data generating process (DGP)

$$\mathbb{P}(\underset{\substack{\text{matrix} \\ \text{r.v.}}}{X}, \underset{\substack{\text{vector} \\ \text{r.v.}}}{Y}) = \mathbb{P}(Y \mid X = \underset{\substack{\text{constant} \\ \text{matrix}}}{x})^{\sigma^2} \, \mathbb{P}(X) \qquad \mathbb{D} = \left(\boxed{X}, \boxed{Y} \right)$$

Imagine you get to see $\mathbb{D}_1, \mathbb{D}_2, \ldots, \mathbb{D}_M$ (iid from $\mathbb{P}(X, Y)$) where M is large.

Now you can fit $g_1 = \mathcal{A}(\mathbb{D}_1), \ldots, g_M = \mathcal{A}(\mathbb{D}_M)$. Then you can use the avg:

$$g_{Avg} = \frac{g_1 + \ldots + g_M}{M} \approx E[G]$$

$g_i = \mathcal{A}(\mathbb{D}_i)$ are iid

$$MSE = \sigma^2 + E_X\left[Bias[G_{avg}]^2\right] + E_X\left[Var[G_{avg}]\right]$$

$$= \sigma^2 + E_X\left[E\left[\left(\frac{g_1 + \ldots + g_M}{M} - f\right)^2\right]\right] + E_X\left[Var\left[\frac{g_1 + \ldots + g_M}{M}\right]\right]$$

$$= \sigma^2 + E_X\left[\left(\frac{1}{M}\left(\underbrace{(E[g_1] - f)}_{Bias[g_1]} + \ldots + \underbrace{(E[g_M] - f)}_{Bias[g_M]}\right)\right)^2\right] + E_X\left[\frac{1}{M^2}\left(Var[g_1] + \ldots + Var[g_M]\right)\right]$$

$$= \sigma^2 + E_X\left[Bias[g_i]\right] + \frac{1}{M} \overbrace{E_X[Var[g_i]]}^{\text{constant}}$$

if M is large

$$\approx \sigma^2 + E_X\left[Bias[g_i]\right] \approx \sigma^2 \quad \text{theoretical minimum MSE.}$$

pick $\mathcal{H}$ very expressive

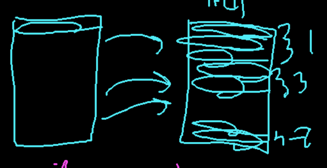
Leo Breiman invented "bagging" (Bootstrap AGGregatING) in 1994.

with one $\mathbb{D}$: what now?

Why not consider a sampling of the $\mathbb{D}$: $\mathbb{D}_{(1)}, \mathbb{D}_{(2)}, \ldots, \mathbb{D}_{(M)}$ where M is large

How to sample? Sample n rows/units with replacement. (Non-parametric bootstrap sampling).

$\mathbb{D}$ $\qquad$ $\mathbb{D}_{(1)}$



original dataset $\qquad$ bootstrap sample

$\mathbb{D}_{(1)}$ has both duplicate rows of $\mathbb{D}$ and omissions.

$$\mathbb{P}(\text{leaving row out}) = 1 - \frac{1}{n}$$

$$\mathbb{P}(\text{''''''} \; n \text{ times}) = \left(1 - \frac{1}{n}\right)^n \underset{n \text{ big}}{\approx} e^{-1} \approx \frac{1}{3}$$

$\Rightarrow$ In each bootstrap sample $\approx \frac{1}{3}$ omissions i.e. $\frac{2}{3}$ of the original $\mathbb{D}$.

Now, fit $g_{(1)} = \mathcal{A}(\mathbb{D}_{(1)}), \ldots, g_{(M)} = \mathcal{A}(\mathbb{D}_M)$

$g_{(1)}, \ldots, g_{(M)}$ are not the same but they will be dependent. $g_{avg} = \frac{g_{(1)} + \ldots + g_{(M)}}{M}$

Back to Math 241 .... let $X_1, \ldots, X_n \overset{iid}{\sim}$

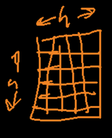
$$Var[\bar{X}] = Var\left[\frac{1}{n}\sum X_i\right] \overset{ind}{=} \frac{1}{n^2}\sum Var[X_i] \overset{\text{ident. distr}}{=} \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

---

let $X_1, \ldots, X_n$ be dependent but identically distr. Let $\sigma^2 = Var[X_i]$ and have the same dependence. Let $\rho = Corr[X_i, X_j]$ $i \neq j$

$$Corr[X_i, X_j] = \frac{Cov[X_i, X_j]}{SE[X_i] SE[X_j]} \Rightarrow \rho = \frac{Cov[X_i, X_j]}{\sigma \cdot \sigma} \Rightarrow Cov[X_i, X_j] = \sigma^2 \rho.$$

$$Var[\bar{X}] = \frac{1}{n^2} Var\left[\sum X_i\right] = \frac{1}{n^2}\left(Var[X_1] + \ldots + Var[X_n] + \sum_{i \neq j} Cov[X_i, X_j]\right)$$



$$= \frac{1}{n^2}\left(n\sigma^2 + (n^2 - n)\sigma^2\rho\right) = \frac{\sigma^2}{n} + \frac{n-1}{n}\sigma^2\rho = \sigma^2\left(\frac{1}{n} + \frac{n-1}{n}\rho\right)$$

if n large... $\approx \sigma^2\rho$.

$$= \frac{1}{n}\left(\sigma^2 + n\sigma^2\rho - \sigma^2\rho\right)$$

$$= \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2$$

if $\rho \approx 0$ $\rightarrow$ $\approx \frac{\sigma^2}{n}$

If $g_{(i)}, g_{(j)}$ have same correlation ρ and we use g_{avg} then...

$$MSE = \sigma^2 + E_X\left[Bias[g_{avg}]^2\right] + E_X\left[Var[g_{avg}]\right]$$

Let $\mathcal{H}$ be very expressive $\Rightarrow Bias[g_{(i)}] = 0 \Rightarrow Bias[g_{avg}] = 0$

$$= \sigma^2 + E_X\left[Var[g_{avg}]\right] = \sigma^2 + E_X\left[\rho \underbrace{Var[g_{(i)}]}_{\text{constant}} + \frac{1-\rho}{M} Var[g_{(i)}]\right]$$

if $M \to \infty$

$$= \sigma^2 + E_X\left[\rho \, Var[g_{(i)}]\right]$$

gain from bagging.

Since $\mathbb{D}_{(i)}, \mathbb{D}_{(j)}$ share observations $\Rightarrow g_{(i)}, g_{(j)}$ will be positively correlated i.e. $\rho \in (0, 1)$.

We want to minimize ρ as much as possible.