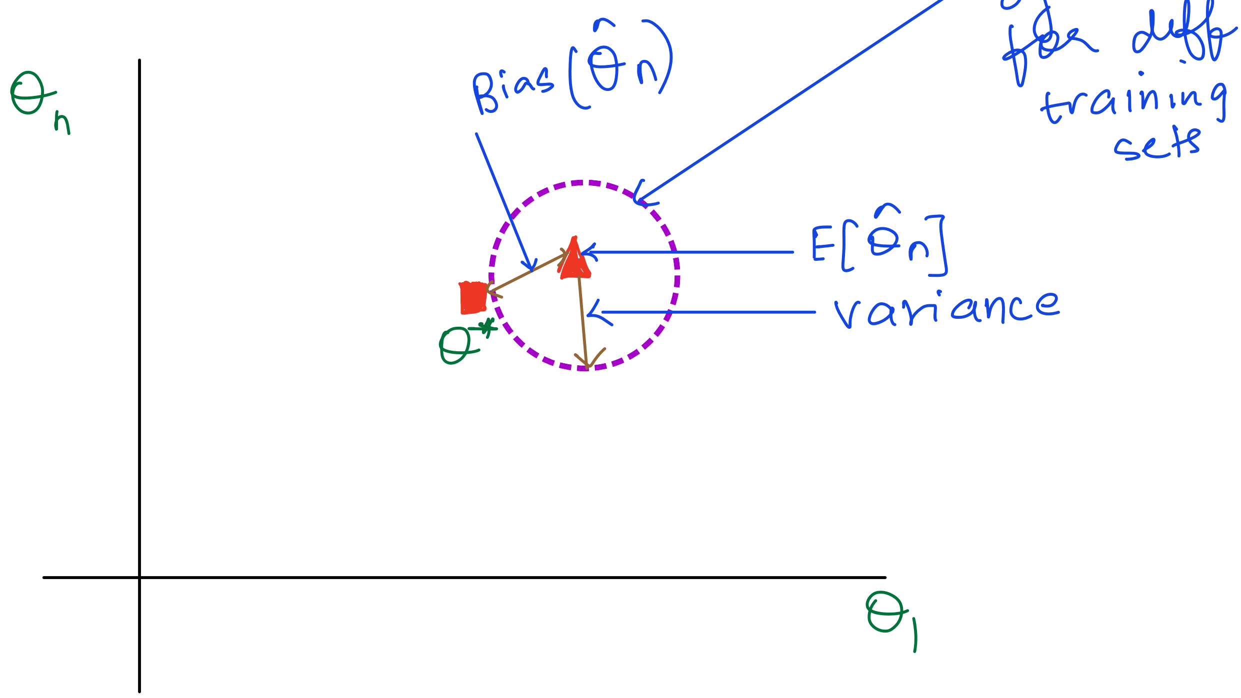
Statistical Learning Uniform Convergence

