Benign Overfitting in Linear Regression

Peter L. Bartlett, Philip M. Long , G´abor Lugosi , and Alexander Tsigler

Linear Regression Setting

$x \in \mathcal{H}$ (Hilbert space) and response $y \in \mathbb{R}$

Assumptions: (x,y) mean-zero, well-specified $E[y|x] = x^T\theta^*$

$x = V\Lambda^{1/2}z$, where $\Sigma = V\Lambda V^\top$ is the spectral decomposition of Σ and z has components that are independent σ_x^2-subgaussian

Define:

$$\Sigma := E[xx^T] = \sum_i \lambda_i v_i v_i^T$$

$$\theta^* := \underset{\theta}{\operatorname{argmin}} E(y - x^T\theta)^2$$

$$\sigma^2 := E(y - x^T\theta^*)^2$$

Minimum norm estimator

We consider overparameterized regime

Data $X \in \mathcal{H}^n$, $y \in \mathbb{R}^n$

Estimator $\hat{\theta}$

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta} \left\{ \|\theta\|^2 : X^\top X\theta = X^\top \boldsymbol{y} \right\} \\ &= \left(X^\top X\right)^\dagger X^\top \boldsymbol{y} \\ &= X^\top \left(XX^\top\right)^{-1} \boldsymbol{y}.\end{aligned}$$

solves

$$\begin{aligned}&\min_{\theta \in \mathbb{H}} && \|\theta\|^2 \\ &\textit{such that} && \|X\theta - \boldsymbol{y}\|^2 = \min_{\beta} \|X\beta - \boldsymbol{y}\|^2 .\end{aligned}$$

Excess Prediction Error

$$
\begin{aligned}
R(\theta) &:= \mathbb{E}_{x,y}\left[\left(y - x^\top\theta\right)^2 - \left(y - x^\top\theta^*\right)^2\right] \\
&= \mathbb{E}_{x,y}\left(y - x^\top\theta^* + x^\top\left(\theta^* - \hat{\theta}\right)\right)^2 - \mathbb{E}\left(y - x^\top\theta^*\right)^2 \\
&= \mathbb{E}_x\left(x^\top\left(\theta^* - \hat{\theta}\right)\right)^2.
\end{aligned}
$$

$$= \left(\hat{\theta} - \theta^*\right)^T \Sigma \left(\hat{\theta} - \theta^*\right)$$

$$\Sigma := E[xx^T] = \sum_i \lambda_i v_i v_i^T$$

$\lambda_1 \geq \lambda_2 \geq \ \ldots \geq \ \lambda_d$ denote the eigenvalues of Σ in descending order

Prediction error has two components:

- $\hat{\theta}$ is the distorted version of θ^*,
  because we have acess to the samples $x_1, \dots x_n$ and not to the covariance of x

$$\left\|\Sigma - \hat{\Sigma}\right\|, \text{where } \hat{\Sigma} := \frac{1}{n} X^T X$$

- $\hat{\theta}$ is corrupted by the noise in $y_1, \dots y_n$

Estimator: $\hat{\theta} = (X^T X)^T X^T y = (X^T X)^T X^T (X\theta^* + \varepsilon)$

Excess Risk: $R(\hat{\theta}) = \left(\hat{\theta} - \theta^*\right)^T \Sigma \left(\hat{\theta} - \theta^*\right)$

$$\approx (\theta^*)^T \left(I - \hat{\Sigma}\hat{\Sigma}^+\right) \left(\Sigma - \hat{\Sigma}\right) \left(I - \hat{\Sigma}^+\hat{\Sigma}\right) \theta^* + \sigma^2 \operatorname{tr}\left((X^T X)^+ \Sigma\right)$$

Theorem 4. *For any σ_x there are $b, c, c_1 > 1$ for which the following holds. Consider a linear regression problem from Definition 1. Define*

$$k^* = \min\left\{k \geq 0 : r_k(\Sigma) \geq bn\right\},$$

where the minimum of the empty set is defined as ∞. Suppose $\delta < 1$ with $\log(1/\delta) < n/c$. If $k^* \geq n/c_1$, then $\mathbb{E}R(\hat{\theta}) \geq \sigma^2/c$. Otherwise,

