A theory of linear models:

Consider, a set of observations of a system, $(x_1, y_1), (x_2, y_2)$ ..... (x_n, y_n)

The underlying function, generating these observations, is non-linear, complicated & unknown:

$$y_n = g(x_n)$$

A linear model seeks a linear approximation to the true 'g' model (which can be generally non linear)

So, we assume that there exists, a set of coefficients $\{\alpha_1, \alpha_2\}$ that describes the linear model.

Page 3

$$y_1 = \alpha_1 x_1 + \alpha_2$$

$$y_2 = \alpha_1 x_2 + \alpha_2$$

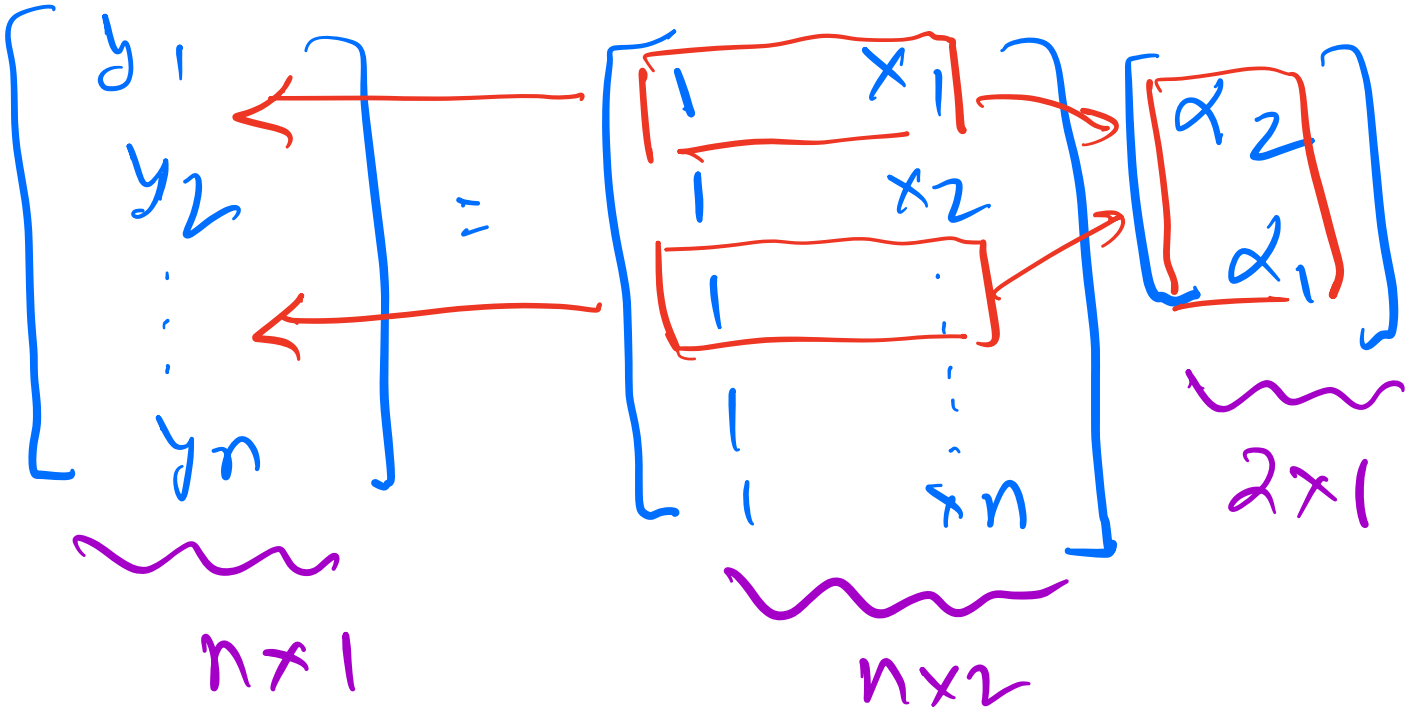
$$\vdots$$

$$y_n = \alpha_1 x_n + \alpha_2$$

We can write this in matrix fashion:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \alpha_1 \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