$$Bias(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n - \theta^*]$$

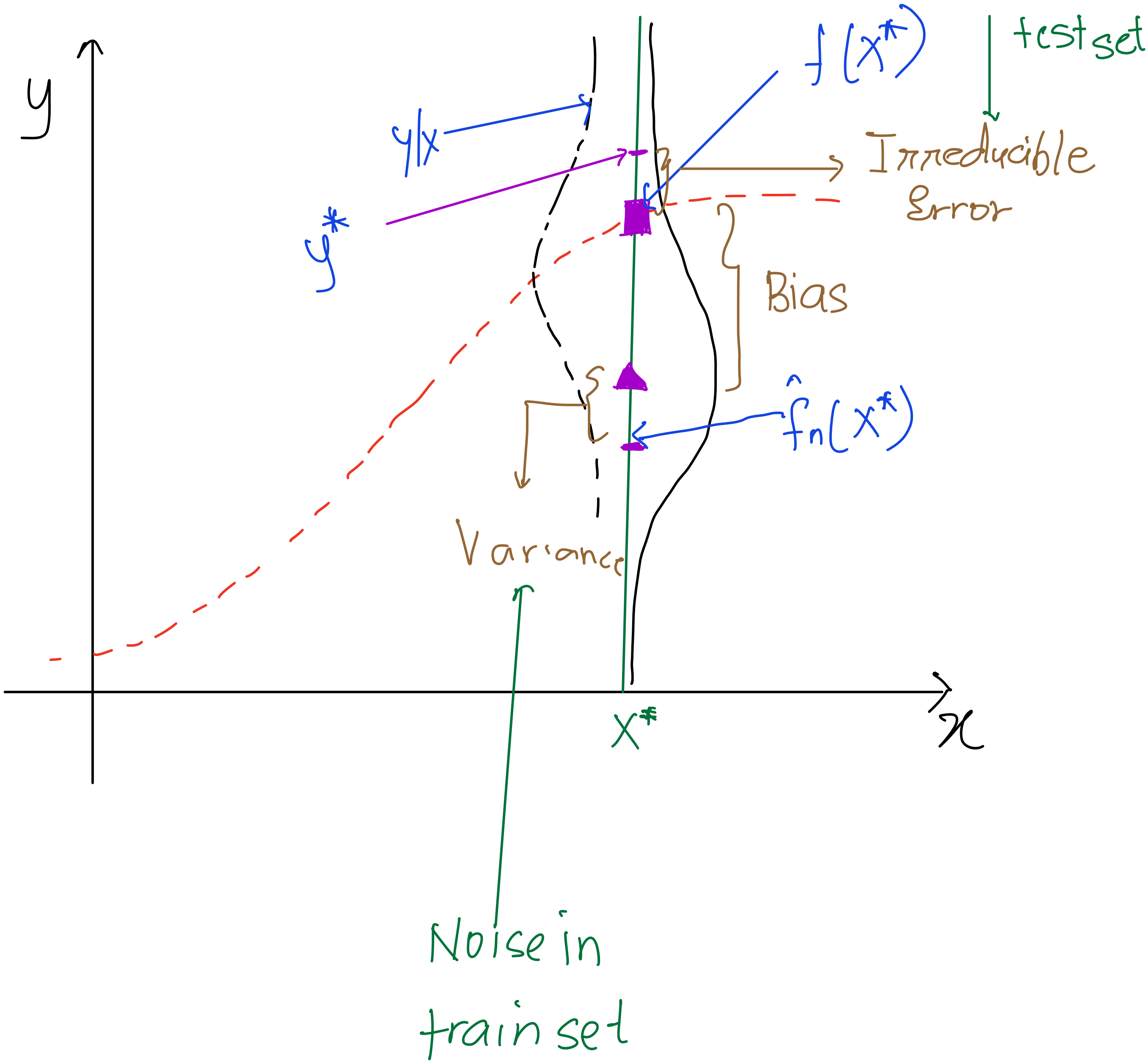


$$Bias(\hat{f}_n) = E[\hat{f}_n(x^*) - f(x^*)]$$

where $f = E[y|x]$, $y = f(x) + \epsilon$

$$Var(\hat{f}_n) = V[\hat{f}_n(x^*)]$$

Noise in



y
x
y|x
$f(x^*)$
test set
y^*
Irreducible Error
Bias
$\hat{f}_n(x^*)$
Variance
x^*
Noise in train set

Regularization

Equivalent to MAP

estimation $J(\theta) = \|X\theta - \vec{y}\|_2^2 + \lambda\|\theta\|^2$

$S \rightarrow$ training set $\{(x^{(1)}, y^{(1)}), \ldots (x^{(i)}, y^{(i)}) \ldots\}$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \underbrace{p(S|\theta)}_{\text{Likelihood}} \underbrace{p(\theta)}_{\text{Regularizer}}$$

MAP - Maximum a posteriori estimate
Instead of calculation the whole posterior, calculate mode of posterior & use that point estimate as output of estimator