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

with probability at least $1-\delta$, and

$$\mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right).$$

Moreover, there are universal constants a_1, a_2, n_0 such that for all $n \geq n_0$, for all Σ, for all $t \geq 0$, there is a θ^* with $\|\theta^*\| = t$ such that for $x \sim \mathcal{N}(0, \Sigma)$ and $y|x \sim \mathcal{N}(x^\top\theta^*, \|\theta^*\|^2\|\Sigma\|)$, with probability at least $1/4$,

$$R(\hat{\theta}) \geq \frac{1}{a_1}\|\theta^*\|^2 \|\Sigma\| \, \mathbb{1}\left[\frac{r_0(\Sigma)}{n\log\left(1 + r_0(\Sigma)\right)} \geq a_2\right].$$

Definition: Effective Rank

$\lambda_1 \geq \lambda_2 \geq \ \dots \geq \ \lambda_d$ denote the eigenvalues of Σ in descending order

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

<u>Examples</u>

1. $\Sigma = I_{dxd}$:

$r_0(I_{dxd}) = R_0(I_{dxd}) =$d

2. $\lambda_1 \geq \lambda_2 = 0 \geq \ \dots \geq \ \lambda_d = 0$:

$r_0(\Sigma) = R_0(\Sigma) = 1$

3. rank(Σ) = d :

$r_0(\Sigma)$=rank(Σ)s(Σ)

$R_0(\Sigma)$=rank(Σ)s(Σ)

$$s(\Sigma) = \frac{\frac{1}{p}\sum_{i=1}^{p}\lambda_i}{\lambda_{k+1}}$$

$$S(\Sigma) = \frac{(\frac{1}{p}\sum_{i=1}^{p}\lambda_i)^2}{\frac{1}{p}\sum_{i=1}^{p}\lambda_i{}^2}$$

Both s and S lie between 1/d ($\lambda_2 \approx 0$) and 1 (λ_i all equal)

$$k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$$

$$r_k(\Sigma) = \frac{\sum_{i>k}\lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k}\lambda_i\right)^2}{\sum_{i>k}\lambda_i^2}.$$

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

Intuition

- The eigenvalues of Σ determines how errors in $\hat{\theta}$ affect prediction accuracy
- To avoid harming prediction accuracy, the noise energy must be distributed across many unimportant directions
- Overparameterization is essential for benign overfitting

➔ Number of small eigenvalues must be large compared to n

➔ Small eigenvalues must be roughly equal

Proof: Upper Bound

Excess Prediction Error: $R(\hat{\theta}) = E\left[x^T(\theta^* - \hat{\theta})\right]^2$

$$\begin{aligned}\hat{\theta} &= X^T(XX^T)^{-1}y \\ &= X^T(XX^T)^{-1}(X\theta^* + \varepsilon)\end{aligned}$$

Using (1), the definition of Σ, and the fact that $\boldsymbol{y} = X\theta^* + \boldsymbol{\varepsilon}$,

$$\begin{aligned}R(\hat{\theta}) &= \mathbb{E}_x\left(x^\top\left(I - X^\top\left(XX^\top\right)^{-1}X\right)\theta^* - x^\top X^\top\left(XX^\top\right)^{-1}\boldsymbol{\varepsilon}\right)^2 \\ &\leq 2\mathbb{E}_x\left(x^\top\left(I - X^\top\left(XX^\top\right)^{-1}X\right)\theta^*\right)^2 + 2\mathbb{E}_x\left(x^\top X^\top\left(XX^\top\right)^{-1}\boldsymbol{\varepsilon}\right)^2 \\ &= 2\theta^{*\top}\left(I - X^\top\left(XX^\top\right)^{-1}X\right)\Sigma\left(I - X^\top\left(XX^\top\right)^{-1}X\right)\theta^* \\ &\qquad + 2\boldsymbol{\varepsilon}^\top\left(XX^\top\right)^{-1}X\Sigma X^\top\left(XX^\top\right)^{-1}\boldsymbol{\varepsilon} \\ &= 2\theta^{*\top}B\theta^* + 2\boldsymbol{\varepsilon}^\top C\boldsymbol{\varepsilon}.\end{aligned}$$

