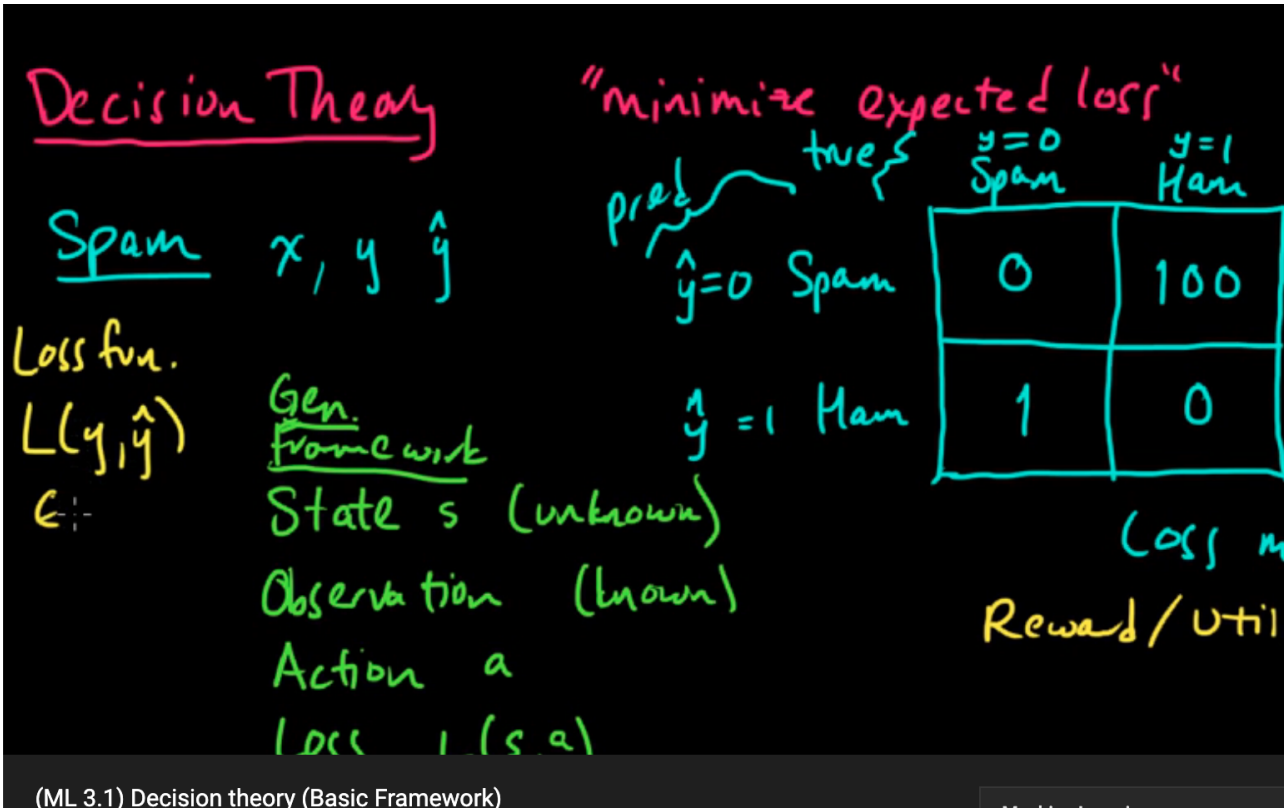
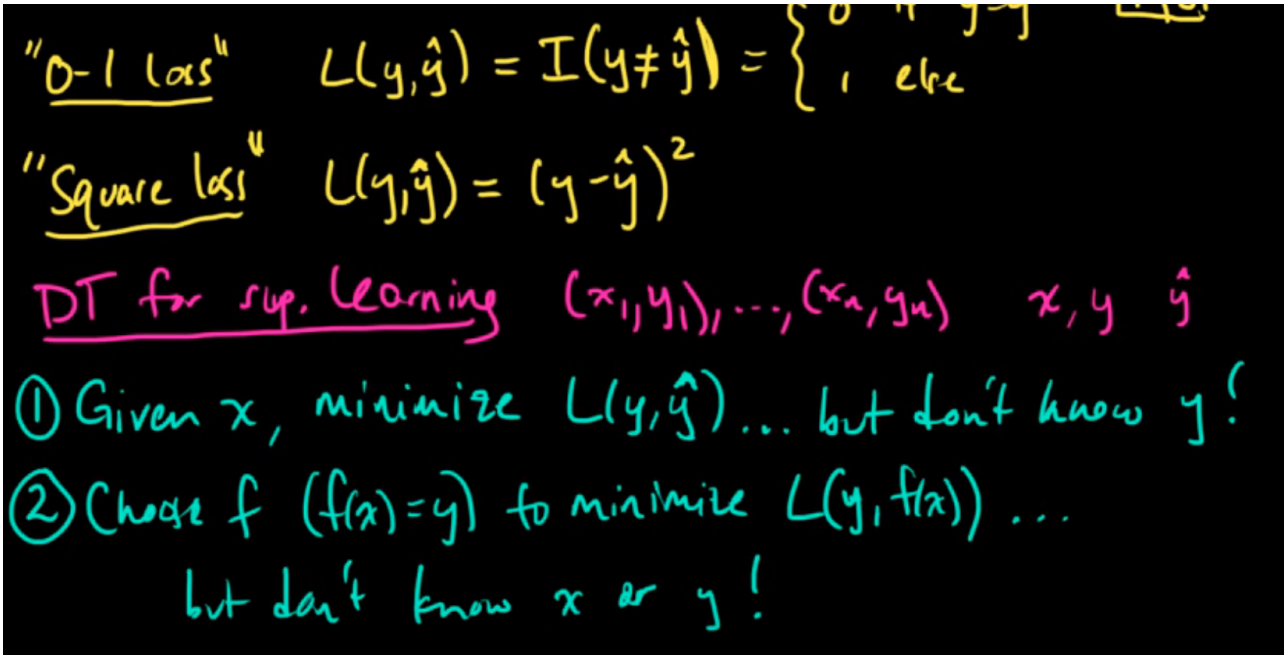
3 Decision Theory

