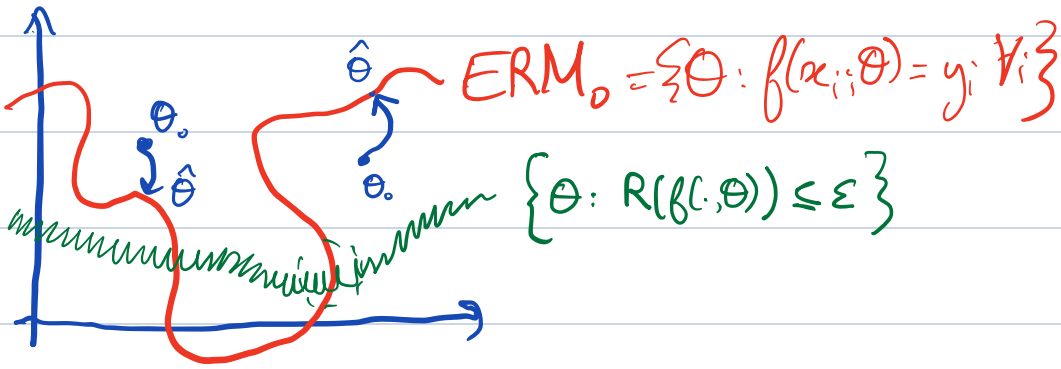
Lecture 9: Benign Overfitting

Last week:



$\hat{\theta}$: depends on algo, initialization, parametrization etc (step size etc.)

What about generalization of $f(\cdot;\hat{\theta})$?

→ model trained with no control on the model complexity (until interpolation of the data)

≠ standard wisdom in statistical learning

→ models that are too complex will not generalize well

→ overfitting is bad and should be avoided

Today: "benign overfitting" → interpolating solution can generalize well

* Classical picture
* min-$\|\cdot\|_2$ interpolating solution in linear models (solution attained by GD)
  ↳ generalizat° via "self-induced regularization"
* Example: inner-product kernel

→ RMT & exact asymptotics
→ matrix concentration

} No time for these → will go back to it in lecture 6

Classical Picture

Parametric family: $F = \{ f(\cdot;\theta) : \theta \in \mathbb{R}^p \}$

Data: $\{(y_i, x_i)\}_{i \le n}$ $\quad x_i \in \mathbb{R}^d \overset{iid}{\sim} \mathbb{P}$ $\quad y_i \in \mathbb{R}$

$y_i = f(x_i; \theta_*) + \varepsilon_i$) noise e.g. $\varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$

SRM: $\hat{\theta}(\lambda) = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i;\theta))^2 + \lambda \|\theta\|_2^2 \right\}$

Excess Test error: $R(\theta_*, \lambda) = \mathbb{E}_x \left[\left(f(x, \theta_*) - f(x, \hat{\theta}(\lambda)) \right)^2 \right]$

Bias - Variance decomposition:

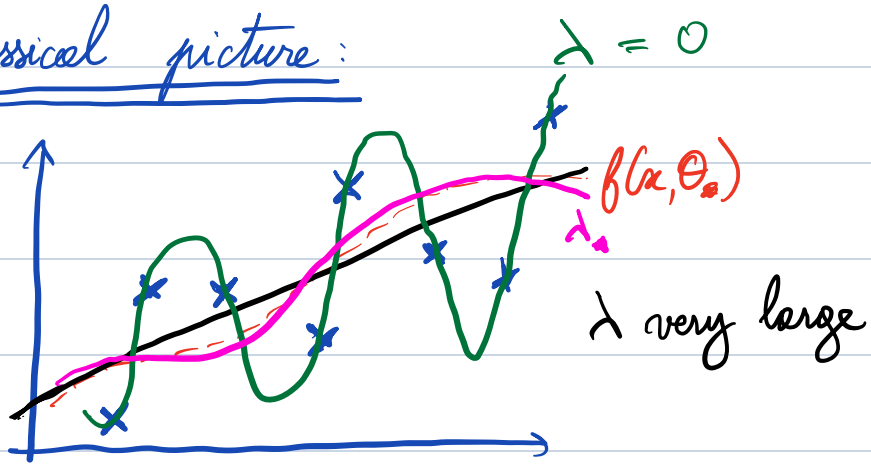
$$\mathbb{E}_\varepsilon [R(\theta_*, \lambda)] = \mathbb{E}_{x,\varepsilon} \left[\left(f(x;\theta_*) - f(x;\hat{\theta}(\lambda)) \right)^2 \right]$$

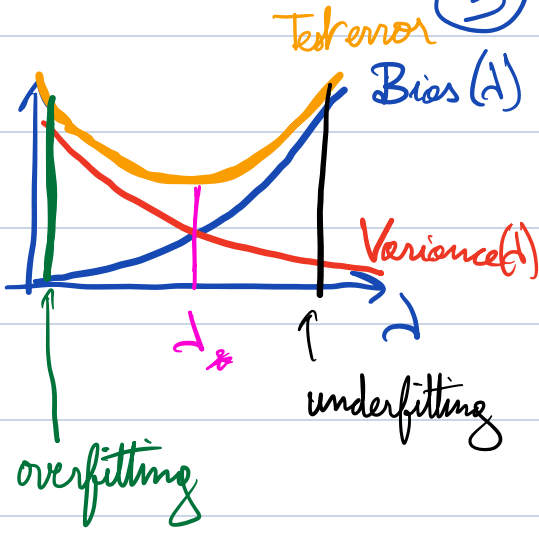
$$= \mathbb{E}_x \left[\left(f(x;\theta_*) - \mathbb{E}_\varepsilon [f(x;\hat{\theta}(\lambda))] \right)^2 \right] + \mathbb{E}_{x,\varepsilon} \left[\left(f(x, \hat{\theta}(\lambda)) - \mathbb{E}_\varepsilon [f(x;\hat{\theta}(\lambda))] \right)^2 \right]$$

$=:$ BIAS (λ) $\quad +$ $\quad$ VARIANCE (λ)

$= \mathbb{E}_x \left[\text{bias}(\mathbb{E}_\varepsilon [f(x,\hat{\theta})])^2 \right]$ $\quad$ $= \mathbb{E}_x \left[\text{Var}_\varepsilon (f(x;\hat{\theta})) \right]$

Classical picture :





Trade-off between Bias and Variance

→ need to carefully choose regularization parameter λ in order to control the complexity of the fitted model $f(x; \theta(\lambda))$

⇒ too complex : small bias, large variance

OVERFITTING

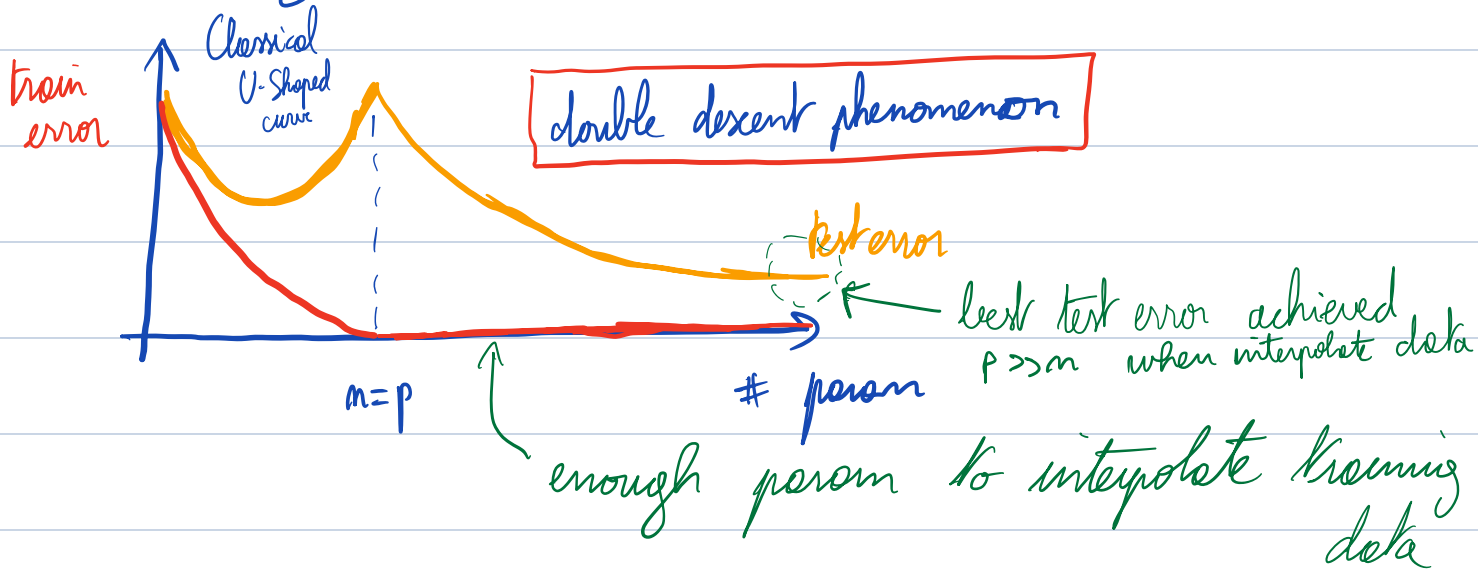
⇒ too simple : large bias, small variance