We showed that $R(\hat{\theta}) = \mathbb{E}_x \left(x^\top \left(\theta^* - \hat{\theta}\right)\right)^2 \leq 2{\theta^*}^\top B \theta^* + 2\varepsilon^\top C \varepsilon$

,where

$$B = \left(I - X^\top \left(XX^\top\right)^{-1} X\right) \Sigma \left(I - X^\top \left(XX^\top\right)^{-1} X\right),$$

$$C = \left(XX^\top\right)^{-1} X \Sigma X^\top \left(XX^\top\right)^{-1}.$$

Bias Term

Lemma 35. *There is a constant c, that depends only on σ_x, such that for any $1 < t < n$, with probability at least $1 - e^{-t}$,*

$$\theta^{*\top} B\theta^* \leq c\|\theta^*\|^2\|\Sigma\| \max\left\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{t}{n}}\right\}.$$

Proof:

$$\theta^{*\top} B\theta^* = \theta^{*\top}\left(I - X^\top\left(XX^\top\right)^{-1}X\right)\Sigma\left(I - X^\top\left(XX^\top\right)^{-1}X\right)\theta^*$$

$$= \theta^{*\top}\left(I - X^\top\left(XX^\top\right)^{-1}X\right)\left(\Sigma - \frac{1}{n}X^\top X\right)\left(I - X^\top\left(XX^\top\right)^{-1}X\right)\theta^*.$$

$$\leq \left\|\Sigma - \frac{1}{n}X^\top X\right\| \|\theta^*\|^2$$

Variance Term Roadmap

$$R(\hat{\theta}) = \mathbb{E}_x \left(x^\top \left(\theta^* - \hat{\theta} \right) \right)^2 \leq 2{\theta^*}^\top B \theta^* + 2\varepsilon^\top C \varepsilon$$

Lemma 19

$$\leq 2{\theta^*}^\top B \theta^* + c\sigma^2 \log \frac{1}{\delta} \operatorname{tr}(C) \qquad \textit{with probability at least } 1 - \delta \textit{ over } \epsilon,$$

Lemma 8 — Lemma 11 — Lemma 17

$$\operatorname{tr}(C) = \sum_i \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \leq c \left(\frac{l}{n} + n \frac{\sum_{i>l} \lambda_i^2}{\left(\sum_{i>k} \lambda_i\right)^2} \right) \leq \frac{k^*}{bn} + \frac{bn}{R_{k^*}(\Sigma)}$$

Lemma 8. *Consider a covariance operator* Σ *with* $\lambda_i = \mu_i(\Sigma)$ *and* $\lambda_n > 0$. *Write its spectral decomposition* $\Sigma = \sum_j \lambda_j v_j v_j^\top$, *where the orthonormal* $v_j \in \mathbb{H}$ *are the eigenvectors corresponding to the* λ_j. *For* i *with* $\lambda_i > 0$, *define* $z_i = Xv_i/\sqrt{\lambda_i}$. *Then*

$$\operatorname{tr}(C) = \sum_i \left[\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top \right)^{-2} z_i \right],$$

and these $z_i \in \mathbb{R}^n$ *are independent* σ_x^2*-subgaussian. Furthermore, for any* i *with* $\lambda_i > 0$, *we have*

$$\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top \right)^{-2} z_i = \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2},$$

where $A_{-i} = \sum_{j \neq i} \lambda_j z_j z_j^\top$.

Proof:

By Assumption 2 in Definition 1, the random variables $x^\top v_i/\sqrt{\lambda_i}$ are independent σ_x^2-subgaussian. We consider X in the basis of eigenvectors of Σ, $Xv_i = \sqrt{\lambda_i} z_i$, to see that

$$XX^\top = \sum_i \lambda_i z_i z_i^\top, \qquad X\Sigma X^\top = \sum_i \lambda_i^2 z_i z_i^\top,$$

$$\operatorname{tr}(C) = \operatorname{tr}\left(\left(XX^\top\right)^{-1} X\Sigma X^\top \left(XX^\top\right)^{-1}\right)$$

$$= \sum_i \left[\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top\right)^{-2} z_i\right].$$