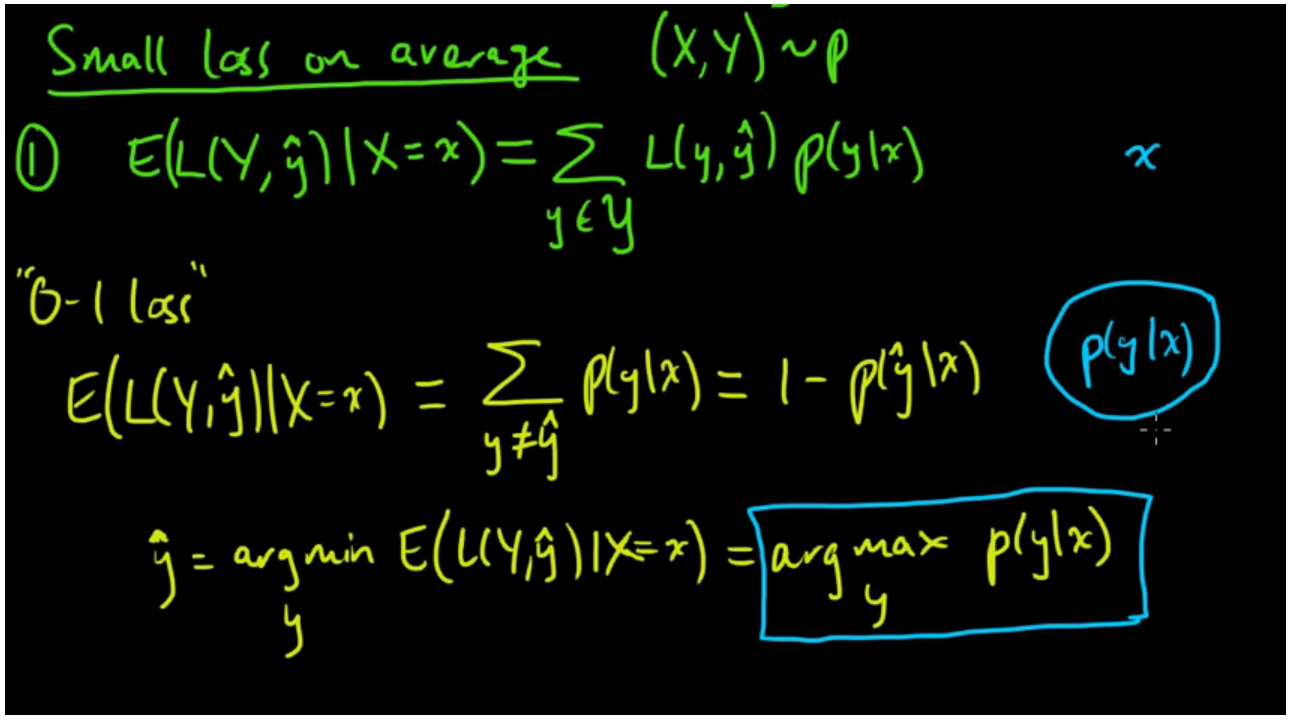
Friday, October 4, 2019 4:29 PM



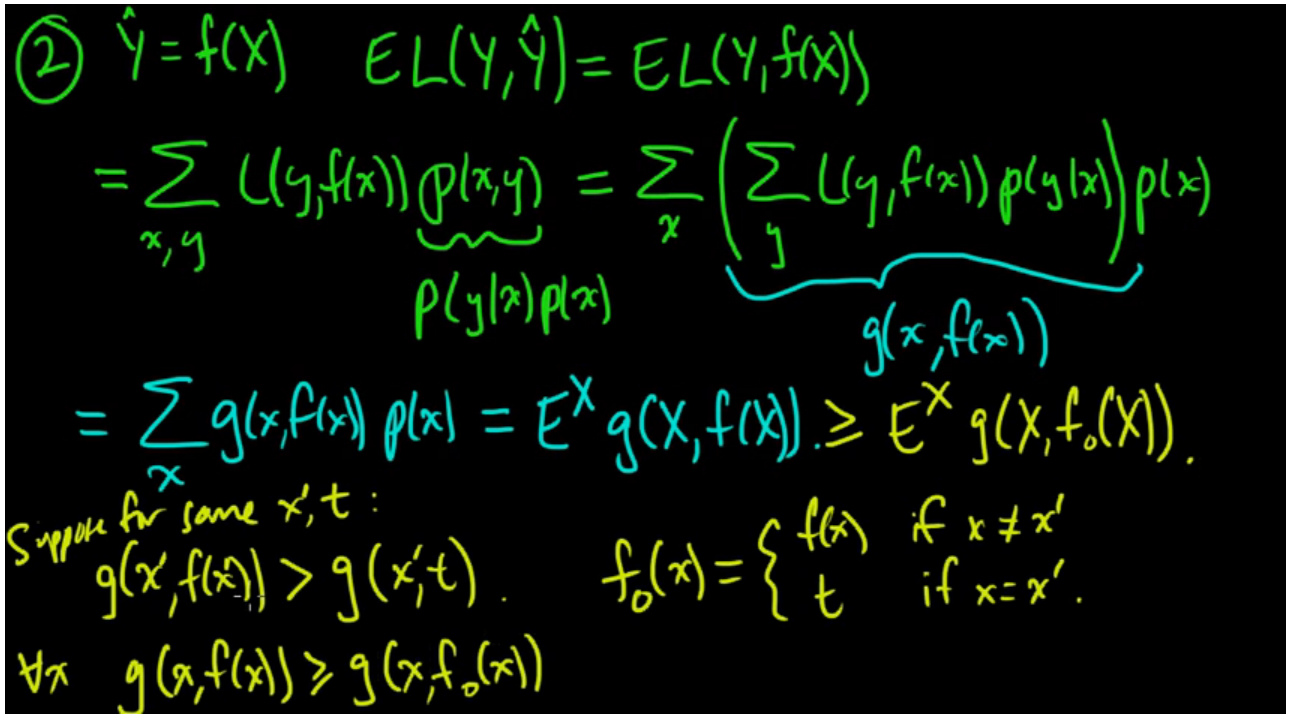


Decision theory as it applies to Supervised Learning

3.2 Minimizing conditional expected loss



3.3 Choosing f to minimize expected loss



$p(y|x)p(x)$ $\qquad$ $g(x, f(x))$

$$= \sum_x g(x, f(x))\, p(x) = E^X g(X, f(X)) \geq E^X g(X, f_0(X)).$$

Suppose for some x', t:

$$g(x', f(x')) > g(x', t). \qquad f_0(x) = \begin{cases} f(x) & \text{if } x \neq x' \\ t & \text{if } x = x'. \end{cases}$$

$\forall x \quad g(x, f(x)) \geq g(x, f_0(x))$ $\qquad$ Choose f to min. $g(x, f(x))$.

Define: $f^*(x) = \arg\min_t g(x, t)$. $\qquad \geq E^X g(X, f^*(X))$

---

② $\hat{Y} = f(X) \qquad EL(Y, \hat{Y}) = EL(Y, f(X))$

$$= \sum_{x,y} L(y, f(x))\, p(x, y) = \sum_x \left(\sum_y L(y, f(x))\, p(y|x) \right) p(x)$$

$p(y|x)p(x)$ $\qquad$ $g(x, f(x))$