For linear Regression,

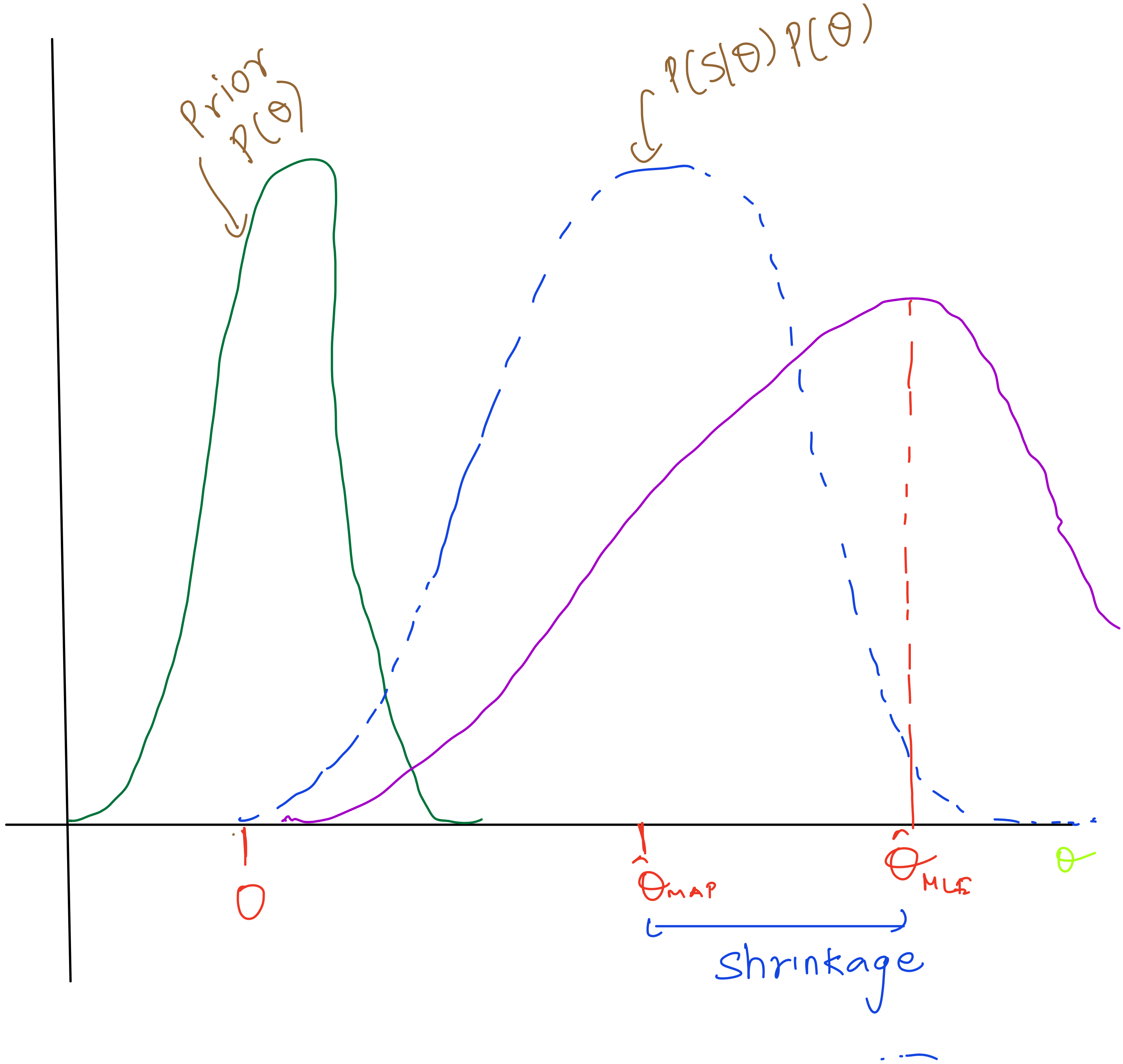
$p(S|\theta) \rightarrow$ Likelihood is some scalar multiple of $\|X\theta - y\|_2^2$

For Gaussian priors, prior takes form of squared error on θ and the λ is directly related to the variance of the prior

$$P(\theta|S) = \frac{P(S|\theta)\,P(\theta)}{P(S)}$$

$$\arg\max_{\theta} P(\theta|S) = \arg\max_{\theta} \frac{P(S|\theta)\,P(\theta)}{P(S)}$$

$$= \arg\max_{\theta} P(S|\theta)\,P(\theta)$$



Prior
P(θ)
P(S|θ) P(θ)
θ
θ̂MAP
θ̂MLE
θ
Shrinkage

L2 Regularized Linear Regression

$$J(\theta) = \left(\sum_{i=1}^{n} \left(y^{(i)} - \theta^T x^{(i)}\right)^2\right) + \lambda \|\theta\|_2^2$$