$$Z^\top(ZZ^\top + A)^{-2}Z = (I + Z^\top A^{-1}Z)^{-1}Z^\top A^{-2}Z(I + Z^\top A^{-1}Z)^{-1}.$$

$$\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top\right)^{-2} z_i = \lambda_i^2 z_i^\top \left(\lambda_i z_i z_i^\top + A_{-i}\right)^{-2} z_i$$

$$= \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2},$$

Lemma 11. *There are constants $b, c \geq 1$ such that if $0 \leq k \leq n/c$, $r_k(\Sigma) \geq bn$, and $l \leq k$ then with probability at least $1 - 7e^{-n/c}$,*

$$\operatorname{tr}(C) \leq c\left(\frac{l}{n} + n\frac{\sum_{i>l}\lambda_i^2}{\left(\sum_{i>k}\lambda_i\right)^2}\right).$$

Proof:

Lemma 8

$$\operatorname{tr}(C) = \sum_i \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1+\lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \leq \sum_{i=1}^{l} \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1+\lambda_i z_i^\top A_{-i}^{-1} z_i)^2} + \sum_{i>l} \lambda_i^2 z_i^\top A^{-2} z_i.$$

Lemma 10 Corollary 13

$$\frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1+\lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \leq \frac{z_i^\top A_{-i}^{-2} z_i}{(z_i^\top A_{-i}^{-1} z_i)^2} \leq c_1^4 \frac{\|z_i\|^2}{\|\Pi_{\mathscr{L}_i} z_i\|^4} \leq c\frac{1}{n} \longrightarrow \sum_{i=1}^{l} \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1+\lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \leq c_4 \frac{l}{n}$$

$\mathscr{L}_i$ is the span of the $n-k$ eigenvectors of A_{-i} corresponding to its smallest $n-k$ eigenvalues

Lemma 10 shows that with probability at least $1 - 2e^{-n/c_1}$, for all $i \leq k$

$$\mu_n(A_{-i}) \geq \lambda_{k+1} r_k(\Sigma)/c_1$$

($\mu_1(A)$ and $\mu_n(A)$ denote the largest and the smallest eigenvalues of the $n \times n$ matrix A.)

lower bounds on the $\mu_n(A_{-i})$'s imply that, for all $z \in \mathbb{R}^n$ and $1 \leq i \leq l$,

$$z^\top A_{-i}^{-2} z \leq \frac{c_1^2 \|z\|^2}{(\lambda_{k+1} r_k(\Sigma))^2},$$

Lemma 10 also shows that for all i, $\mu_{k+1}(A_{-i}) \leq c_1 \lambda_{k+1} r_k(\Sigma)$

$$z^\top A_{-i}^{-1} z \geq (\Pi_{\mathscr{L}_i} z)^\top A_{-i}^{-1} \Pi_{\mathscr{L}_i} z \geq \frac{\|\Pi_{\mathscr{L}_i} z\|^2}{c_1 \lambda_{k+1} r_k(\Sigma)},$$

where $\mathscr{L}_i$ is the span of the $n-k$ eigenvectors of A_{-i} corresponding to its smallest $n-k$ eigenvalues. So for $i \leq l$,

$$\frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \leq \frac{z_i^\top A_{-i}^{-2} z_i}{(z_i^\top A_{-i}^{-1} z_i)^2} \leq c_1^4 \frac{\|z_i\|^2}{\|\Pi_{\mathscr{L}_i} z_i\|^4}. \tag{3}$$

Lemma 11. *There are constants $b, c \geq 1$ such that if $0 \leq k \leq n/c$, $r_k(\Sigma) \geq bn$, and $l \leq k$ then with probability at least $1 - 7e^{-n/c}$,*

$$\operatorname{tr}(C) \leq c \left(\frac{l}{n} + n \frac{\sum_{i>l} \lambda_i^2}{\left(\sum_{i>k} \lambda_i\right)^2} \right).$$

Lemma 10

$$\sum_{i>l} \lambda_i^2 z_i^\top A^{-2} z_i \leq \frac{c_1^2 \sum_{i>l} \lambda_i^2 \|z_i\|^2}{\left(\lambda_{k+1} r_k(\Sigma)\right)^2}$$

