality for rectangular matrices)Let $X_1, ...X_N$ be independent,mean zero,$m \times n$ random matrices,such that $\|X_i\| \leq K$ almost surely for all i. Then we have with probability at least $1 - 2(m+n)e^{\frac{-0.5t^2}{\sigma^2 + \frac{Kt}{3}}}$, $\|\Sigma_{i=1}^N X_i\| \leq t$,where $\sigma^2 = \max(\|\Sigma_{i=1}^N EX_i^T X_i\|, \|\Sigma_{i=1}^N EX_i X_i^T\|)$*

Proof. Simply apply matrix Bernstein's inequality for the sum of $(m+n) \times (m+n)$ symmetric matrices $\begin{bmatrix} 0 & X_i^T \\ X_i & 0 \end{bmatrix}$, and we will get the result. □

Then, by using this lemma, we have:

$$\|\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T X_{k:\infty} \theta_{k:\infty}^*\| \lesssim \sqrt{n}\sqrt{k}\sqrt{\Sigma_{i>k}\mu_i {\theta_i^*}^2}\sqrt{ln(k)} \tag{26}$$

holds with high probability. Next we turn to another part, namely $\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T (I_n - A_k^{-1}) X_{k:\infty}\theta_{k:\infty}$ in 15, through concentration inequalities analogous to the lemmas mentioned above, we can show:

$$\mu_1(A_k) \lesssim (\sqrt{n} + \sqrt{p-k})^2 \max_{i>k} \frac{\mu_i}{\lambda_i} + 1 \tag{27}$$

Hence

$$\mu_1(I - A_k^{-1}) \leq \frac{(\sqrt{n} + \sqrt{p-k})^2 max_{i>k} \frac{\mu_i}{\lambda_i}}{1 + (\sqrt{n} + \sqrt{p-k})^2 max_{i>k} \frac{\mu_i}{\lambda_i}} \tag{28}$$

Using it together with the bounds for $\|\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T\|$ and $\|X_{k:\infty}\theta_{k:\infty}^*\|$, we get

$$\|\Sigma_{0:k}^{-\frac{1}{2}} X_{0:k}^T (I_n - A_k^{-1}) X_{k:\infty}\theta_{k:\infty}\| \lesssim \frac{(\sqrt{n} + \sqrt{p-k})^2 \max_{i>k} \frac{\mu_i}{\lambda_i}}{1 + (\sqrt{n} + \sqrt{p-k})^2 \max_{i>k} \frac{\mu_i}{\lambda_i}} n \|\theta_{k:\infty}^*\|_{\Sigma_{k:\infty}} \tag{29}$$

Plugging 26 and 29 into 15, we get 25 can be improved as

$$\frac{\mu_1(A_k)^2 (n+p)^2 (\max_{i>k} \frac{\mu_i}{\lambda_i})^2}{(1 + (n+p)\max_{i>k} \frac{\mu_i}{\lambda_i})^2} \|\theta_{k:\infty}^*\|_{\Sigma_{k:\infty}}^2 + \frac{\mu_1(A_k)^2 k \Sigma_{i>k}\mu_i {\theta_i^*}^2 ln(k)}{n} + \frac{\Sigma_{i \leq k} \frac{\lambda_i^2}{\mu_i} {\theta_i^*}^2}{n^2} \tag{30}$$