$$= \sum_x g(x, f(x))\, p(x) = E^X g(X, f(X)) \geq E^X g(X, f_0(X)).$$

Suppose for some x', t:

$$g(x', f(x')) > g(x', t). \qquad f_0(x) = \begin{cases} f(x) & \text{if } x \neq x' \\ t & \text{if } x = x'. \end{cases}$$

$\forall x \quad g(x, f(x)) \geq g(x, f_0(x))$ $\qquad$ Choose f to min. $g(x, f(x))$.

Define: $f^*(x) = \arg\min_t g(x, t)$. $\qquad \geq E^X g(X, f^*(X))$

$g(x, f(x))$

$$= \sum_x g(x, f(x))\, p(x) = E^X g(X, f(X)) \geq E^X g(X, f_0(X)).$$

Suppose for some x', t:

$$g(x', f(x')) > g(x', t). \qquad f_0(x) = \begin{cases} f(x) & \text{if } x \neq x' \\ t & \text{if } x = x'. \end{cases}$$

$\forall x \quad g(x, f(x)) \geq g(x, f_0(x))$ Choose f to min. $g(x, f(x))$.

Define: $f^*(x) = \arg\min_t g(x, t)$.

$\geq E^X g(X, f^*(X))$

$\Rightarrow$ $p(y|x)$ is the key quantity

$\forall x \quad g(x, f(x)) \geq g(x, f_0(x))$ Choose f to min. $g(x, f(x))$.