Lemma 12

$$\begin{aligned}\sum_{i>l} \lambda_i^2 \|z_i\|^2 &\leq n \sum_{i>l} \lambda_i^2 + a\sigma_x^2 \max\left(\lambda_{l+1}^2 t, \sqrt{tn \sum_{i>l} \lambda_i^4} \right) \\ &\leq n \sum_{i>l} \lambda_i^2 + a\sigma_x^2 \max\left(t \sum_{i>l} \lambda_i^2, \sqrt{tn} \sum_{i>l} \lambda_i^2 \right) \\ &\leq c_5 n \sum_{i>l} \lambda_i^2,\end{aligned}$$

$$\sum_{i>l} \lambda_i^2 z_i^\top A^{-2} z_i \leq c_6 n \frac{\sum_{i>l} \lambda_i^2}{\left(\lambda_{k+1} r_k(\Sigma)\right)^2}$$

Lemma 17. *For any $b \geq 1$ and $k^* := \min\{k : r_k(\Sigma) \geq bn\}$, if $k^* < \infty$, we have*

$$\min_{l \leq k^*} \left(\frac{l}{bn} + \frac{bn \sum_{i>l} \lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} \right) = \frac{k^*}{bn} + \frac{bn \sum_{i>k^*} \lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} = \frac{k^*}{bn} + \frac{bn}{R_{k^*}(\Sigma)}$$

Proof:

We can write the function of l being minimized as

$$\begin{aligned}
\frac{l}{bn} + \frac{bn \sum_{i>l} \lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} &= \sum_{i=1}^{l} \frac{1}{bn} + \sum_{i>l} \frac{bn\lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} \\
&\geq \sum_{i=1}^{k^*} \min\left\{ \frac{1}{bn}, \frac{bn\lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} \right\} \\
&\quad + \sum_{i>k^*} \frac{bn\lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} \\
&= \sum_{i=1}^{l^*} \frac{1}{bn} + \sum_{i>l^*} \frac{bn\lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2},
\end{aligned}$$

where l^* is the largest value of $i \leq k^*$ for which

$$\frac{1}{bn} \leq \frac{bn\lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2},$$

where l^* is the largest value of $i \leq k^*$ for which

$$\frac{1}{bn} \leq \frac{bn\lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2},$$

$$\lambda_1 \geq \lambda_2 \geq \ \dots \geq \ \lambda_d$$

since the λ_i^2 are non-increasing. This condition holds iff

$$\lambda_i \geq \frac{\lambda_{k^*+1} r_{k^*}(\Sigma)}{bn}.$$

The definition of k^* implies $r_{k^*-1}(\Sigma) < bn$. So we can write

$$k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$$

$$\begin{aligned} r_{k^*}(\Sigma) &= \frac{\sum_{i>k^*} \lambda_i}{\lambda_{k^*+1}} \\ &= \frac{\sum_{i>k^*-1} \lambda_i - \lambda_{k^*}}{\lambda_{k^*+1}} \\ &= \frac{\lambda_{k^*}}{\lambda_{k^*+1}} (r_{k^*-1}(\Sigma) - 1) \\ &< \frac{\lambda_{k^*}}{\lambda_{k^*+1}} (bn - 1), \end{aligned}$$

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}$$

and so the minimizing l is k^*. Also,

$$\frac{\sum_{i>k^*} \lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} = \frac{\sum_{i>k^*} \lambda_i^2}{\left(\sum_{i>k^*} \lambda_i\right)^2} = \frac{1}{R_{k^*}(\Sigma)}.$$

Proof: Lower Bound

Excess Prediction Error: $R(\hat{\theta}) = E\left[x^T(\theta^* - \hat{\theta})\right]^2$