UNDERFITTING

Modern approach :

No careful control of the complexity of the model

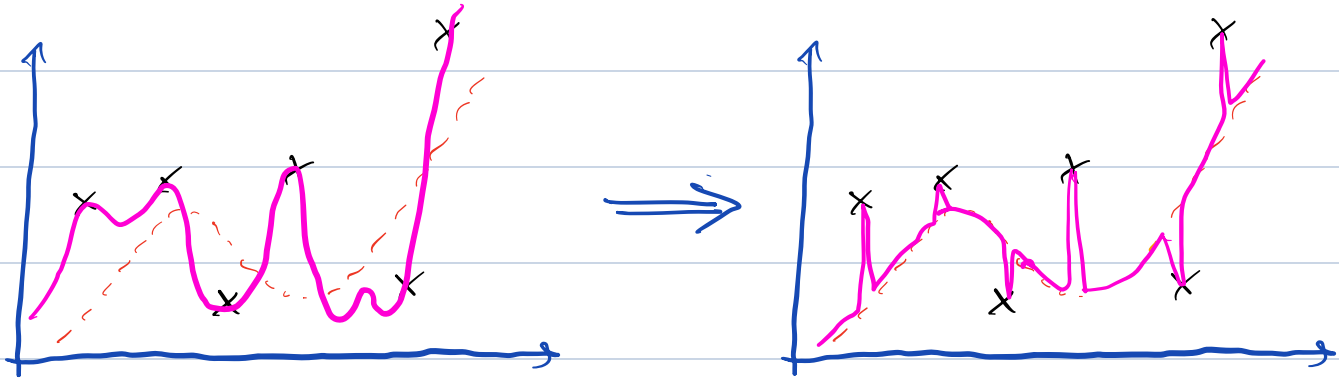
In fact, generalize well even when trained until interpolation of the training data $f(x_i, \hat{\theta}) = y_i = f(x_i, \theta_0) + \varepsilon_i$



Today: I will present a scenario where the balance between the bias and the variance is achieved not by carefully tuning an explicit parameter but by a novel phenomena

Self-induced regularization

⇒ Show sufficient & necessary conditions for benign overfitting in linear models



Expected picture for overfitting

Sometimes benign

Benign overfitting

only way to get $\hat{R}_n(\hat{f}) = 0$ and $R(\hat{f}) \ll 1$ when noise

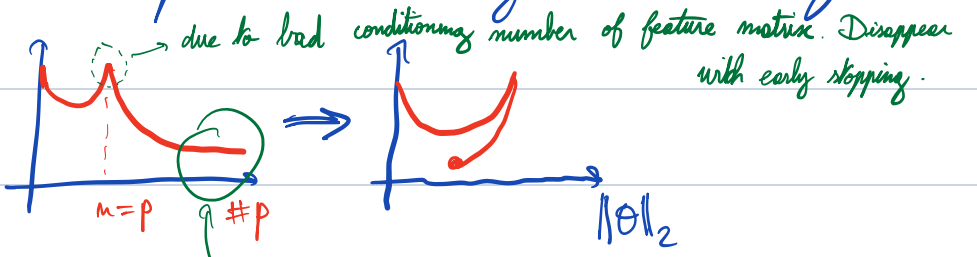
estimator: $\hat{f}(x, \hat{\theta}) = \hat{f}_0(x, \hat{\theta}_0) + \Delta(x)$

$\hat{f}_0(x, \hat{\theta}_0)$: good for prediction because smooth

$\Delta(x)$: spikes that are useful for interpolation but do not harm prediction because $\mathbb{E}_x[\Delta(x)^2] \ll 1$.

Rmk 1: Double descent: we saw that # param is not a good measure of model complexity

→ vs $\|\theta\|_2$



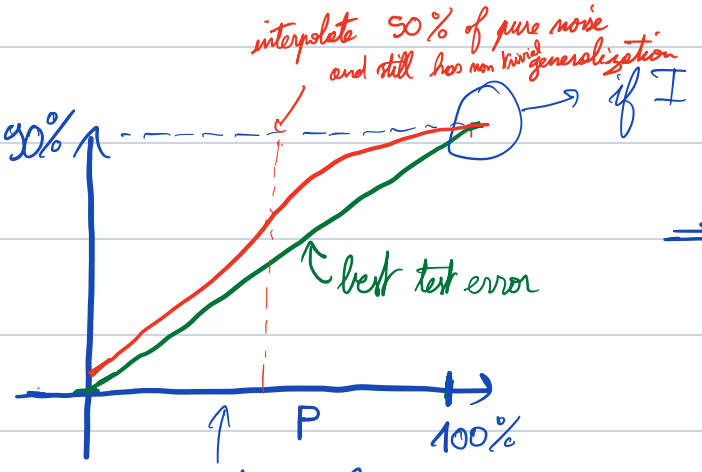
recover classical U-shaped curve

interesting phenomenon here is that we can interpolate noisy data and still generalize well.

Rmk 2: Can UC explain benign overfitting?

→ bounds from lecture 2 independent of the data distribution

"Understanding deep learning requires rethinking generalization" Zhang et al 2017



if I guess one of the 10 classes at random

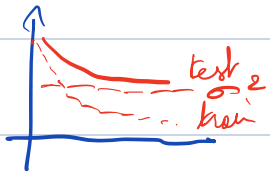
⇒ same NN trained to interpolate data with different level of corrupted samples

$Rd_m(F) = 1$) can interpolate pure noise data

Simple rule: $y_i = f_*(x_i) + \varepsilon_i$

$$\hat{R}_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i)) \leq (1-\kappa)\sigma^2$$

$$R(\hat{f}) = \sigma^2 + \mathbb{E}[(\hat{f}(x) - f_*(x))^2]$$



$$\varepsilon_n(F) \geq |\hat{R}_n(\hat{f}) - R(\hat{f})| \geq \kappa\sigma^2$$

bounded away from 0.