Page 4



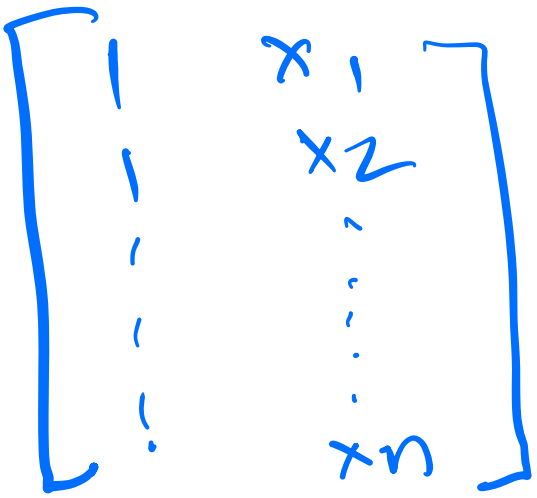
Note: Each row is a constraint.

(α_1, α_2) → is a tuple of parameters

We want to find optimal

$(\alpha_1, \alpha_2) = \theta$

Page 5

In machine learning, the matrix  is called

the feature matrix where each column is a feature.

This is also the design matrix. We will call it A.

$\therefore \quad y = A\theta \rightarrow$ linear problem

Page 6

Now, we could had written this as:

$$\forall n, \quad y_n = \alpha_1 x_n + \alpha_2 x_n^2 + \alpha_3$$

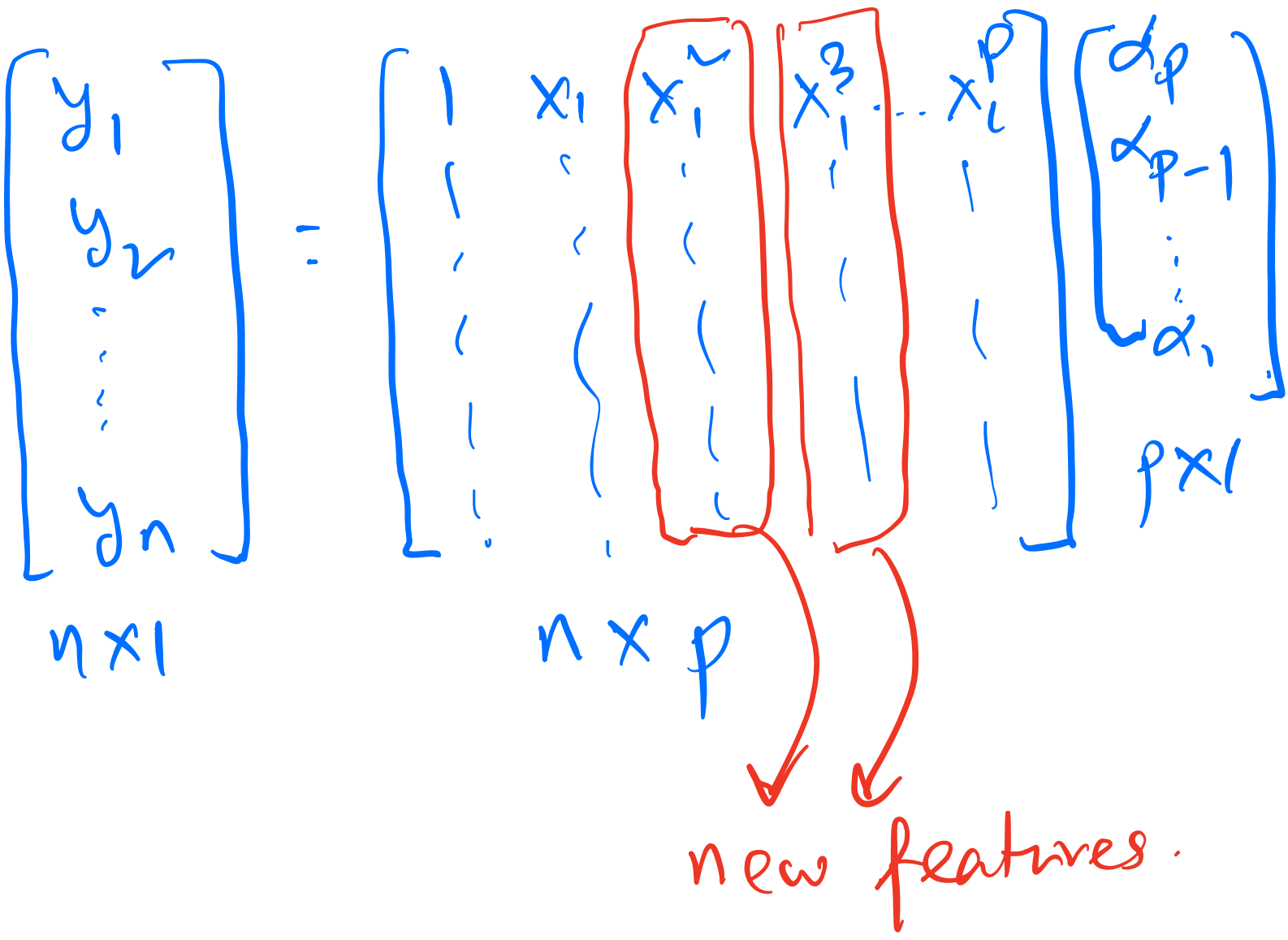
Then:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{n \times 1} = \underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}}_{n \times 3} \underbrace{\begin{bmatrix} \alpha_3 \\ \alpha_2 \\ \alpha_1 \end{bmatrix}}_{3 \times 1}$$