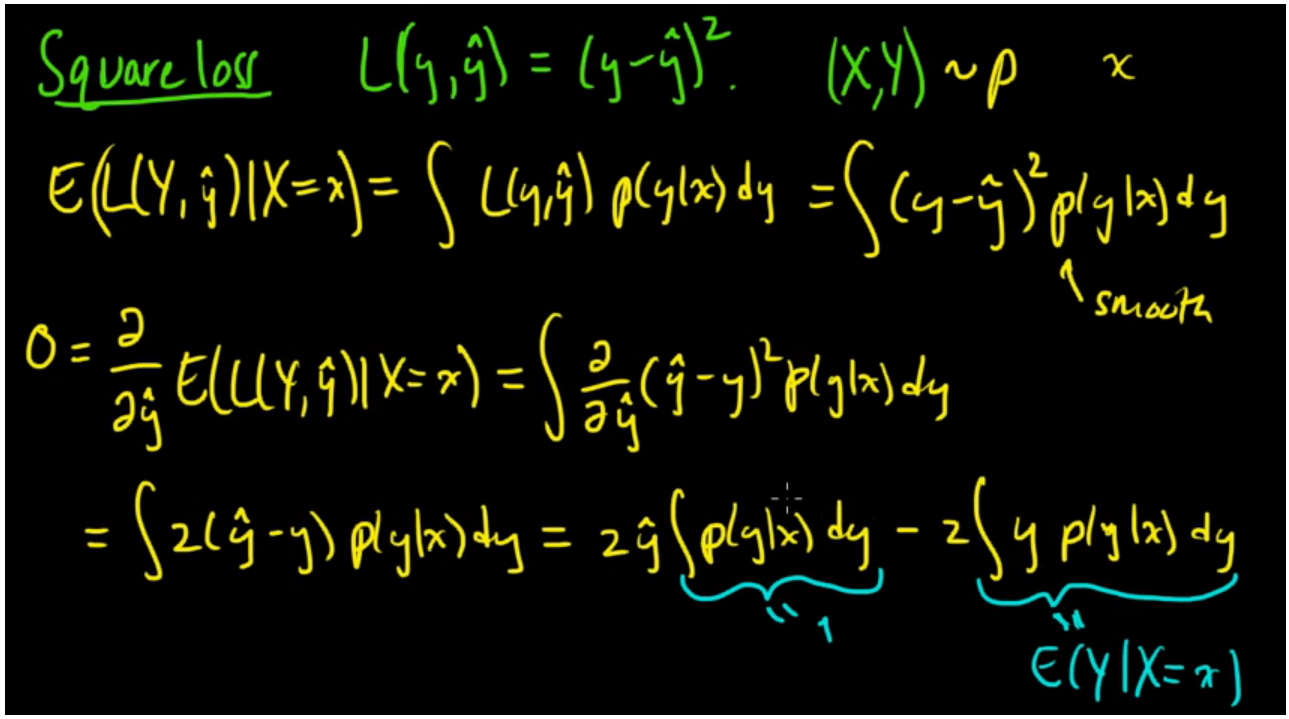
Define: $f^*(x) = \arg\min_t g(x, t)$.

