CS725

Bayesian Parameter Estimation: Let there be a prior distribution over the weights $P(w)$. The posterior distribution over the weights $P(w|\mathcal{D})$ updates prior beliefs given observations or data $\mathcal{D}$.

Maximum Aposteriori (MAP) estimate: $\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\, P(\theta|\mathcal{D})$

$$= \underset{\theta}{\operatorname{argmax}} \log P(\mathcal{D}|\theta) + \log P(\theta)$$

Recall the coin example: $P(\mathcal{D}|\theta) = \theta^{n_H}(1-\theta)^{n_T}$

What is a good prior over θ?

CONJUGATE PRIOR: Let the likelihood $P(\mathcal{D}|\theta)$ come from a family of distributions d_1. Let the prior $P(\theta)$ come from a family of distributions d_2. The prior is said to be CONJUGATE if the posterior also comes from the same family d_2 as the prior.

Beta distribution is a conjugate prior for the binomial likelihood (Bernoulli)

Back to the coin example: $P(\mathcal{D}|\theta) = \theta^{n_H}(1-\theta)^{n_T}$

Prior: $P(\theta) = \mathrm{Beta}(\theta; \alpha, \beta) = \frac{1}{C}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

Posterior: $P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)\,P(\theta)$

$$\propto \theta^{n_H+\alpha-1}(1-\theta)^{n_T+\beta-1}$$

MAP estimate of θ, $\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} \underbrace{(n_H+\alpha-1)\log\theta + (n_T+\beta-1)\log(1-\theta)}_{L(\theta)}$

Setting $\frac{\partial L(\theta)}{\partial \theta} = 0$ & solving for $\theta \Rightarrow$ $\boxed{\theta_{MAP} = \frac{n_H + \alpha - 1}{N + \alpha + \beta - 2}}$

Pseudo coin counts coming from the Beta prior

MAP for linear regression

Assume a Gaussian prior over w, i.e., $P(w) = \mathcal{N}\left(0, \frac{I}{\lambda}\right)$

Recall multivariate Gaussians: $\mathcal{N}(\mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$

$$P(w) = \mathcal{N}\left(0, \frac{I}{\lambda}\right) = \frac{1}{(2\pi)^{d/2} \frac{1}{\lambda^{d/2}}} \cdot \exp\left(-\frac{\lambda}{2} w^T w\right) = \left(\frac{\lambda}{2\pi}\right)^{d/2} \exp\left(-\frac{\lambda}{2} \|w\|_2^2\right)$$

MAP estimate for linear regression:

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} \log P(\mathcal{D}|\theta) + \log P(\theta)$$

$$= \underset{\theta}{\operatorname{argmax}} -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 - \frac{\lambda}{2}\|w\|_2^2$$

$$= \underset{\theta}{\operatorname{argmin}} \underbrace{\sum_{i}(y_i - w^T x_i)^2 + \lambda' \|w\|_2^2}_{L_2\text{-regularized or ridge regression}}$$

BIAS and VARIANCE of ESTIMATORS

How do we measure the goodness of a predictor function $h_w(x)$?

Test error or Generalization error is measured and comprises of:

(A) Bias

(B) Variance

(C) Noise (Irreducible or irrecoverable error)