Page 7

We can keep adding more columns like thes, $x_n^2, x_n^3 \ldots$

e^{x_n}, $\log(x_n)$ ....



Page 8

This is still a linear model.

Because it is linear in θ

Note: Does this remind you of the basis functions in Sin D y?

So, every linear model is

$$\underset{n\times 1}{y} = \underset{n\times p}{A}\ \underset{p\times 1}{\theta}$$

We can have $n > p$ or $n < p$.

Over-determined

When $n \gg p$, we have a tall skinny matrix. In such a situation, "generally", we have no unique solution. This is because we have more equations than unknowns. So, instead, we search for a least square solution!

$$L(\theta) = \|A\theta - y\|_2^2$$

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\ L(\theta)$$

In the under-determined case, where $n << p$, we have a short & fat matrix. There are infinitely many solutions to $A\theta = y$. We find the solution, where $\|\theta\|_2$ is minimized.

Role of SVD in computing θ^*

$A = U \Sigma V^*$. We can invert each of these matrices to compute the moore-penrose inverse of rectangular A.

$$A^{-1} \equiv A^{+} = V \Sigma^{-1} U^{*}$$

$$A^{+} A = I_{(m \times m)}$$

$$A^{+} A \theta = A^{+} y \Rightarrow$$

$$\theta = \boxed{V \Sigma^{-1} U^{*} y.}$$

Another way of solving this:

$$y = A\theta \quad \text{for} \quad n \gg p$$

The next few pages assumes some knowledge on Random Matrix Theory

$$y A^T = (A^T A)\,\theta$$

$$\theta^* = \underbrace{(A^T A)^{-1}}_{P \times P}\,(A^T y)$$

→ normal equation

Suppose $y_n = g(x_n)$ is an actual linear process with noise:

$$[y_n] = A\hat{\theta} + \eta \quad \rightarrow \eta \sim N(0, \sigma^2 I)$$

Page 13

Then:

$$\|\hat{\theta} - \theta^*\|^2$$

$$= \|\hat{\theta} - (A^TA)^{-1}A^Ty\|^2$$

$$= \|\hat{\theta} - (A^TA)^{-1}A^T(A\theta^* + \eta)\|^2$$

$$= \|\hat{\theta} - \underbrace{(A^TA)^{-1}(A^TA)}_{I}\hat{\theta} - (A^TA)^{-1}A^T\eta\|^2$$

$$= \|(A^TA)^{-1}A^T\eta\|^2$$

$$= (A^T A)^{-1} A^T \eta \eta^T A (A^T A)^{-1}$$

We can re-write this as:

$$\|\hat{\theta} - \theta^*\|^2 = (\eta A (A^T A)^{-2} A^T \eta^T)$$

For multiple samples of $\eta \sim N(0, \sigma^2 I)$

$$E_N[\|\hat{\theta} - \theta^*\|^2]$$

$$= E_N[\eta A (A^T A)^{-2} A^T \eta^T]$$

Now, we will use two results from classical statistics:

$$E_x\left[x^T B x\right] = Tr(B)$$

↳ Trace

$$\eta \; x \sim N(0, I)$$

So: $E_N\left[\eta A (A^T A)^{-2} A^T \eta^T\right]$

$$= \sigma^2 Tr\left(A (A^T A)^{-2} A^T\right)$$

$$= \sigma^2 \, Tr\left((A^T A)^{-1}\right)$$

$$= \sigma^2 \frac{p}{n} tr\left(\left(\frac{A^T A}{n}\right)^{-1}\right)$$

tr : normalized trace

$$= \frac{Tr}{n}$$

So we have:

$$E_N\left[\|\hat{\theta} - \theta^*\|^2\right]$$

$$= \sigma^2 \frac{p}{n} \operatorname{tr}\left(\left(\frac{1}{n} A^T A\right)^{-1}\right)$$

At this point, we need a primer on random matrix theory

Wishart matrices: A matrix $\hat{\Sigma}$ is a wishart matrix if $\hat{\Sigma} = \frac{1}{n} X^T X$

where each row of the matrix X is drawn from a Gaussian

$$N(0, \Sigma_{p \times p})$$

For wishart matrices, the histogram of the eigenvalues of $\hat{\Sigma} = \frac{1}{n} X^T X$ converges to the

Marchenko-Pastur

distribution.

If $\frac{p}{n} \to \gamma \in [0, 1]$

as $n \to \infty$

$\psi_{MP}(t) \to$ pdf of marchenko-pastur

$$= \frac{1}{2\pi\gamma t} \sqrt{(\gamma_+ - t)(t - \gamma_-)}$$

on $[\gamma_-, \gamma_+]$

Page 21

$$\gamma_- = (1 - \sqrt{\gamma})^2$$

$$\gamma_+ = (1 + \sqrt{\gamma})^2$$

Stieltjes Transform!

For Wishart matrices, Z

$$S_n(z) = \sum \left[\mathrm{tr}\left[\left(\hat{\Sigma} - z I_p \right)^{-1} \right] \right]$$

Page 22

For the marchenko-pastur distribution, the Stieltjes transform is given as:

$$S_{MP}(z) = \int \frac{1}{(t-z)} \psi_{MP} \, dt$$

$$\forall \; z \in \mathbb{C} \setminus \mathbb{R}$$

Page 23

We come back to our estimated θ^* in linear regression

$$E_N\left[\|\hat{\theta} - \theta^*\|^2\right]$$

$$= \sigma^2 \frac{p}{n} \operatorname{tr}\left(\left(\frac{1}{n} A^T A\right)^{-1}\right)$$

If A is a standard Gaussian random matrix, then:

Page 24

$\hat{\Sigma} = \frac{1}{n} A^T A$ is a Wishart matrix, which as $n \to \infty$ converges to $\sigma^2 \gamma S(0)$

We know:

$$S(z) = \int \frac{1}{t - z} \psi_\mu \, dt$$

and

$$1 + z S(z) = \frac{S(z)}{1 + \gamma S(z)}$$

Page 25

of $z = 0$,

$$S(0) = \frac{1}{1-\gamma}$$

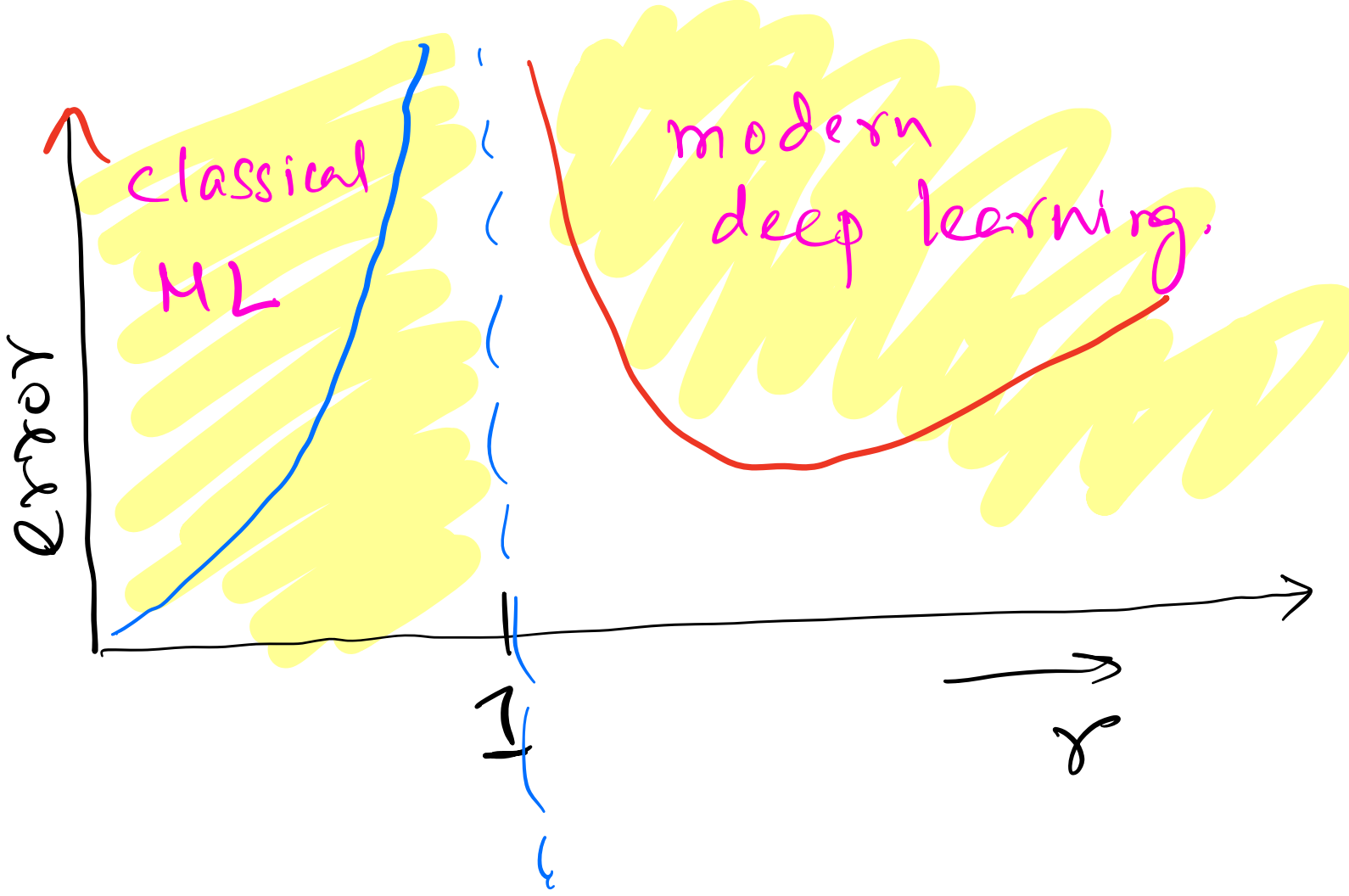
$n \gg p$ range

$$\therefore \quad E\left[\|\hat{\theta} - \theta^*\|^2\right] = \sigma^2 \frac{\gamma}{1-\gamma}$$

↳ A lot of analysis to prove that the risk in estimated parameters is a function of parameters vs samples (γ)

For $p >> n$, we have.

$$\mathbb{E}\left[\|\hat{\theta} - \theta^*\|^2\right] = \|\hat{\theta}\|^2 \left(1 - \frac{1}{\gamma}\right)$$



Page 27

Bias - Variance Trade-off

How do we design a model?