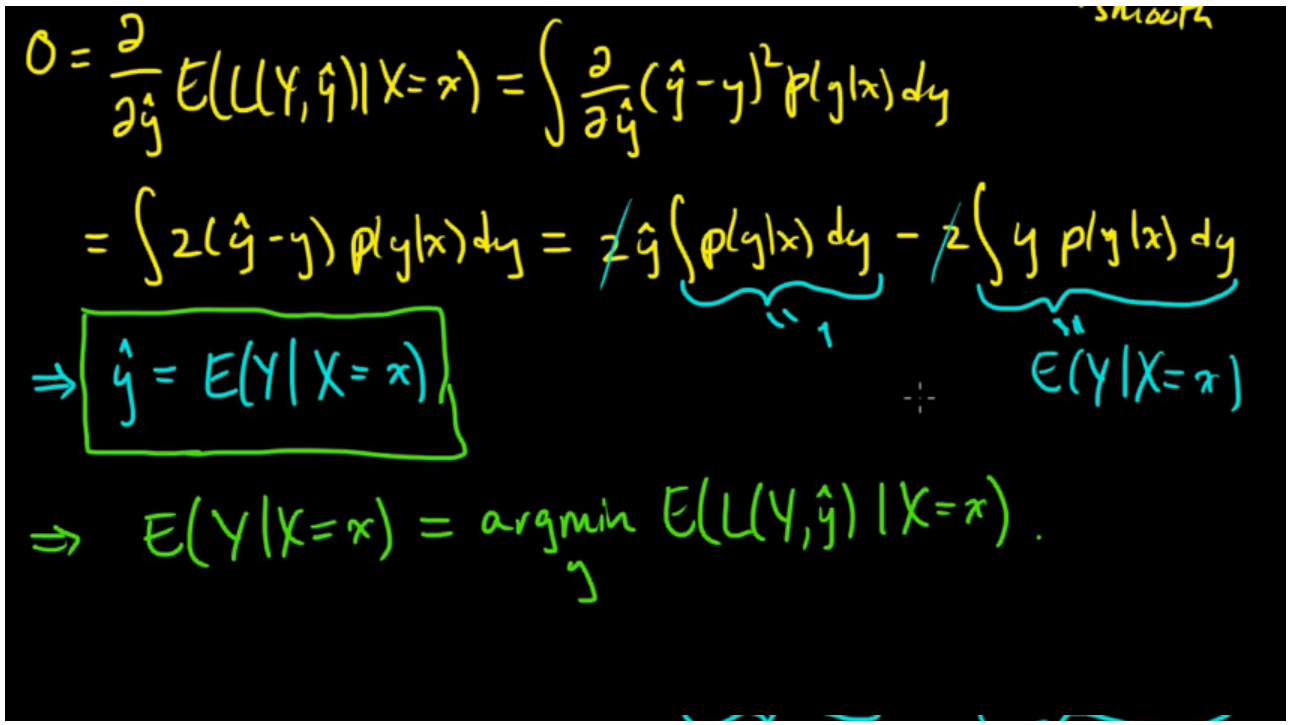
$\geq E^X g(X, f^*(X))$

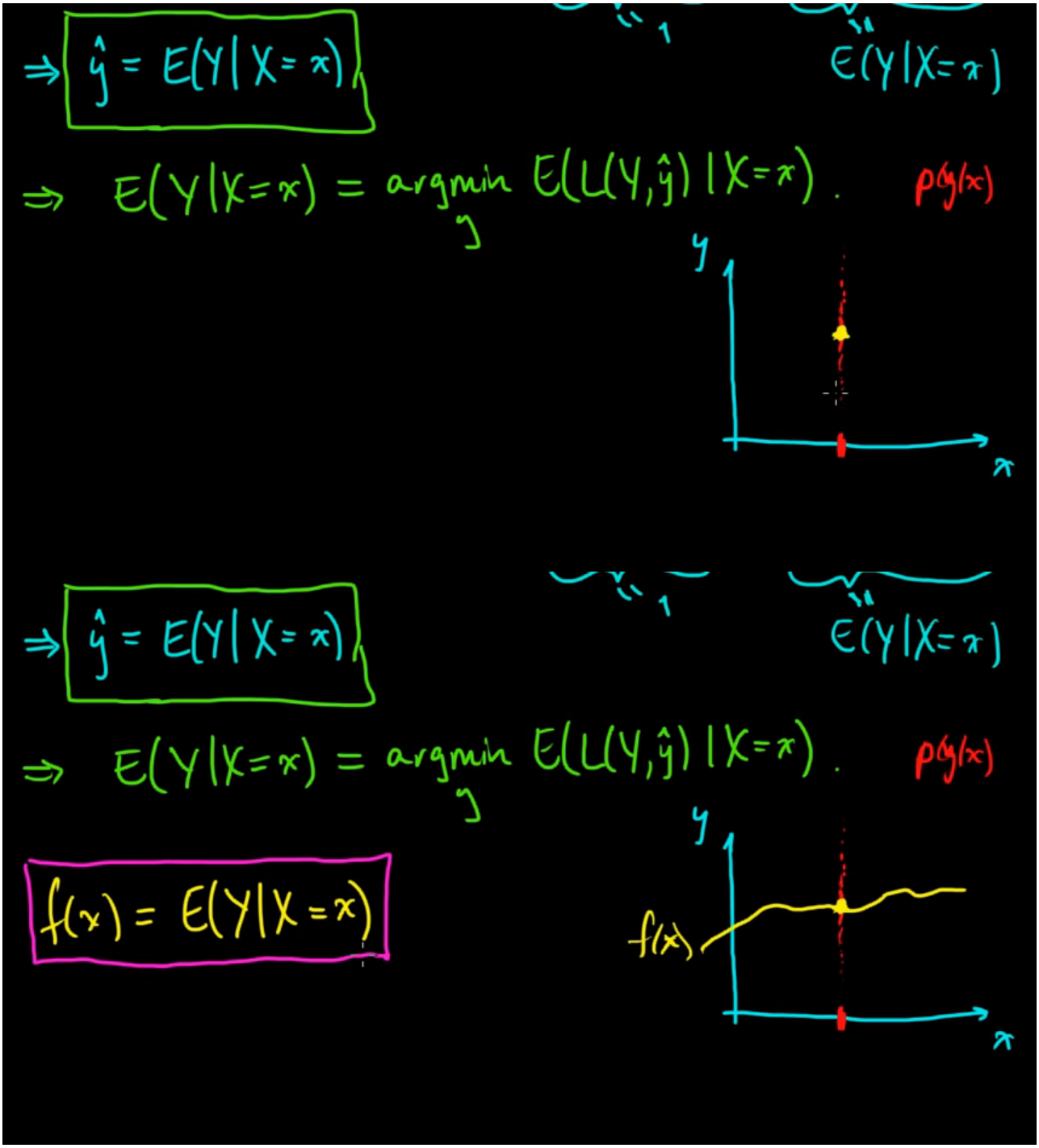
$\Rightarrow$ $p(y|x)$ is the key quantity

$E(Y, \hat{Y}) = E^X(E(Y, \hat{Y}|X))$ ← L.I.E.