RIDGE REGRESSION

Setting: * Data $\{(y_i, x_i)\}_{i \le n}$ $\quad x_i \in \mathbb{R}^p$ $\quad y_i \in \mathbb{R}$

* $\mathbb{E}[x_i] = 0$ $\quad \mathbb{E}[x_i x_i^\top] = \Sigma$

$x_i = \Sigma^{1/2} z_i$

* x_i are iid $\quad y_i = \langle \theta_*, x_i \rangle + \varepsilon_i$

noise ε_i independent $\quad \mathbb{E}[\varepsilon_i] = 0 \quad \mathbb{E}[\varepsilon_i^2] = \sigma_\varepsilon^2$

Fit a linear model: $f(x, \theta) = \langle x, \theta \rangle \quad \theta \in \mathbb{R}^p$

Ridge regression: $\hat{\theta}(\lambda) = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|y - X\theta\|_2^2 + \lambda \|\theta\|_2^2 \right\}$

$X = \begin{matrix} \overleftarrow{\ p\ } \rightarrow \\ \uparrow \\ n \\ \downarrow \end{matrix} \begin{bmatrix} - x_1 - \\ \vdots \\ - x_n - \end{bmatrix} \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

$= X^\top S\, y$

where $S = (XX^\top + \lambda \mathrm{Id})^{-1}$

Limit $\lambda \searrow 0^+$: $\hat{\theta}(0^+) = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\theta\|_2 : y = X\theta \right\}$

$= X^\top S\, y \qquad S = (XX^\top)^+$

$\Rightarrow$ Minimum-$\|\cdot\|_2$ norm interpolating solution

↳ GD with $\theta^0 = 0$ converges to $\hat{\theta}(0^+)$

Excess test error: $R(\hat{\theta}) = \mathbb{E}_x\left[(f(x;\hat{\theta}) - f(x;\theta_*))^2\right]$

$= \mathbb{E}_x\left[(\langle x, \hat{\theta} - \theta_* \rangle)^2\right]$

$= \|\hat{\theta} - \theta_*\|_\Sigma^2$

$\|u\|_\Sigma = \|\Sigma^{\frac{1}{2}} u\|_2$ $\quad = (\hat{\theta} - \theta_*)^T \Sigma (\hat{\theta} - \theta_*)$

Bias - Variance decomposition:

$\overline{R}(\hat{\theta}(\lambda)) = \mathbb{E}_\varepsilon\left[R(\hat{\theta}(\lambda))\right] = \mathbb{E}_\varepsilon\left[\|\hat{\theta}(\lambda) - \theta_*\|_\Sigma^2\right]$

$X^T S y = X^T S X \theta_* + X^T S \varepsilon$

$= B(\lambda) + V(\lambda)$

where $B(\lambda) = \|(Id - X^T S X)\theta_*\|_\Sigma^2$

$V(\lambda) = \sigma^2 \operatorname{Tr}(S X \Sigma X^T S)$

Self-induced regularization : high level explanation

WLOG : assume $\Sigma = \mathrm{diag}(\lambda_1, \lambda_2, \dots \lambda_p)$ $\quad \lambda_1 \geq \lambda_2 \geq \dots$

Consider top k eigenspaces (corresponding to top k eigenspaces $\lambda_1 \dots \lambda_k$)

write $x = \begin{bmatrix} x_0 \\ x_+ \end{bmatrix}$ } first k coordinates
} $p-k$ last coordinates

$$X = \begin{bmatrix} — & x_1 & — \\ & \vdots & \\ — & x_n & — \end{bmatrix} = \overset{n}{\downarrow}\left[\begin{array}{c|c} \overset{k}{\longleftarrow} & \overset{p-k}{\longleftarrow} \\ X_0 & X_+ \end{array}\right]$$

$$\Sigma = \overset{k}{\downarrow}\left(\begin{array}{c|c} \Sigma_0 & \\ \hline & \Sigma_+ \end{array}\right) \quad \overset{p-k}{\longleftarrow}$$

$\theta = (\theta_0, \theta_+)$ $\quad \theta_0 = (\theta_1, \dots, \theta_k)$ $\quad \theta_+ = (\theta_{k+1}, \dots, \theta_p)$

$$XX^T = X_0 X_0^T + \underbrace{X_+ X_+^T}_{\text{assume} \approx \gamma \, \mathrm{Id}_n}$$

(later, assume $\frac{1}{c}\mathrm{Id} \lesssim M \lesssim c\,\mathrm{Id}$)

Recall: $\hat{\theta}(\lambda) = X^T S y \approx X^T (X_0 X_0^T + (\gamma + \lambda)\mathrm{Id})^{-1} y$

$$\longrightarrow \quad \hat{\theta}_0 = X_0^\top (X_0 X_0^\top + (\lambda+\gamma)\mathrm{Id})^{-1} y$$

top k coordinates

$$\longrightarrow \text{the same as} \quad \hat{\theta}_0 = \underset{\theta_0 \in \mathbb{R}^k}{\operatorname{argmin}} \left\{ \|y - X_0\theta_0\|_2^2 + (\lambda+\gamma)\|\theta_0\|_2^2 \right\}$$

"self induced regularization" from the high-degree part of the features $X_+ X_+^\top$

Even when $\lambda = 0^+$: * $\hat{\theta}_0$ solution of a lower-dim problem with effective regularization $\gamma > 0$.

* $\hat{\theta}_+$ can show that it is small and does not contribute to the test error but interpolate the training data

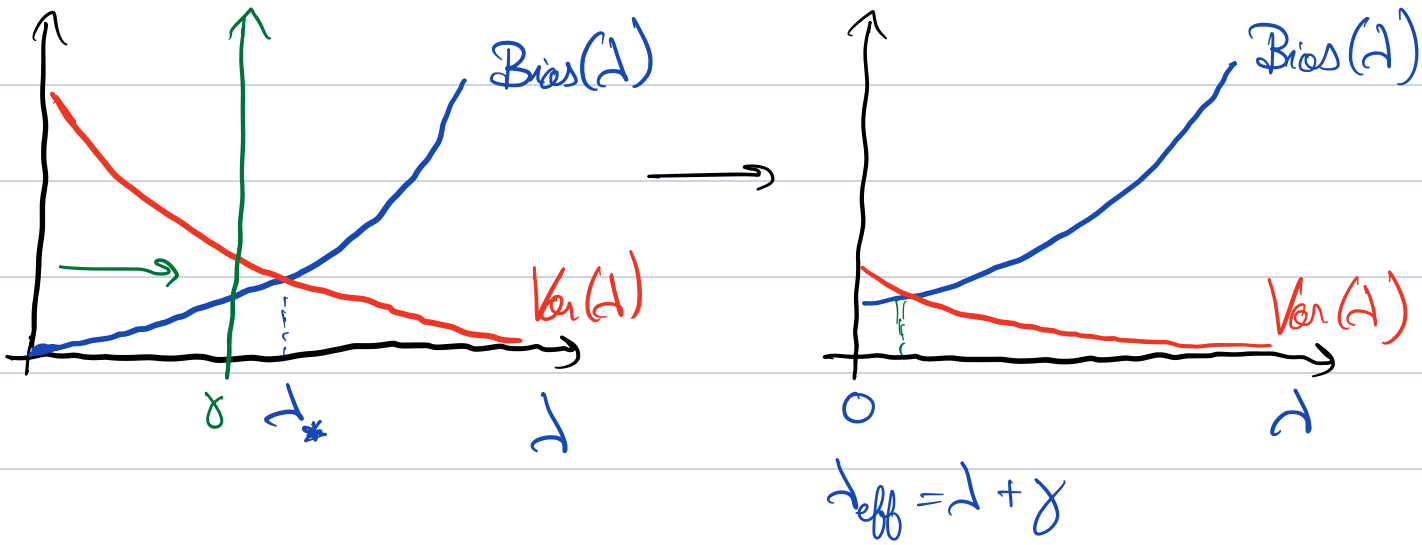
$$f(x, \hat{\theta}) = \underbrace{\langle \hat{\theta}_0, x_0 \rangle}_{\hat{f}_0(x)} + \underbrace{\langle \hat{\theta}_+, x_+ \rangle}_{\text{spiky part } \Delta(x)}$$

"signal part" of the features → generalize well thanks to γ

"noise part" of feature → interpolate data, doesn't contribute to test error

⇒ Noise part acts as an effective ridge regularization

Effective regularization $\lambda \longrightarrow \lambda_{\text{eff}} = \lambda + \gamma$



"Shifted Bias-Variance curve"

$\lambda = 0$ close to optimal regularization

"Benign Overfitting in Ridge Regression" – Bartlett, Trigler, 2020

$x_i = \Sigma^{\frac{1}{2}} z_i$ $\quad z_i$ iid sub-Gaussian (for simplicity) ↓ more general even for non linear kernels