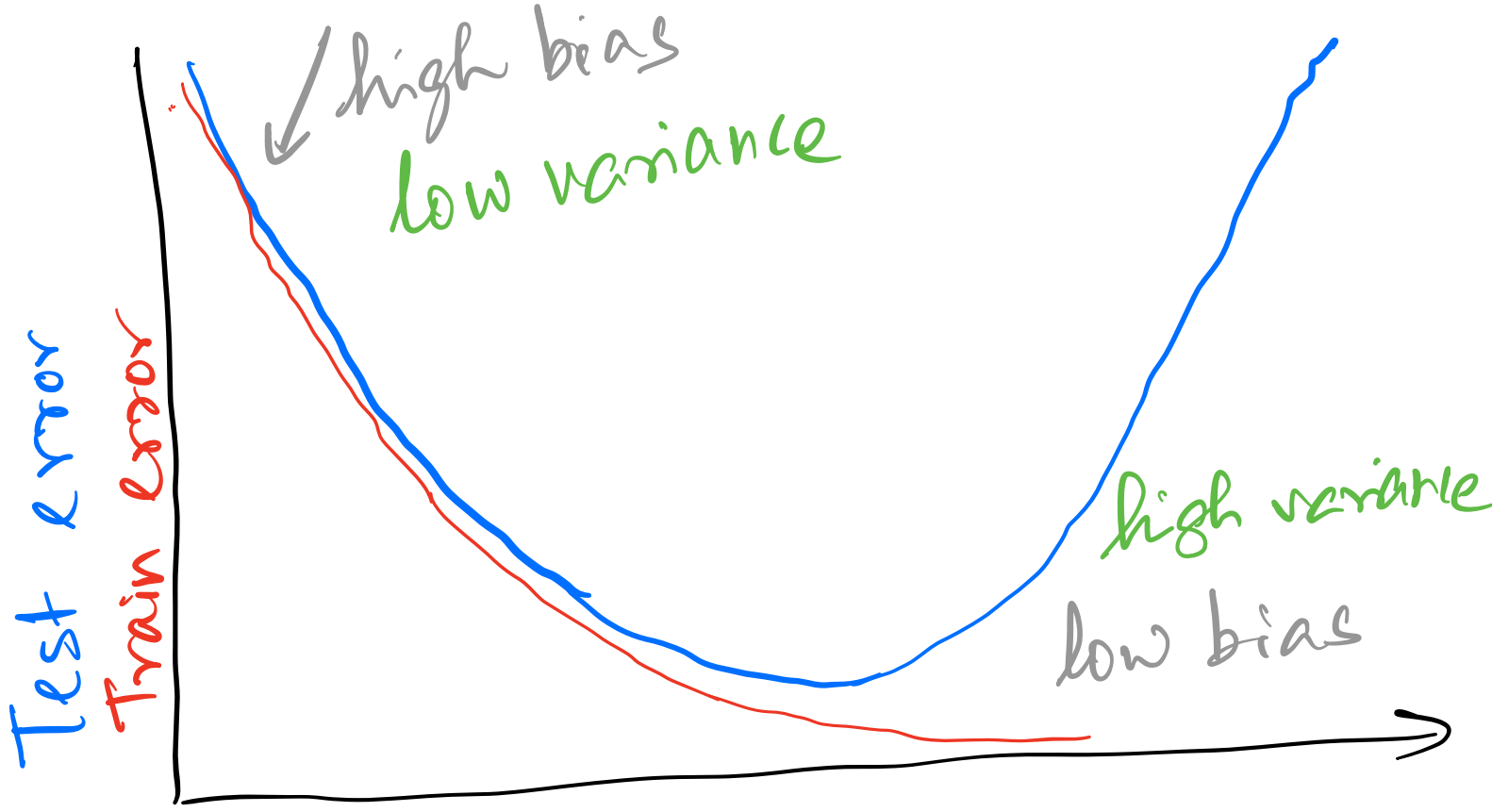
How many parameters should we have?

Classical Wisdom:

- Start with a small model
- Keep increasing the model as test error decreases
- Stop, when the test error increases again.

Page 28

See the slide deck to understand visually how bias decreases and variance increases



Page 29

We will prove that:

$$\text{MSE error} = (\text{bias})^2 + \text{variance} + \text{irreducible error}$$