Also, since ε has zero mean conditionally on X, and is independent of x, we have

$$\begin{aligned}
\mathbb{E}_{x,\varepsilon} R(\hat{\theta}) &= \mathbb{E}_{x,\varepsilon}\left[\left(x^\top\left(I - X^\top\left(XX^\top\right)^{-1}X\right)\theta^*\right)^2 + \left(x^\top X^\top\left(XX^\top\right)^{-1}\varepsilon\right)^2\right] \\
&= \theta^{*\top}\left(I - X^\top\left(XX^\top\right)^{-1}X\right)\Sigma\left(I - X^\top\left(XX^\top\right)^{-1}X\right)\theta^* \\
&\quad + \operatorname{tr}\left(\left(XX^\top\right)^{-1}X\Sigma X^\top\left(XX^\top\right)^{-1}\mathbb{E}\left[\varepsilon\varepsilon^\top \middle| X\right]\right) \\
&\geq \theta^{*\top}B\theta^* + \sigma^2\operatorname{tr}(C).
\end{aligned}$$

$$\mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)$$

Variance Term Roadmap

Lemma 8

Lemma 14

$$\operatorname{tr}(C) = \sum_i \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \geq \sum_i \frac{1}{cn}\left(1 + \frac{\sum_{j>k} \lambda_j + n\lambda_{k+1}}{n\lambda_i}\right)^{-2}$$

Lemma 16.2

$$\geq \frac{1}{cb^2} \min_{l \leq k} \left(\frac{l}{n} + \frac{b^2 n \sum_{i>l} \lambda_i^2}{(\lambda_{k+1} r_k(\Sigma))^2} \right)$$

Lemma 17

$$\geq \frac{k^*}{bn} + \frac{bn}{R_{k^*}(\Sigma)}$$

Lemma 14. *There is a constant c such that for any $i \geq 1$ with $\lambda_i > 0$, and any $0 \leq k \leq n/c$, with probability at least $1 - 5e^{-n/c}$,*

$$\frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \geq \frac{1}{cn} \left(1 + \frac{\sum_{j>k} \lambda_j + n\lambda_{k+1}}{n\lambda_i}\right)^{-2}.$$

$$A_{-i} = \sum_{j \neq i} \lambda_j z_j z_j^\top$$

Proof:

Fix $i \geq 1$ with $\lambda_i > 0$ and $0 \leq k \leq n/c$. By Lemma 10, with probability at least $1 - 2e^{-n/c_1}$,

$$\mu_{k+1}(A_{-i}) \leq c_1 \left(\sum_{j>k} \lambda_j + \lambda_{k+1} n \right),$$

and hence

$$z_i^\top A_{-i}^{-1} z_i \geq \frac{\|\Pi_{\mathscr{L}_i} z_i\|^2}{c_1 \left(\sum_{j>k} \lambda_j + \lambda_{k+1} n\right)}.$$

$\mathscr{L}_i$ is the span of the $n-k$ eigenvectors of A_{-i} corresponding to its smallest $n-k$ eigenvalues

By Corollary 13, with probability at least $1-3e^{-t}$,

$$\|\Pi_{\mathscr{L}_i} z_i\|^2 \geq n - a\sigma_x^2(k+t+\sqrt{tn}) \geq n/c_2,$$

provided that $t < n/c_0$ and $c > c_0$ for some sufficiently large c_0. Thus, with probability at least $1-5e^{-n/c_3}$,

$$z_i^\top A_{-i}^{-1} z_i \geq \frac{n}{c_3\left(\sum_{j>k}\lambda_j + \lambda_{k+1}n\right)},$$

hence

$$1+\lambda_i z_i^\top A_{-i}^{-1} z_i \leq \left(\frac{c_3\left(\sum_{j>k}\lambda_j + \lambda_{k+1}n\right)}{\lambda_i n} + 1\right)\lambda_i z_i^\top A_{-i}^{-1} z_i.$$

Dividing $\lambda_i^2 z_i^\top A_{-i}^{-2} z_i$ by the square of both sides, we have

$$\frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1+\lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \geq \left(\frac{c_3\left(\sum_{j>k}\lambda_j + \lambda_{k+1}n\right)}{\lambda_i n} + 1\right)^{-2} \frac{z_i^\top A_{-i}^{-2} z_i}{(z_i^\top A_{-i}^{-1} z_i)^2}.$$