3.4 Square loss





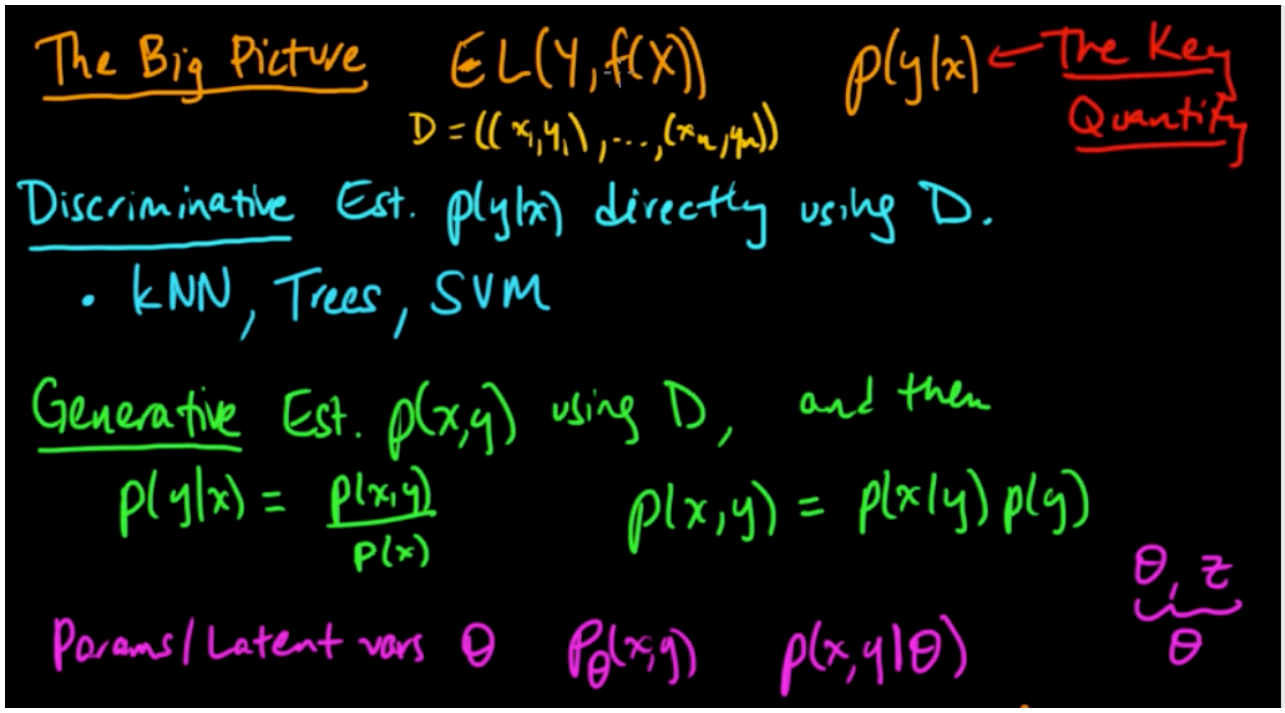


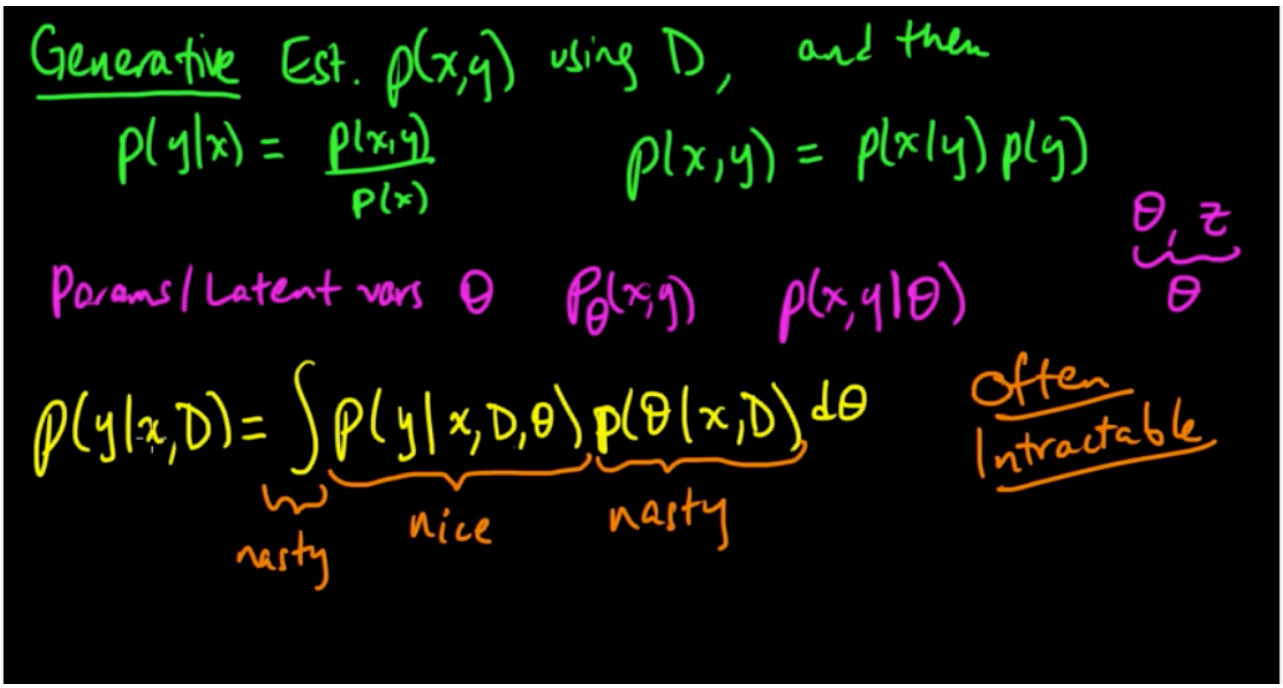
3.5 The Big Picture in machine learning (part1)