Recall: $S = (\lambda Id + XX^T)^{-1} = (\lambda Id + X_+ X_+^T + X_0 X_0^T)^{-1}$

$S_+ = (\lambda Id + X_+ X_+^T)^{-1}$ $\qquad S = (S_+^{-1} + X_0 X_0^T)^{-1}$

THM [Bartlett, Trigler, '20] Assume w.h.p $\frac{\lambda_{max}(S_+)}{\lambda_{min}(S_+)} \leq L$

Then whp:

$$Bias(\lambda) \lesssim L^4 \left[\|\theta_0^*\|^2_{\Sigma_0^{-1}} \left(\frac{\lambda + \sum_{i>k} \lambda_i}{n} \right)^2 + \|\theta_+^*\|^2_{\Sigma_+} \right]$$

$$Var(\lambda) \lesssim \sigma_\varepsilon^2 L^2 \left[\frac{k}{n} + \frac{n \sum_{i>k} \lambda_i^2}{\left(\lambda + \sum_{i>k} \lambda_i\right)^2} \right]$$

RMK:

* Matching lower bound (up to multiplicative constants / minimax)
* In particular, interpolators $\lambda = 0$ can be near optimal.
* Does not fit θ_+^* at all
* Does not depend on p (if sums $\sum \lambda_i$ cv : applies to $p = \infty$)

B.O: small variance: need to choose k so that

(1) $k \ll n$ (2) $\dfrac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2} \gg n$

E.g. $d_{k+1} := \#$ eigenvalues of Σ in $\left[\dfrac{\lambda_{k+1}}{2}, \lambda_{k+1}\right]$

and assume eigenvalues $< \dfrac{\lambda_{k+1}}{2}$ are negligible

$\Rightarrow$ (2) requires $\dfrac{\left(d_{k+1} \cdot \lambda_{k+1}\right)^2}{d_{k+1} \cdot \lambda_{k+1}^2} = \boxed{d_{k+1} \gg n}$

and we have Variance $\lesssim \dfrac{k}{n} + \dfrac{n}{d_{k+1}}$

$\dfrac{k}{n}$: Variance of fitting θ_0^* (actually this is the parametric rate!)

$\dfrac{n}{d_{k+1}}$: variance of fitting the rest

effect of overfitting $\Downarrow$ negligible

Rmk: $\lambda_i = \frac{1}{i^\alpha (1+\log i)^\beta}$ decay of eigenvalues

Test error $(\lambda = 0^+) \longrightarrow 0$ as $n \longrightarrow \infty$ "consistency of interpolating solution"

IFF $\alpha = 1$ $\beta > 1$

To get benign overfitting here:

(1) Overparametrization: $p >> n$

directions in parameter space unimportant for prediction >> sample size

(2) Smallest eigenvalues of Σ must decay slowly

$d_{k+1} >> n$

Self-induced regularization $\Sigma^{\frac{1}{2}} z_i$ z_i iid c-sub-G coordinates hold way more generally

Lemma: $\exists c > 0$ s.t with proba at least $1 - 2\exp(-\frac{n}{c})$

$$\frac{\lambda_{max}(X_+ X_+^T)}{\lambda_{min}(X_+ X_+^T)} \leq c \frac{\sum_{i>k} \lambda_i + n\lambda_{k+1}}{\sum_{i>k} \lambda_i - cn\lambda_{k+1}} \leq L$$

$n\lambda_{k+1} \ll \sum_{i>k} \lambda_i$

Example: Kernel regression with inner product kernel on the sphere

$u_i \overset{iid}{\sim} \mathrm{Unif}\left(S^{d-1}(\sqrt{d})\right)$ $\qquad y_i = f_*(u_i) + \varepsilon_i$ → general squared integrable function

Linear model: $\hat{f}(u;\theta) = \langle \theta, \phi(u) \rangle$

$$\phi(u) = \Big(\xi_0, \xi_1 u_1, \dots \xi_1 u_d \;\Big|\; \xi_2 Y_{21}(u), \dots \xi_2 Y_{2d^2}(u) \;\Big|\; \vdots \;\Big|\; \xi_k Y_{k1}(u), \dots, \xi_k Y_{k\mathcal{O}(d^k)}(u) \;\Big|\; \vdots \Big)$$

$\xi_k = \mathcal{O}(d^{-\frac{k}{2}})$

) basis of degree k polynomials in u. $\mathcal{O}(d^k)$

$x_i = \phi(u_i) = \Sigma^{\frac{1}{2}} z_i \qquad \mathbb{E}[z_i z_i^\top] = \mathrm{Id}_\infty$

$$\Sigma = \begin{pmatrix} \xi_0^2 & & & & \\ & \xi_1^2 \;(\text{d times}) & & & \\ & & \xi_2^2 \;(\mathcal{O}(d^2)\text{ times}) & & \\ & & & \ddots & \\ & & & & \ddots \end{pmatrix}$$

In $d^\ell \ll n \ll d^{\ell+1}$

Take $k := d^\ell$ $\qquad X_+ X_+^T \approx \left(\sum_{s=\ell+1}^{\infty} B_{d,s}\, \xi_s^2\right) \mathrm{Id}_n$

$$f_*(u) = \underbrace{\langle x_0, \theta_{*,0}\rangle}_{\text{best degree-}\ell\text{ polynomial approx}} + \langle x_+, \theta_{*,+}\rangle$$

$$\hat\theta_0 \simeq \theta_{*,0} \qquad \mathbb{E}_x\left\{\langle \hat\theta_+, x_+\rangle^2\right\} = \|\hat\theta_+\|_{\Sigma_+}^2 \simeq 0$$

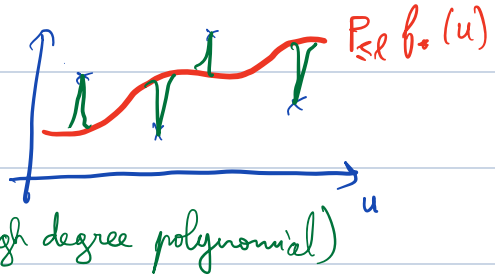
[needs very different tools to analyze $\longrightarrow$ not sub-Gaussian]

Thm [M., Ghorbani, Mei, Montanari, '19]

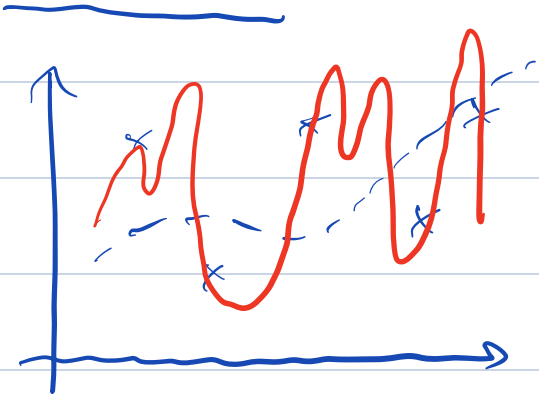
If $d^{\ell+\delta} \le n \le d^{\ell+1-\delta}$ as $n, d \to \infty$, then

$$\text{Test error} = \|P_{>\ell} f_*\|_{L^2}^2 + o(1)$$

$$\hat f(u) = P_{\le \ell} f_*(u) + \underbrace{\Delta(u)}_{\text{spiky part}}$$



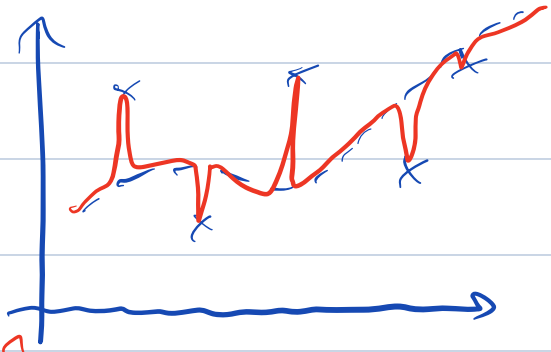
Conclusion:



classical picture

BAD OVERFITTING

$\Downarrow$ Modern picture



sometimes

BENIGN OVERFITTING

$\hat{f} = f_0 + \Delta$

Above "low-dim" (classical) picture misleading

$\longrightarrow$ much easier to interpolate benignly in high dim

(Of course, it required a lot of work to realize it)