Also, from the Cauchy-Schwarz inequality and Corollary 13 again, we have that on the same event,

$$\begin{aligned}\frac{z_i^\top A_{-i}^{-2} z_i}{(z_i^\top A_{-i}^{-1} z_i)^2} &\geq \frac{z_i^\top A_{-i}^{-2} z_i}{\left\|A_{-i}^{-1} z_i\right\|^2 \|z_i\|^2} \\ &= \frac{1}{\|z_i\|^2} \geq \frac{1}{n + a\sigma_x^2(t+\sqrt{nt})} \geq \frac{1}{c_4 n}.\end{aligned}$$

Lemma 16. *There are constants c such that for any $0 \leq k \leq n/c$ and any $b > 1$ with probability at least $1 - 10e^{-n/c}$,*

1. *If $r_k(\Sigma) < bn$, then $\operatorname{tr}(C) \geq \frac{k+1}{cb^2 n}$.*
2. *If $r_k(\Sigma) \geq bn$, then*

$$\operatorname{tr}(C) \geq \frac{1}{cb^2} \min_{l \leq k} \left(\frac{l}{n} + \frac{b^2 n \sum_{i>l} \lambda_i^2}{(\lambda_{k+1} r_k(\Sigma))^2} \right).$$

Proof:

$$\begin{aligned}
\operatorname{tr}(C) &\geq \frac{1}{c_1 n} \sum_i \left(1 + \frac{\sum_{j>k} \lambda_j + n\lambda_{k+1}}{n\lambda_i} \right)^{-2} \\
&\geq \frac{1}{c_2 n} \sum_i \min \left\{ 1, \frac{n^2 \lambda_i^2}{\left(\sum_{j>k} \lambda_j \right)^2}, \frac{\lambda_i^2}{\lambda_{k+1}^2} \right\} \\
&\geq \frac{1}{c_2 b^2 n} \sum_i \min \left\{ 1, \left(\frac{bn}{r_k(\Sigma)} \right)^2 \frac{\lambda_i^2}{\lambda_{k+1}^2}, \frac{\lambda_i^2}{\lambda_{k+1}^2} \right\}.
\end{aligned}$$

$$\operatorname{tr}(C) \geq \frac{1}{c_2 b^2 n} \sum_i \min\left\{1, \left(\frac{bn}{r_k(\Sigma)}\right)^2 \frac{\lambda_i^2}{\lambda_{k+1}^2}, \frac{\lambda_i^2}{\lambda_{k+1}^2}\right\}$$

if $r_k(\lambda) \geq bn$,

$$\begin{aligned}
\operatorname{tr}(C) &\geq \frac{1}{c_2 b^2} \sum_i \min\left\{\frac{1}{n}, \frac{b^2 n \lambda_i^2}{(\lambda_{k+1} r_k(\Sigma))^2}\right\} \\
&= \frac{1}{c_2 b^2} \min_{l \leq k} \left(\frac{l}{n} + \frac{b^2 n \sum_{i>l} \lambda_i^2}{(\lambda_{k+1} r_k(\Sigma))^2}\right),
\end{aligned}$$

where the equality follows from the fact that the λ_is are non-increasing. $\square$

Theorem 4. *For any σ_x there are $b, c, c_1 > 1$ for which the following holds. Consider a linear regression problem from Definition 1. Define*

$$k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\},$$

where the minimum of the empty set is defined as ∞. Suppose $\delta < 1$ with $\log(1/\delta) < n/c$. If $k^* \geq n/c_1$, then $\mathbb{E}R(\hat{\theta}) \geq \sigma^2/c$. Otherwise,

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

with probability at least $1-\delta$, and

$$\mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right).$$

Moreover, there are universal constants a_1, a_2, n_0 such that for all $n \geq n_0$, for all Σ, for all $t \geq 0$, there is a θ^* with $\|\theta^*\| = t$ such that for $x \sim \mathcal{N}(0, \Sigma)$ and $y|x \sim \mathcal{N}(x^\top\theta^*, \|\theta^*\|^2\|\Sigma\|)$, with probability at least $1/4$,

$$R(\hat{\theta}) \geq \frac{1}{a_1}\|\theta^*\|^2 \|\Sigma\| \mathbb{1}\left[\frac{r_0(\Sigma)}{n\log(1+r_0(\Sigma))} \geq a_2\right].$$