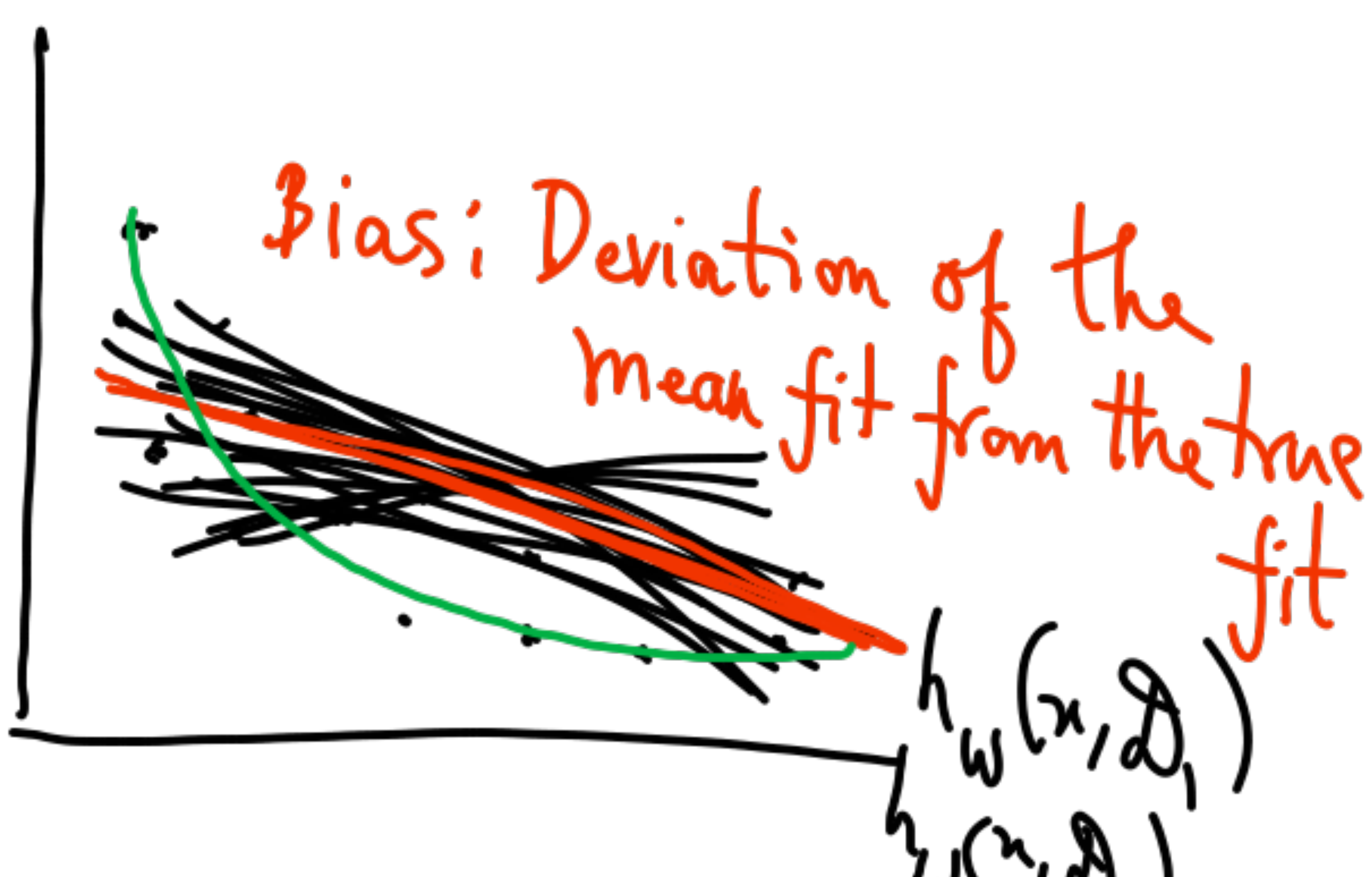
Let $y = f(x) + \epsilon,\ \epsilon \sim N(0, \sigma^2)$

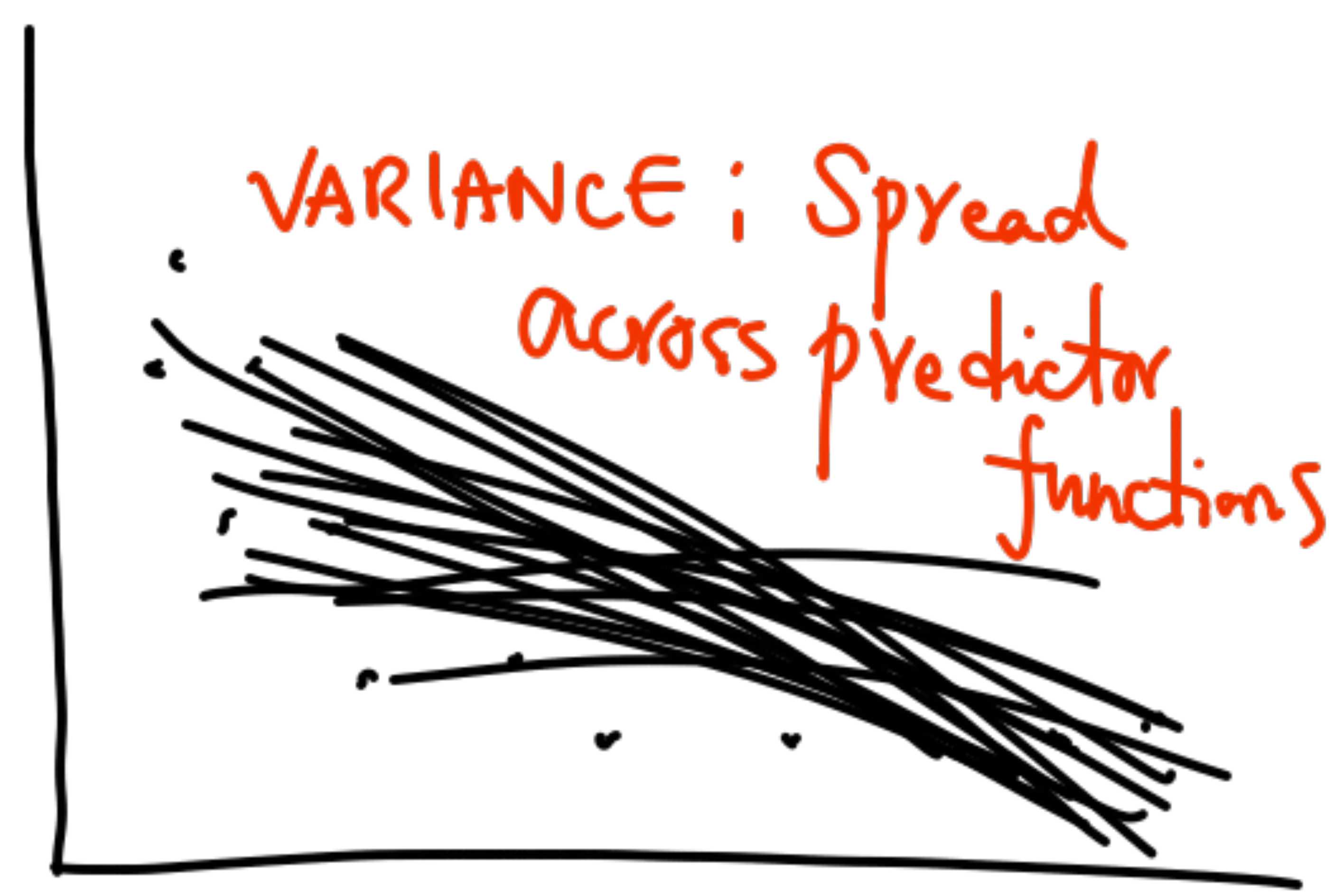
(A) BIAS can be defined as $\boxed{\underset{\mathcal{D}}{E}\left[h_w(x;\mathcal{D})\right] - f(x)}$