Proof of Bartlett / Tsigler

$$S = (X_0 X_0^T + X_+ X_+^T + \lambda)^{-1} \qquad S_+ = (X_+ X_+^T + \lambda)^{-1}$$

$$\frac{\lambda_{max}(S_+)}{\lambda_{min}(S_+)} \leq L$$

Here only prove Variance term $(\sigma_\varepsilon^2 = 1$ WLOG $)$

$$V(X, \lambda) = \mathrm{Tr}(S X \Sigma X S) = V_0 + V_+ \qquad V_0 := \mathrm{Tr}(S X_0 \Sigma_0 X_0^T S)$$
$$V_+ := \mathrm{Tr}(S X_+ \Sigma_+ X_+^T S)$$

1st Term: V_0 = variance along top k directions $\lesssim L^2 \frac{k}{n}$

Proof:

$$S X_0 = (X_0 X_0^T + S_+^{-1})^{-1} X_0$$
$$= S_+ X_0 (I + X_0^T S_+ X_0)^{-1}$$

Introduce $M := \Sigma_0^{\frac{1}{2}} (I + X_0 S_+ X_0^T)^{-1} \Sigma_0^{\frac{1}{2}}$

$$V_0 = \mathrm{Tr}(S_+ X_0 \Sigma_0^{-\frac{1}{2}} M^2 \Sigma_0^{-\frac{1}{2}} X_0^T S_+) \qquad (= Z_0)$$

$$V_0 \le \|M\|_{op}^2 \; \mathrm{Tr}\left(S_+ Z_0 Z_0^T S_+\right)$$

$$\|M\|_{op} = \left\|\left(\Sigma_0^{-1} + \Sigma_0^{-\frac{1}{2}} X_0^T S_+ X_0 \Sigma_0^{-\frac{1}{2}}\right)\right\|_{op} \le \frac{1}{\lambda_{min}\left(Z_0^T S_+ Z_0\right)}$$

$$\le \frac{1}{\lambda_{min}(S_+)\,\lambda_{min}\left(Z_0^T Z_0\right)}$$

Hence

$$V_0 \le \frac{\lambda_{max}(S_+)^2 \; \mathrm{Tr}\left(Z_0^T Z_0\right)}{\lambda_{min}(S_+)^2 \cdot \lambda_{min}\left(Z_0^T Z_0\right)^2}$$

$$\frac{\mathrm{Tr}(A)}{\lambda_{min}(A)} \le \frac{\lambda_{max}(A)}{\lambda_{min}(A)} \cdot \mathrm{rank}(A)$$

$$\le L^2 \; \frac{\lambda_{max}\left(Z_0^T Z_0\right)}{\lambda_{min}\left(Z_0^T Z_0\right)} \times \frac{k}{\lambda_{min}\left(Z_0^T Z_0\right)}$$

$$\left[Z_0 \in \mathbb{R}^{n \times p} \quad \text{iid} \quad O(1)\text{-sub Gaussian} \quad k \le c\,n \right.$$
$$\left. \lambda_{min}(Z_0),\ \lambda_{max}(Z_0) \asymp \sqrt{n} \right]$$

$$V_0 \lesssim L^2 \cdot \frac{k}{n} \qquad \square$$

2nd term : V_t = variance due to overfitting

$$\lesssim L^2 \frac{n \sum_{i>k} d_i^2}{\left(\lambda + \sum_{i>k} d_i\right)^2}$$

Rmk : $\frac{\lambda_{max}(S_t)}{\lambda_{min}(S_t)} \leq L \Rightarrow \lambda_{max}(S_t) \lesssim \frac{L}{\lambda + \sum_{i>k} d_i}$

$$\lambda_{max}(S_t) \leq L\, \lambda_{min}(S_t) \leq L \frac{n}{Tr(S_t^{-1})}$$

$$\frac{1}{n} Tr(S_t^{-1}) = \frac{1}{n} Tr(\lambda Id + X_t X_t^T) = \lambda + \frac{1}{n} \sum_{i=1}^{n} \|x_{i,t}\|_2^2$$

WLLN

$$\simeq \lambda + Tr(\Sigma_t)$$

Proof : $V_t = Tr(S X_t \Sigma_t X_t^T S) \leq \lambda_{max}(S)^2 Tr(X_t \Sigma X_t^T)$

$$\lesssim \lambda_{max}(S_t)^2 \sum_{i=1}^{n} \langle z_{i,t}, \Sigma_t^2 z_{i,t} \rangle$$

$$\lesssim \lambda_{max}(S_t)^2 \, n \cdot Tr(\Sigma_t^2)$$

Using the above remark

$$V_+ \leq L^2 \frac{n \cdot \mathrm{Tr}(\Sigma_+^2)}{\left(\lambda + \sum_{i>k} \lambda_i\right)^2} \qquad \square$$

Sharp asymptotics using RMT

So far we only used matrix concentration
↳ in particular: bounds hold for $n \geq C$.

⇒ Lots of work have considered to derive sharp formulas for the bias and variance as $n, p \to \infty$ using RMT

Below a very brief presentation of these results

[Hastie, Montanari, Rosset, Tibshirani, 2019]
(Follow Andrea's presentation)

Setting: * $\{(x_i, y_i)\}_{i \leq n}$ $y_i \in \mathbb{R}$ $x_i \in \mathbb{R}^p$ iid

* $y_i = \langle \theta_*, x_i \rangle + \varepsilon_i$ $\mathbb{E}\varepsilon_i = 0$ $\mathbb{E}\varepsilon_i^2 = \sigma_\varepsilon^2$

* $x_i = \Sigma^{\frac{1}{2}} z_i$ $\mathbb{E} z_i = 0$ $\mathbb{E}[z_i z_i^T] = I_p$

* z_i independent coordinates $\mathbb{E}[|z_{ij}|^k] \leq C_k < \infty$ $\forall i, k$

Ridge solution $\hat{\theta}(\lambda) = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_2^2 \right\} = \frac{1}{n} X^T \left(\lambda + \frac{XX^T}{n}\right)^{-1} y$

$$R(\hat{\theta}) = \|\hat{\theta}(\lambda) - \theta_*\|_\Sigma^2 = B_n + V_n$$

$$B_n = \lambda^2 \langle \theta_*, S_\lambda \Sigma S_\lambda \theta_* \rangle \qquad \hat{\Sigma} = \frac{1}{n} X^T X$$

$$V_n = \frac{\sigma^2}{n} \operatorname{Tr}(\Sigma \hat{\Sigma} S_\lambda^2) \qquad S_\lambda = (\hat{\Sigma} + \lambda I)^{-1}$$

Thm [Hastie et al, 19] $\Sigma = I$ $\lambda = 0^+$ (min-norm)

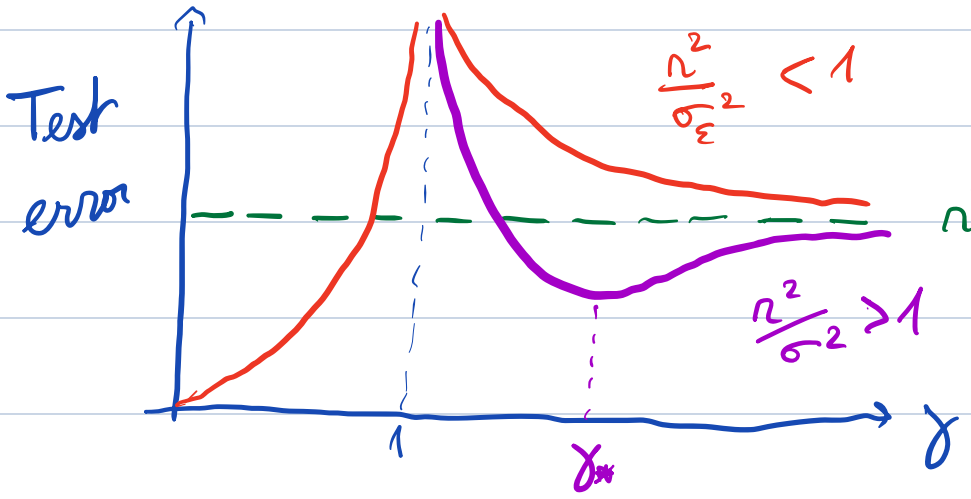
$n, p \to \infty$ $\frac{p}{n} \to \gamma \in (0, \infty)$ $\|\theta_*\|_2^2 \to r^2$

