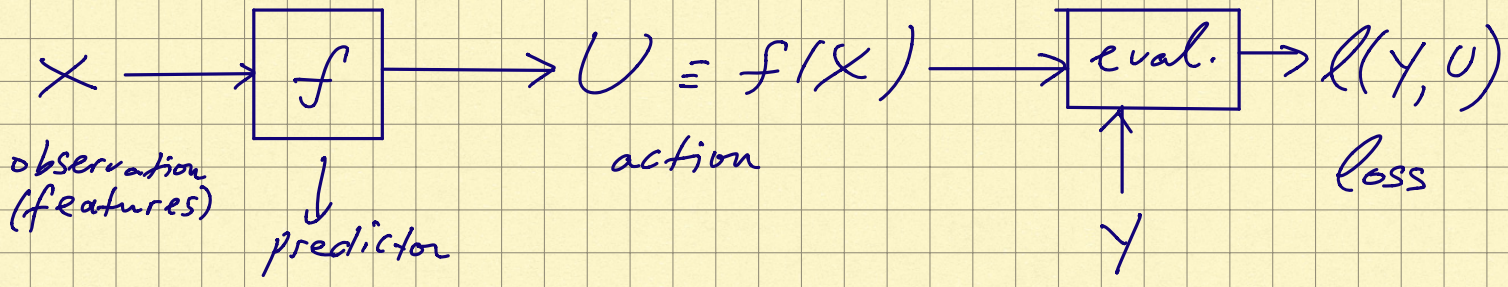
Recap: prediction, known distribution

$(X, Y) \sim P$



Loss (or risk) of f: $L_P(f) := \mathbb{E}_P[\ell(Y, f(X))]$

Optimality: $f_P^* = \arg\min_f L_P(f)$

Examples:

1) binary classification

Y, U take values in $\{0, 1\}$

$\ell(y, u) = 1_{\{y \neq u\}}$

$L_P^* = \min_f L_P(f) = \mathbb{E}_P[\min\{1 - \eta(X), \eta(X)\}]$

$f_P^*(x) = 1_{\{\eta(x) \geq 1/2\}}$

where $\eta(x) := \mathbb{P}[Y = 1 \mid X = x] = \mathbb{E}[Y \mid X = x]$

2) MMSE estimation

$X \in \mathbb{R}^d$; $Y, U \in \mathbb{R}$

$\ell(y, u) = (y - u)^2$

$L_P^* = \mathbb{E}_P(Y - \underbrace{\mathbb{E}[Y \mid X]}_{f_P^*(X)})^2$

3) "soft" binary classification

$Y \in \{0,1\}$

$U \in [0,1]$ $\quad \ell(y,u) = |y-u|$

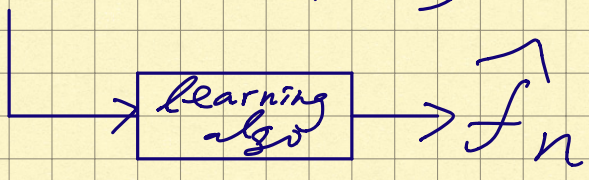
Takeaway message: if P (dist. of (X,Y)) is known, no learning takes place

Learning: P unknown

$(X_1,Y_1),\dots,(X_n,Y_n) \overset{iid}{\sim} P$ – data

$Z = (X,Y)$

$Z_1,\dots,Z_n \overset{iid}{\sim} P$ $\quad Z^n := (Z_1,\dots,Z_n)$



$\hat{f}_n$ is a random function from features X into actions U

$$L_P(\hat{f}_n) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y,\hat{f}_n(x))\, P(dx,dy)$$

$$= \mathbb{E}[\ell(Y,\hat{f}_n(X)) \mid Z^n]$$

where $(X,Y) \sim P$, indep. of Z^n

$\underbrace{Z_1,\dots,Z_n}_{\text{data}}, Z_{n+1} \overset{iid}{\sim} P$

$\Rightarrow L_P(\hat{f}_n) = \mathbb{E}[\ell(Y_{n+1},\hat{f}_n(X_{n+1})) \mid Z^n]$

Important: $L_P(\hat{f}_n)$ cannot be computed w/o knowledge of P!

$Z_1, \dots, Z_n \xrightarrow{LA} \hat{f}_n$

$Z_1, \dots, Z_n$ — training data

$Z'_1, \dots, Z'_n \overset{iid}{\sim} P, \; \perp Z^n$ (validation/test set)

$$\hat{L}_P(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y'_i, \hat{f}_n(X'_i))$$

Cond. on Z^n, $\hat{L}_P(\hat{f}_n) \approx L_P(\hat{f}_n)$

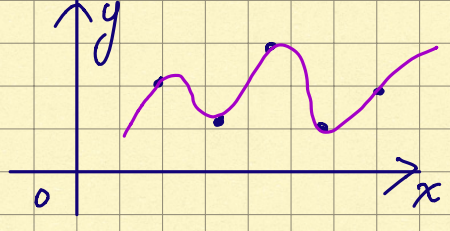
Goals: $L_P(\hat{f}_n)$ vs. $L_P^* \equiv \min_f L_P(f)$

$L_P(\hat{f}_n) \geq L_P^*$

- want to make excess loss $L_P(\hat{f}_n) - L_P^*$ small

Interpolation / memorization?

$\hat{f}_n(X_i) = Y_i \qquad i = 1, \dots, n$



overfitting: $\hat{f}_n$ interpolates but $L_P(\hat{f}_n) \gg L_P^*$

- inductive bias: learning algo is restricted to some structured class of predictors

$\hat{f}_n \in \mathcal{H}$ (hypothesis space)
- typically restricted in comp.

$$L_p(\hat{f}_n) - L_p^*(\mathcal{H}) =: \text{excess risk relative to } \mathcal{H}$$

where $L_p^*(\mathcal{H}) := \min_{f \in \mathcal{H}} L_p(f)$

$$L_p(\hat{f}_n) - L_p^*$$

$$= \underbrace{L_p(\hat{f}_n) - L_p^*(\mathcal{H})}_{\text{estimation error}} + \underbrace{L_p^*(\mathcal{H}) - L_p^*}_{\text{appx. error}}$$

Belkin-Rakhlin-Tsybakov (AISTATS 2019):
"Does interpolation contradict statistical optimality?"

Exist settings where $\hat{f}_n$ interpolates the data, but is still asymptotically consistent:

$$\mathbb{E}(\hat{f}_n(X) - Y)^2 - \min_{f \in \mathcal{H}} \mathbb{E}(f(X) - Y)^2 \to 0 \text{ as } n \to \infty$$

provided $\eta(x) = \mathbb{E}[Y | X = x] \in \mathcal{H}$.

Double descent (Belkin et al.)