$$\hat{\theta}_n = \left(X^T X + \lambda I\right)^{-1} X^T \vec{y}$$

$$X^T X + \lambda I = U \begin{bmatrix} \sigma_1^2 + \lambda & & \\ & \ddots & \\ & & \sigma^2 + \lambda \end{bmatrix} U^T$$

$$\left(X^T X + \lambda I\right)^{-1} = U \begin{bmatrix} (\sigma_1^2 + \lambda)^{-1} & & \\ & \ddots & \\ & & (\sigma_d^2 + \lambda)^{-1} \end{bmatrix} U^T$$

$$U U^T = I = U^T U$$

$$E[\hat{\theta}_n] = \left[U \begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2+\lambda} & & 0 \\ & \ddots & \\ 0 & & \frac{\sigma_d^2}{\sigma_d^2+\lambda} \end{bmatrix} U^T \right] \theta^*$$

$$\hat{\theta}_n = (X^T X + \lambda I)^{-1} X^T (X\theta^* + \varepsilon)$$

Standard L.R. is unbiased
$E[\hat{\theta}_n] = \theta^*$

But when $\lambda > 0$, $\theta^* \neq E[\hat{\theta}_n]$ so bias added.

Also, see eigen values are all < 1, so they have a shrinkage effect

$$\mathrm{cov}(\hat{\theta}_n) = U \begin{bmatrix} \frac{\tau^2 \sigma_1^2}{(\sigma_1^2+\lambda)^2} & & \\ & \ddots & \\ & & \frac{\tau^2 \sigma_d^2}{(\sigma_d^2+\lambda)^2} \end{bmatrix} U^T$$

$$\varepsilon \sim N(0, \tau^2)$$

$$y = \theta^{*T} x + \varepsilon$$

$$\mathrm{Cov}(\hat{\theta}_n) = U \begin{bmatrix} \frac{\tau^2 \sigma_1^2}{(\sigma_1^2+\lambda)^2} & & \\ & \ddots & \\ & & \frac{\tau^2 \sigma_d^2}{(\sigma_d^2+\lambda)^2} \end{bmatrix} U^T$$

As $\lambda \uparrow$, variance reduces, bias increases.

$$MSE[\hat{f}_n] = \underbrace{\gamma^2}_{\text{Irreducible error}} + \underbrace{E[\hat{f}_n(x^*) - f(x^*)]^2}_{\text{Bias}^2} + \underbrace{V[\hat{f}_n(x^*)]}_{\text{variance}}$$

$$f(x) = \theta^{*T} x$$

$$\begin{aligned}
\text{Bias}(\hat{f}_n(x^*)) &= E[\hat{f}_n(x^*) - f(x^*)] \\
&= E[\hat{\theta}_n^T x^* - \theta^{*T} x^*] \\
&= E[\hat{\theta}_n - \theta^*]^T x^* \\
&= \text{Bias}(\hat{\theta}_n)^T x^*
\end{aligned}$$