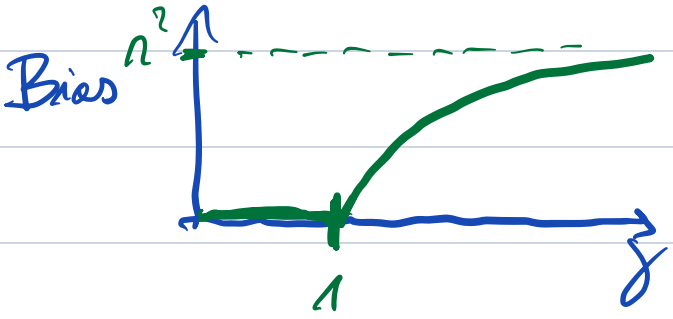
Then $$\lim_{n,p \to \infty} B_n = r^2 \left(1 - \frac{1}{\gamma}\right)_+$$

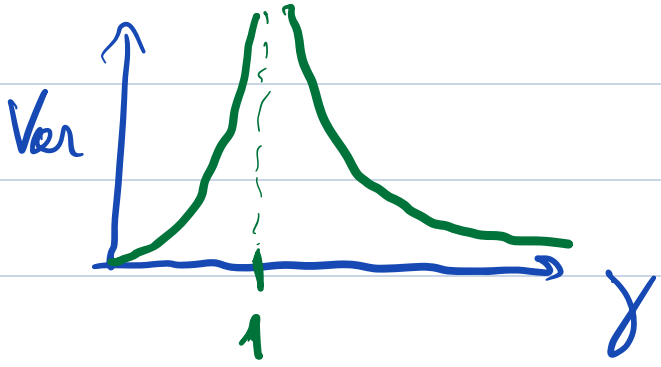
$$\lim_{n,p \to \infty} V_n = \sigma_\varepsilon^2 \frac{1}{\left(\frac{1}{\gamma} \vee \gamma - 1\right)}$$

In particular,

$$\lim_{n,p \to \infty} R(\hat{\theta}) = \begin{cases} \dfrac{\gamma \sigma_\varepsilon^2}{1-\gamma} & \text{if } \gamma < 1 \\ r^2\left(1 - \frac{1}{\gamma}\right) + \dfrac{\sigma^2}{\gamma - 1} & \text{if } \gamma > 1 \end{cases}$$







Rmk:

* Prove way more stuff in their paper
* Double descent
  → however not minimizer $\gamma \to \infty$
* Not a good model $\underline{p = d}$
  ↳ true fct $f_*(x) = \langle \theta_*, x \rangle$ becomes more complex as p grows
* Better models
  [Mei, Montanari, '19] ...

Crash course Random Matrix Theory (RMT)

$\hat{\Sigma} = \frac{XX^T}{n} \in \mathbb{R}^{p \times p}$ empirical spectral distribution (ESD)

$$\mu_n = \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(\hat{\Sigma})}$$

Marchenko - Pastur law: $\mu_n \xrightarrow[\substack{n,p \to \infty \\ n/p \to \gamma}]{\text{(weak cv)}} \mu_\infty$

To prove it: Stieltjes transform $z \in \mathbb{C}$

$$M_n(z) = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{\lambda_i(\hat{\Sigma}) - z} = \frac{1}{p} \mathrm{Tr}\left((\hat{\Sigma} - z I_p)^{-1}\right)$$

Thm: [MP] Let $z \mapsto m(z)$ be the analytic fct on $z \in \mathbb{C}_+$ s.t $|m(z)| \leq \frac{C}{|z|}$ as $|z| \geq C$ and solution of

$$\frac{m}{1+zm} = 1 + \gamma m$$

Then a.s. $M_n(z) \longrightarrow m(z)$ as $\substack{n,p \to \infty \\ n/p \to \gamma}$

Stieltjes inversion :

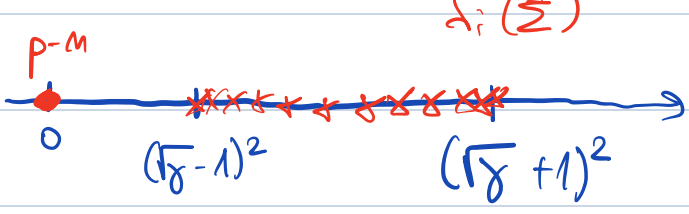
$$\mu([a,b]) = \lim_{\eta \to 0} \int_a^b \frac{1}{\pi} \operatorname{Im}\left(m(\lambda + i\eta) \right) d\lambda$$

Rmk : $m(z) = \dfrac{1 - z - \gamma - \sqrt{(1-z-\gamma)^2 - 4\gamma z}}{2\gamma z}$ by solving the fixed pt equat°

Thm : [Bai-Yin law] $\gamma > 1$ $\operatorname{rank}(\hat{\Sigma}) = n$ w.h.p

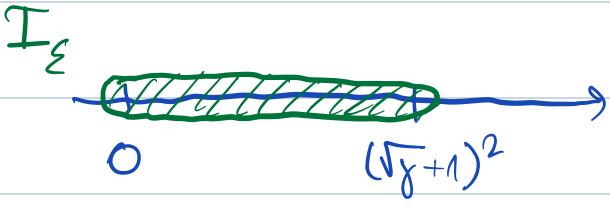
$$\begin{cases} \lambda_{max}(\hat{\Sigma}) = (\sqrt{\gamma} + 1)^2 + o_p(1) \\ \lambda_{min}(\hat{\Sigma}) = (\sqrt{\gamma} - 1)^2 + o_p(1) \end{cases}$$

For $\gamma \leq 1$ $\lambda(\hat{\Sigma}) = 0$ $p - n$



Rmk: $M_n(z)$ is unif. cont. (differentiable) on $\mathbb{C} \setminus I_\varepsilon$

$$I_\varepsilon = \left\{ z : \begin{array}{l} Re(z) \in \left[-\varepsilon, (\sqrt{\gamma}+1)^2 + \varepsilon\right] \\ |Im(z)| \leq \varepsilon \end{array} \right\}$$



All derivatives unif. bounded outside I_ε

(convince yourself using expression of $M_n(z)$)

$M_n(z) \xrightarrow{a.s.} m(z) \quad \forall z \notin I_\varepsilon$.

and some for derivatives

Take $z = -\lambda$ $\quad M_n(-\lambda) = \frac{1}{p} Tr\left((\hat{\Sigma} + \lambda)^{-1}\right) \longrightarrow m(-\lambda)$

Proof of Thm:

Variance $\frac{1}{\sigma_\varepsilon^2} V_n = \frac{1}{n} \mathrm{Tr}\left(\hat{\Sigma}(\hat{\Sigma}+\lambda)^{-2}\right)$

$$= \frac{p}{n}\left[\frac{1}{p}\mathrm{Tr}\left((\hat{\Sigma}+\lambda)^{-1}\right) - \frac{1}{p}\lambda\,\mathrm{Tr}\left((\hat{\Sigma}+\lambda I)^{-2}\right)\right]$$

$$= \frac{p}{n}\left[M_n(-\lambda) + \lambda \frac{\partial}{\partial\lambda} M_n(-\lambda)\right]$$

$$\longrightarrow \gamma\left[m(-\lambda) + \lambda \frac{\partial}{\partial\lambda} m(-\lambda)\right]$$

$$= \gamma \frac{\partial}{\partial\lambda}\left(\lambda\, m(-\lambda)\right)$$

$$m(-\lambda) = \frac{-1-\lambda-\gamma + \sqrt{(1+\lambda-\gamma)^2 + 4\lambda\gamma}}{2\gamma\lambda}$$