- focusing on supervised learning, but same concept for others
- By starting from the decision theoretic idea that we want to minimize the expected loss betweer value y and our prediction, f(x), how many of the core concepts and methods in ML fall out natu trying to solve this problem
- key quantity: p(y|x) that we needed to solve this minimization problem

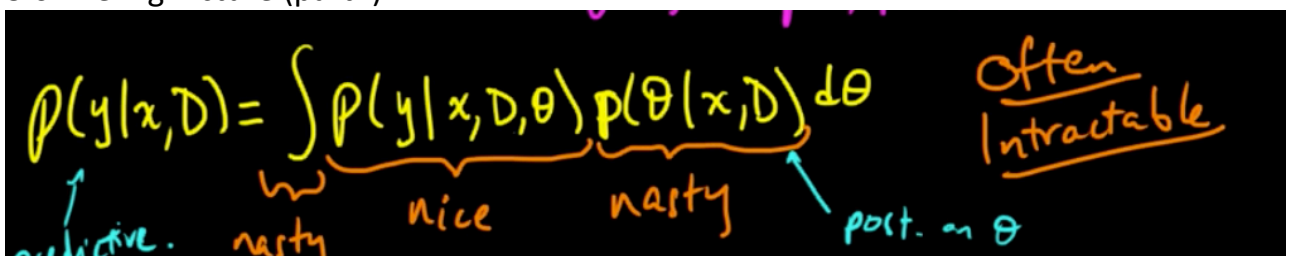
n the true
rally while

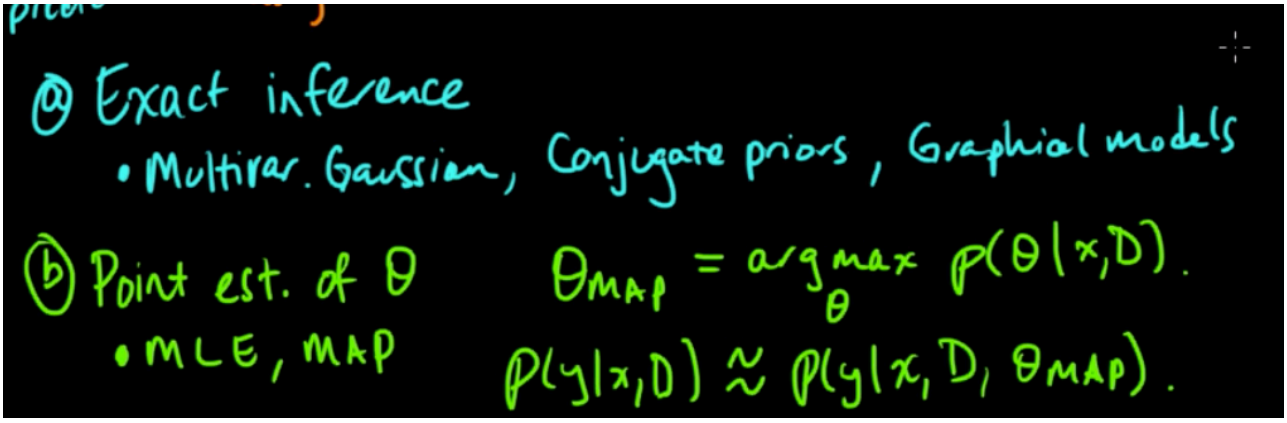
1. Discriminative

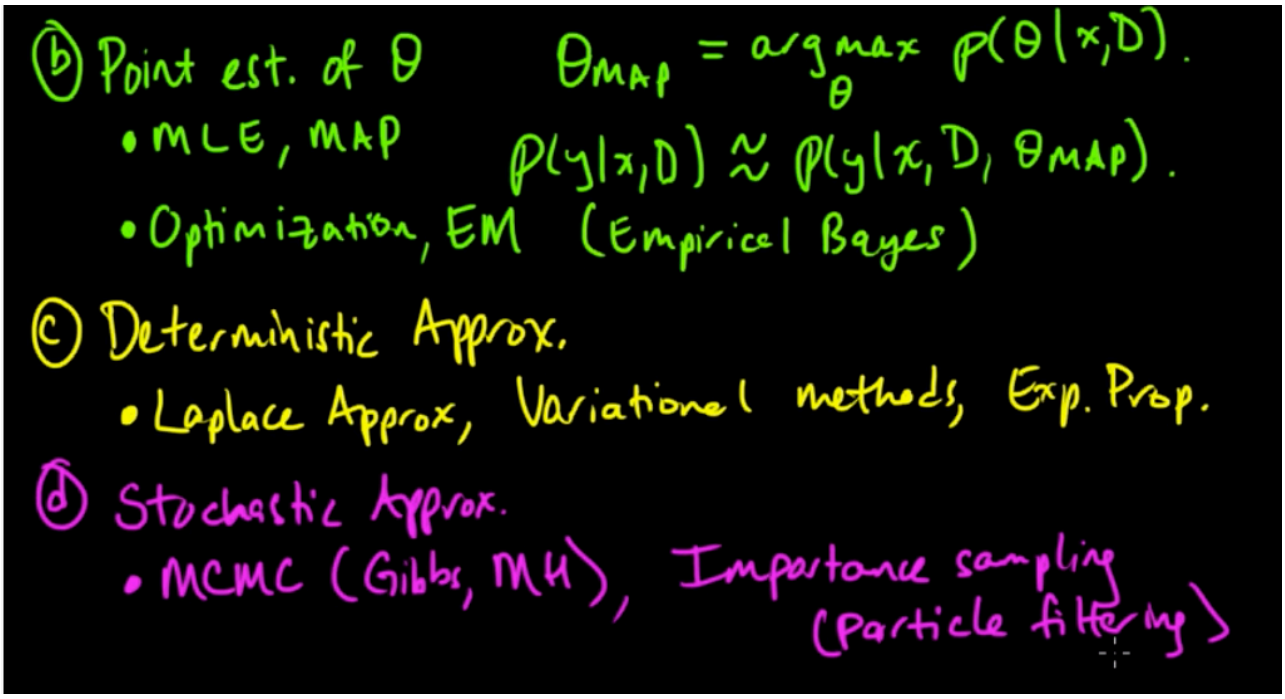




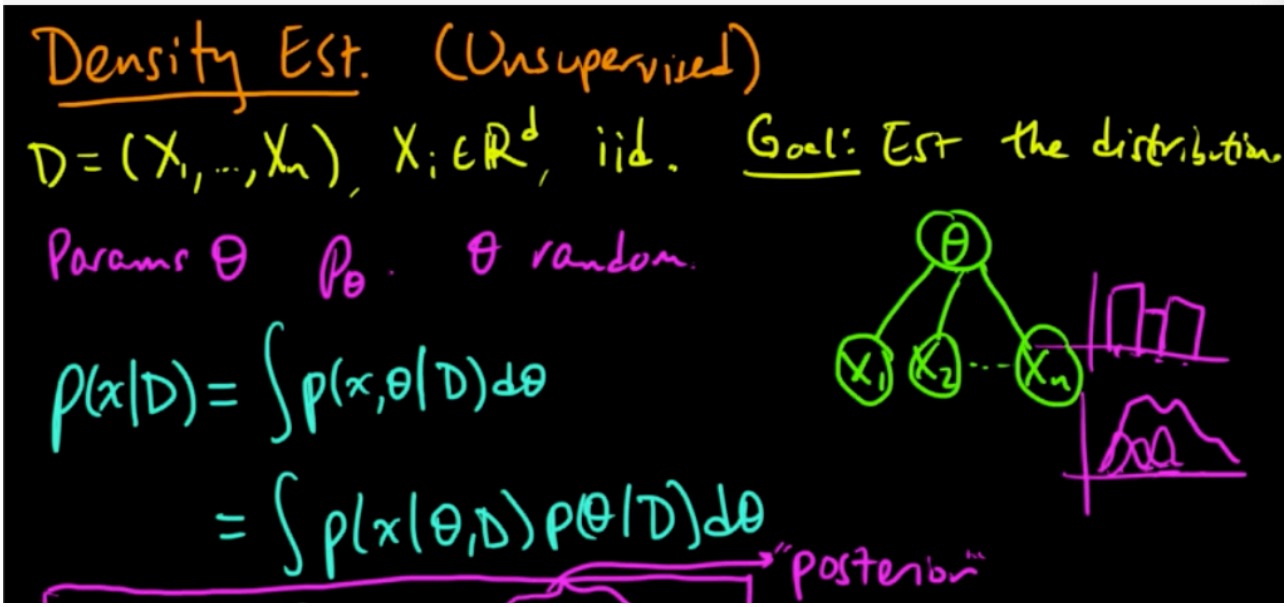
3.6 The Big Picture (part2)







3.7 The Big Picture (part3)



Importa

- Ge
- Let
- His
- Mo
  -
  -

nce sampling is to estimate the expected value of the predictive p. distribution, p(y|x, Data)

neralize the discussion so far (which was based on the supervised setting),
's take a look at the problem of density estimation (Unsupervised)
togram estimation, Kernel density estimation
re prob. approach: parameters parametrizing the distribution on X's

- Then, think of those parameters (thetha) as random var
- Then, suppose these data are generated from the model in this way
  - Pick a theta from p_theta
  - Draw X1, X2, ..., Xn from that sampled theta
  - Estimation the deistibution over X means, to specify "p(x|all data D)"

$$p(x|D) = \int p(x|\theta)\, p(\theta|D)\, d\theta$$

<--- Know

this because we choose the form of this p. distribution