$$\text{Var}[\hat{f}_n] = x^T \text{Cov}[\hat{\theta}_n] x$$

Heuristics for Bias & Variance

Training Error $\approx$ Bias

Cross Val. Error - Training Error $\approx$ Variance

To fight Bias

* Make model larger
* Reduce regularization

To fight Variance

* Collect more data
* Increase regularization

LEARNING THEORY

1. Train & Test Data ~ Same Distribution

2. Examples are sampled I.I.D.

Risk of Hypothesis

$$\mathcal{E}(h) = \mathbb{E}_{(x,y)\sim D}\left[\text{loss}(y, h(x))\right]$$

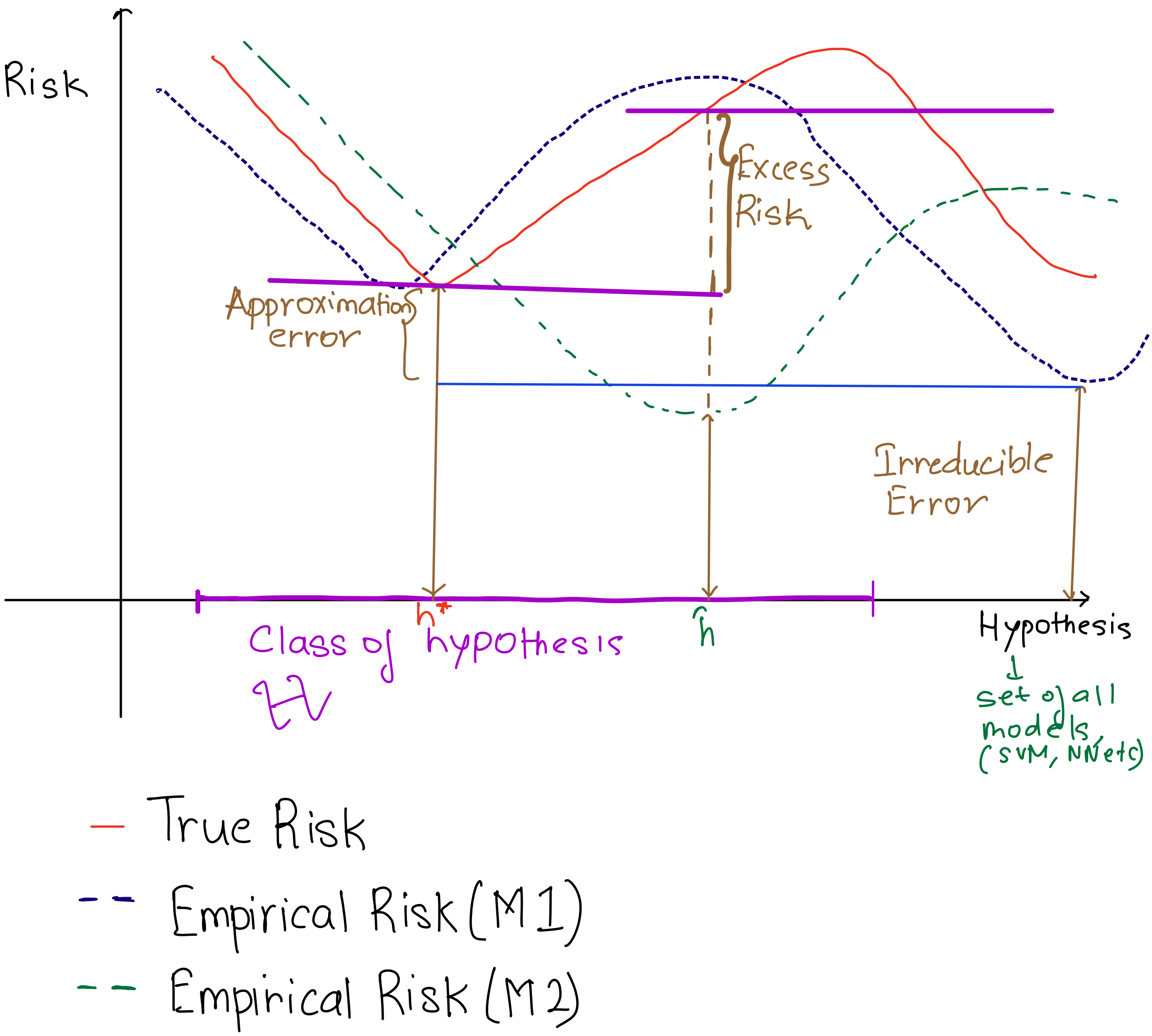
$$S = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$$

Empirical Risk

$$\hat{\mathcal{E}}(h) = \frac{1}{|S|}\sum_{(x,y)\in S}\text{loss}(y, h(x))$$

- For a given h, relation bet^n $\hat{\mathcal{E}}(h)$ & $\mathcal{E}(h)$?

- How does our G.E. compare to the best G.E.?



As you ↑ size of dataset, the dist^n bet^n curves decreases i.e. becomes tighter.

h^* Best in class hypothesis

Approximation Error - Penalty you pay by limiting to a class of models

Excess Risk - Penalty due to smaller dataset

Step 1 : Uniform Convergence

Step 2: Excess Risk Bound

Uniform Convergence : w.p. $\geq 1-\delta$, all $h \in \mathcal{H}$

$$\underbrace{|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma}_{\gamma \,=\, \text{Term}(n,\, \delta,\, \mathcal{H})}$$

Eg. Suppose $\delta = 0.1$, then we say

with 90% probability, the gap betⁿ true & empirical risk is less than $\sqrt{}$