Ridgeless limit: $\lambda \downarrow 0 \quad \gamma > 1 \quad m(-\lambda) = \frac{\gamma-1}{\gamma\lambda} + \frac{1}{\gamma(\gamma-1)} + O(\lambda)$

$$\frac{\partial}{\partial\lambda}\left(\lambda\, m(-\lambda)\right) = \frac{1}{\gamma(\gamma-1)} + O(\lambda)$$

$$\lim_{\lambda\to 0} \lim_{n,p\to\infty} V_n(\lambda) = \frac{\sigma_\varepsilon^2}{\gamma-1} \qquad \gamma > 1$$

Rmk: * need to invert the limits (doable with some work)

* same for $\gamma < 1$

Bias $\quad B_n = \lambda^2 \langle \theta_*, (\hat{\Sigma} + \lambda)^{-2} \theta_* \rangle$

x isotropic : $\quad B_n = \lambda^2 \|\theta_*\|_2^2 \frac{1}{p} \mathrm{Tr}\left((\hat{\Sigma} + \lambda)^{-2}\right) + o_p(1)$

[e.g. prove it for $x \sim N(0, \mathrm{Id}_d)$]

$$B_n = \lambda^2 \|\theta_*\|_2^2 \left(- \frac{\partial}{\partial \lambda} M_n(-\lambda) \right)$$

$$\longrightarrow r^2 \left(1 - \frac{1}{\gamma}\right)_+ \quad \text{with some calculus}$$

□

Anisotropic case

$\Sigma \neq I$ risk will depend on both Σ, θ_*

Interpolating solution $\lambda = 0^+$

"Effective regularization" λ_* unique positive solution of

$$\frac{1}{\gamma} = \frac{1}{p} \operatorname{Tr}\left(\Sigma (\Sigma + \lambda_* I)^{-1}\right)$$

$$\overline{B}_n := \frac{\lambda_*^2 \, \langle \theta_*, (\Sigma + \lambda_*)^{-2} \Sigma \theta_* \rangle}{1 - \frac{1}{n} \operatorname{Tr}\left(\Sigma^2 (\Sigma + \lambda_*)^{-2}\right)}$$

$$\overline{V}_n := \frac{\sigma^2}{n} \frac{\operatorname{Tr}\left(\Sigma^2 (\Sigma + \lambda_*)^{-2}\right)}{1 - \frac{1}{n} \operatorname{Tr}\left(\Sigma^2 (\Sigma + \lambda_*)^{-2}\right)}$$

Thm [Mastie et al, '19]

$$\begin{cases} \lim_{n,p \to \infty} |B_n - \overline{B}_n| = 0 \\ \lim_{n,p \to \infty} |V_n - \overline{V}_n| = 0 \end{cases} \quad \text{a.s.}$$

Interpretation: test error:

Ridge regression $\longleftrightarrow$ gaussian sequence model

$$y^s = \Sigma^{\frac{1}{2}}\theta_* + \frac{\omega}{\sqrt{n}} g \qquad g \sim N(0, I_p)$$

$$\hat{\theta}^s = \operatorname*{argmin}_{\theta} \left\{ \| y^s - \Sigma^{\frac{1}{2}}\theta \|_2^2 + \lambda_* \|\theta\|_2^2 \right\}$$

ω fixed pt of: $\omega^2 = \frac{\sigma_\varepsilon^2}{n} + \mathbb{E}_g\left[\|\hat{\theta}^s - \theta_*\|_\Sigma^2 \right]$

$$R_n^s = \mathbb{E}\left[\|\hat{\theta}^s - \theta_*\|_\Sigma^2 \right] = \underset{\|}{B_n^s} + \underset{\|}{V_n^s}$$

$$\bar{B}_n \qquad \bar{V}_n$$

	Original problem	Sequence model
design	X random	$\Sigma^{\frac{1}{2}}$ deterministic
reg.	$\lambda = 0^+$	$\lambda_* > 0$!!
noise	σ_ε^2	ω^2

self induced regularization !! more regularization in the model than what is naively expected from $\lambda = 0$.

Appendix : Concentration of $X_+ X_+^\top$

I'll present a nice decoupling argument that applies under quite general conditions (much beyond sub-Gaussian e.g. inner-product kernels)

$\{u_i\}_{i \in [n]}$ iid $\quad K = (K(u_i, u_j))_{i,j \in [n]} \in \mathbb{R}^{n \times n}$

$K : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ p.s.d. fct $\to$ p.s.d operator $\mathbb{K}$ (lecture on kernels)

$$K(u_i, u_j) = \sum_{k=1}^{\infty} \underbrace{\xi_k^2}_{\text{eigenvalues}} \underbrace{\phi_k(u_i)\,\phi_k(u_j)}_{\text{eigenvectors}}$$

EIGENDECOMPOSITION

$\xi_1^2 \geq \xi_2^2 \geq \dots$

$\mathbb{E}[\phi_k(u)\phi_\ell(u)] = \delta_{k\ell}$

$\|\mathbb{K}\|_{op} = \xi_1^2$

$\mathrm{Tr}(\mathbb{K}) = \sum_{k=1}^{\infty} \xi_k^2 < \infty$ ("trace-class")

$= \mathbb{E}[K(u,u)]$

Assumptions :

(A) Diagonal: $\mathbb{E}\left[\max_{i\in[n]} |K(u_i,u_i) - \mathrm{Tr}(K)|\right] \leq C\sqrt{n\|K\|_{op}\mathrm{Tr}(K)}$

(B) Hypercontractivity: $\forall p$ integer ≥ 2, $\exists C_p$ s.t.

$$\forall v \in \mathbb{R}^\infty \quad \mathbb{E}\left[\left(\sum v_k \phi_k(u)\right)^p\right]^{\frac{1}{p}} \leq C_p \|v\|_2$$

Thm: [M., Mei, Montanari, 2021] For all $\delta > 0$, $\exists C$ s.t.

$$\mathbb{E}\left[\|K - \mathrm{Tr}(K) I_n\|_{op}\right] \leq C\sqrt{n^{1+\delta}\|K\|_{op}\mathrm{Tr}(K)}$$

Rmk: * Bounds on expectation : use Markov to get w.h.p
↳ to get better dependency in proba → sub-G : union bound
↘ hyper : moment method

can also improve $n^{1+\delta} \to n$

* $(\phi_k(u_i))_{k\geq 1} \to z_i$ i.i.d sub-Gaussian entries
→ easy to check (A) & (B)

* $\lambda_{max}(K) \leq \mathrm{Tr}(K)\left(1 + C\sqrt{\frac{n^{1+\delta}\|K\|_{op}}{\mathrm{Tr}(K)}}\right)$

$\lambda_{min}(K) \geq \mathrm{Tr}(K)\left(1 - C\sqrt{\frac{n^{1+\delta}\|K\|_{op}}{\mathrm{Tr}(K)}}\right)$

We will use the following result in random matrix concentration

Prop [Vershynin 2010 : Theorem 5.48] Let $A \in \mathbb{R}^{n \times k}$ with $A = [a_1, \dots, a_n]^T$ where $a_i \in \mathbb{R}^k$ are indep random vectors with common second moment matrix $\Sigma = \mathbb{E}[a_i a_i^T]$. Let $\Gamma = \mathbb{E}\left[\max_{i \in [n]} \|a_i\|_2^2\right]$. Then there exists a universal constant C such that

$$\mathbb{E}\left[\|A\|_{op}^2\right]^{1/2} \leq \left(\|\Sigma\|_{op} \cdot n\right)^{1/2} + C \cdot \left(\Gamma \cdot \log(n \wedge k)\right)^{1/2}$$