(B) VARIANCE can be defined as $\underset{\mathcal{D}}{E}\left[\left(h_w(x) - E\left[h_w(x)\right]\right)^2\right]$

(C) NOISE can be defined as $E\left[\left(y - f(x)\right)^2\right] = E\left[\epsilon^2\right]$
$= \sigma^2$

$\left[\because \sigma^2 = E\left[\left(\epsilon - E(\epsilon)\right)^2\right] = E(\epsilon^2) - \cancel{E(\epsilon)}^2\right]$





BIAS/VARIANCE Decomposition for Linear Regression

Given $y = f(x) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Let $\tilde{x}$ be a test point with $\tilde{y} = f(\tilde{x}) + \tilde{\epsilon}$, $\tilde{\epsilon} \sim \mathcal{N}(0, \sigma^2)$.

Then, the expected test error : $\underset{\mathcal{D},\tilde{\epsilon}}{E}\left[\left(\tilde{y} - h_w(\tilde{x}, \mathcal{D})\right)^2\right]$

$$E\left[\left(\tilde{y} - h_w(\tilde{x})\right)^2\right] = E\left[\tilde{y}^2 + h_w(\tilde{x})^2 - 2\tilde{y}\, h_w(\tilde{x})\right]$$

$$= E\left[\tilde{y}^2\right] + E\left[h_w(\tilde{x})^2\right] - 2\, E\left[\tilde{y}\right] E\left[h_w(\tilde{x})\right]$$