Proof of theorem: $K = \mathrm{ddiag}(K) + \Delta$

where $\mathrm{ddiag}(K) = \mathrm{diag}\left(K(u_i, u_i)\right)_{i \in [n]}$

$$\Delta_{ij} = \begin{cases} K(u_i, u_j) & i \neq j \\ 0 & \text{if } i = j. \end{cases}$$

so that $\mathbb{E}\left[\|K - \mathrm{Tr}(K) \cdot I_n\|_{op}\right]$

$$\leq \underbrace{\mathbb{E}\left[\|\mathrm{ddiag}(K) - \mathrm{Tr}(K) \cdot I_n\|_{op}\right]}_{(I)} + \underbrace{\mathbb{E}\left[\|\Delta\|_{op}\right]}_{(II)}$$

Using assumption (A):

$$(I) = \mathbb{E}\left[\max_{i \in [n]} |K(u_i, u_i) - \mathrm{Tr}(\mathbb{K})| \right] \leq \sqrt{n \|\mathbb{K}\|_{op} \mathrm{Tr}(\mathbb{K})}$$

For the second term: note that Δ does not have independent rows or columns and we can't use the above proposition directly.
We will use a standard decoupling argument:

Lemma: Let $\Delta_{T_1, T_2} = (\Delta_{ij})_{i \in T_1, j \in T_2} \in \mathbb{R}^{|T_1| \times |T_2|}$

$$\mathbb{E}[\|\Delta\|_{op}] \leq 4 \max_{T \subseteq [n]} \mathbb{E}[\|\Delta_{T, T^c}\|_{op}]$$

Proof of lemma: For $\|v\|_2 = 1$

$$v^T \Delta v = \sum_{i \neq j} v_i \Delta_{ij} v_j = 4 \mathbb{E}_T \left[\sum_{\substack{i \in T \\ j \in T^c}} v_i \Delta_{ij} v_j \right]$$

where T is a random subset of $\{1, \dots, n\}$ where each element is selected with prob $1/2$

Therefore $\mathbb{E}\left[\|\Delta\|_{op}\right] = \mathbb{E}\left[\sup_{v \in S^n} v^T \Delta v\right]$

$$\lesssim 4\, \mathbb{E}_T\, \mathbb{E}\left[\sup_{v \in S^n} \sum_{\substack{i \in T \\ j \in T^c}} v_i \Delta_{ij} v_j\right] \leq 4 \sup_{T \subseteq [n]} \mathbb{E}\left[\|\Delta_{T,T^c}\|_{op}\right]$$

□

Using this lemma, it suffices to bound $\|\Delta_{T,T^c}\|_{op}$ for all $T \subseteq n$ $\quad \Delta_{T,T^c} = (\Delta_{ij})_{i \in T,\, j \in T^c}$

Each row is $\Delta_{i,T^c} = (K(u_i, u_j))_{j \in T^c}$

so conditional on $(u_j)_{j \in T^c}$, Δ_{T,T^c} has iid rows!!

Denote $\mathbb{E}_T$ the expectation over $(u_i)_{i \in T}$ conditional on $(u_j)_{j \in T^c}$. We can use the above general concentration bound

$$\mathbb{E}_T\left[\|\Delta_{T,T^c}\|_{op}\right] \leq \left\{\|\Sigma_T\|_{op}\, |T|\right\}^{1/2} + C\left\{\Gamma_T \cdot \log\left(|T| \wedge |T^c|\right)\right\}^{1/2}$$

where we denoted

$$\Sigma_T := \mathbb{E}_{u_i}\left[\Delta_{i,T^c} \Delta_{i,T^c}^T\right] \qquad \Gamma_T := \mathbb{E}_T\left[\max_{i \in [n]} \|\Delta_{i,T^c}\|_2^2\right]$$

Hence

$$\mathbb{E}\left[\|\Delta\|_{op}\right] \lesssim q \sup_{T \subseteq [m]} \left\{ \left(\mathbb{E}_{T^c}\left[\|\Sigma_T\|_{op}\right] \cdot n\right)^{\frac{1}{2}} + C \left(\mathbb{E}_{T^c}\left[\Gamma_T\right] \cdot \log n\right)^{\frac{1}{2}} \right\}$$

Bounding $\mathbb{E}_{T^c}\left[\|\Sigma_T\|_{op}\right]$

$$\|\Sigma_T\|_{op} = \left\| \mathbb{E}_{u_i}\left[\Delta_{iT^c} \Delta_{iT^c}^\top\right] \right\|_{op} = \sup_{\|v\|_2 = 1} \sum_{ij \in T^c} \sum_{k}^{q} \xi_k^2 \phi_k(u_i) \phi_k(u_j) v_i v_j$$

$$\lesssim \|K\|_{op} \sup_{\|v\|_2 = 1} \sum_{ij \in T^c} \xi_k^2 \phi_k(u_i) \phi_k(u_j) v_i v_j$$

$$\lesssim \|K\|_{op} \left\| \left(K(u_i, u_j)\right)_{ij \in T^c} \right\|_{op}$$

$$\lesssim \|K\|_{op} \left(\|\operatorname{ddiag} K\|_{op} + \|\Delta\|_{op} \right)$$

By hypercontractivity:

$$\mathbb{E}\left[\|\operatorname{ddiag} K\|_{op}\right] \leq \mathbb{E}\left[\max_{i \in [n]} K(u_i, u_i)^p\right]^{\frac{1}{p}} \leq \mathbb{E}\left[\sum_{i=1}^{n} K(u_i, u_i)^p\right]^{\frac{1}{p}}$$

$$\lesssim n^{\frac{1}{p}} \mathbb{E}\left[K(u_i, u_i)^p\right]^{\frac{1}{p}} \lesssim C_p \, n^{\frac{1}{p}} \, \mathrm{Tr}(K)$$

Hence

$$\mathbb{E}_{T^c}[\Sigma_T] \leq C_p\, n^{\frac{1}{p}} \|K\|_{op}\, \mathrm{Tr}(K) + \|K\|_{op}\, \mathbb{E}[\|\Delta\|_{op}]$$

Bounding $\mathbb{E}_{T^c}[\Gamma_T]$ By hypercontractivity,

$$\mathbb{E}_{T^c}[\Gamma_T] = \mathbb{E}\Big[\max_{i\in T} \|\Delta_{i T^c}\|_2^2\Big]$$

$$\leq n \cdot \mathbb{E}\Big[\max_{i\in T,\, j\in T^c} \Delta_{ij}^2\Big]$$

$$\leq n \cdot \mathbb{E}\Big[\sum_{ij\in[n]} \Delta_{ij}^{2p}\Big]^{\frac{1}{p}}$$

$$\leq n^{1+\frac{2}{p}} \cdot \mathbb{E}[\Delta_{ij}^{2p}]^{1/p}$$

$$\leq n^{1+\frac{2}{p}}\, C_p^2\, \mathbb{E}[\Delta_{ij}^2]$$

$$\mathbb{E}_{T^c}[\Gamma_T] \leq n^{1+\frac{2}{p}}\, C_p^2\, \|K\|_{op}\, \mathrm{Tr}(K)$$

Putting everything together:

$$\mathbb{E}\left[\|\Delta\|_{op}\right] \le K_p \left\{ \|K\|_{op} \, \mathrm{Tr}(K) \, n^{1+\frac{2}{p}} \log n \right\}^{1/2}$$

$$+ K_p \left\{ n \cdot \|K\|_{op} \, \mathbb{E}\left[\|\Delta\|_{op}\right] \right\}^{1/2}$$

[Use that $x^2 - \varepsilon_1 x - \varepsilon_2 \le 0 \Rightarrow x \le \dfrac{\varepsilon_1 + (\varepsilon_1^2 + 4\varepsilon_2)^{1/2}}{2}$

$$\le (\varepsilon_1^2 + 4\varepsilon_2)^{\frac{1}{2}}$$

$$\mathbb{E}\left[\|\Delta\|_{op}\right] \le K_p \left\{ n \|K\|_{op} + \left\{ \|K\|_{op} \mathrm{Tr}(K) \, n^{1+\frac{2}{p}} \log n \right\}^{\frac{1}{2}} \right\}$$

□