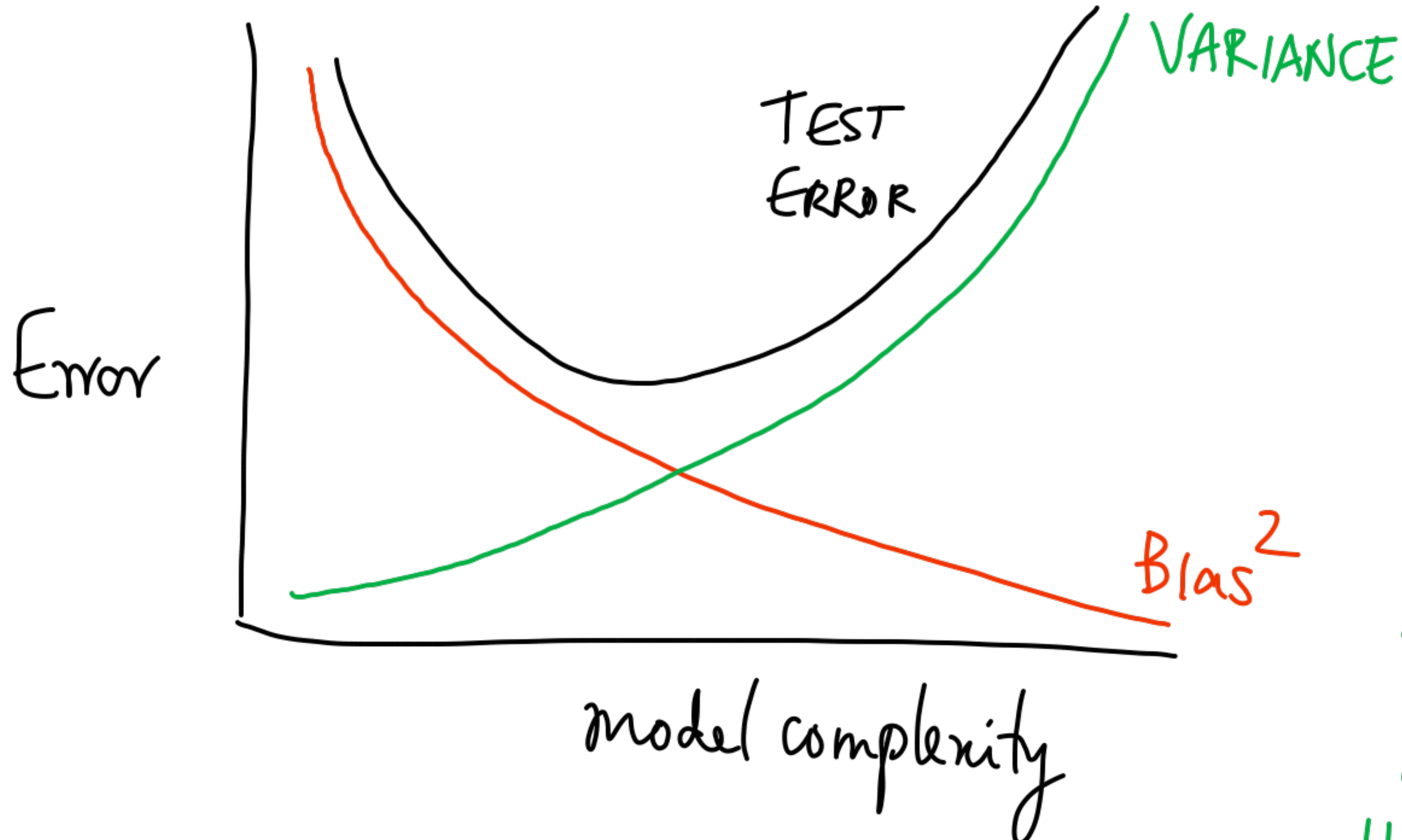
(Recall: $E[(X-E[X])^2] = E[X^2]-E[X]^2$)

$$E\left[\left(\tilde{y}-h_w(\tilde{x})\right)^2\right] = E\left[\tilde{y}^2\right] + E\left[h_w(\tilde{x})^2\right] - 2E[\tilde{y}]E\left[h_w(\tilde{x})\right]$$

$$= E\left[\left(\tilde{y}-E[\tilde{y}]\right)^2\right] + E[\tilde{y}]^2 + E\left[\left(h_w(\tilde{x})-E[h_w(\tilde{x})]\right)^2\right] + E\left[h_w(\tilde{x})\right]^2 - 2E[\tilde{y}]E\left[h_w(\tilde{x})\right]$$

$$= \underbrace{\left(E\left[h_w(\tilde{x})\right]-E[\tilde{y}]\right)^2}_{\text{BIAS}^2} + \underbrace{E\left[\left(h_w(\tilde{x})-E[h_w(\tilde{x})]\right)^2\right]}_{\text{VARIANCE}} + \underbrace{E\left[\left(\tilde{y}-f(\tilde{x})\right)^2\right]}_{\text{NOISE}}$$

$$\boxed{E[\tilde{y}] = E\left[f(\tilde{x})+\tilde{\epsilon}\right] = f(\tilde{x})}$$



High model complexity:
Low bias, high variance

Low model complexity:
High bias, low variance

MAP Problem

Suppose you are given n samples $x_1, \dots, x_n$ drawn i.i.d. from an exponential distribution given by $g(x|\theta) = \theta \exp(-\theta x)$ for $x \geq 0$

(A) What is MLE of θ?

(B) Which of the foll density fns gives a conjugate prior for the exponential likelihood?

Beta dist: $p(x; \alpha, \beta) = K_1 \, x^{\alpha-1} (1-x)^{\beta-1}$

Gamma dist: $P(x; \alpha, \beta) = K_2 \exp(-\beta x) \, x^{\alpha-1}$

Inverse gamma dist: $p(x; \alpha, \beta) = K_3 \exp(-\beta/x) \, x^{-\alpha-1}$

K_1, K_2, K_3 are constants

(C) What is the MAP estimate of θ?