Important Announcements

March 4

- BGD: $\nabla \hat{L}(\vec{w}) = \left(\frac{\partial}{\partial w_0} \hat{L}(\vec{w}), \ldots \right.$
- $O(n^3)$ for inverse calculation
- MBGD $\nabla \hat{L}_m(\vec{w})$

Evaluating Predictors/Models

Is $\hat{L}(\hat{f}_D)$ really a good estimate of $L(\hat{f}_D)$?

where $\mathcal{A}(D) = f_D \in \mathcal{F}$ $\quad$ D is a r.v. $\quad$ $(\vec{X}, Y) \sim \mathbb{P}_{\vec{X},Y}$

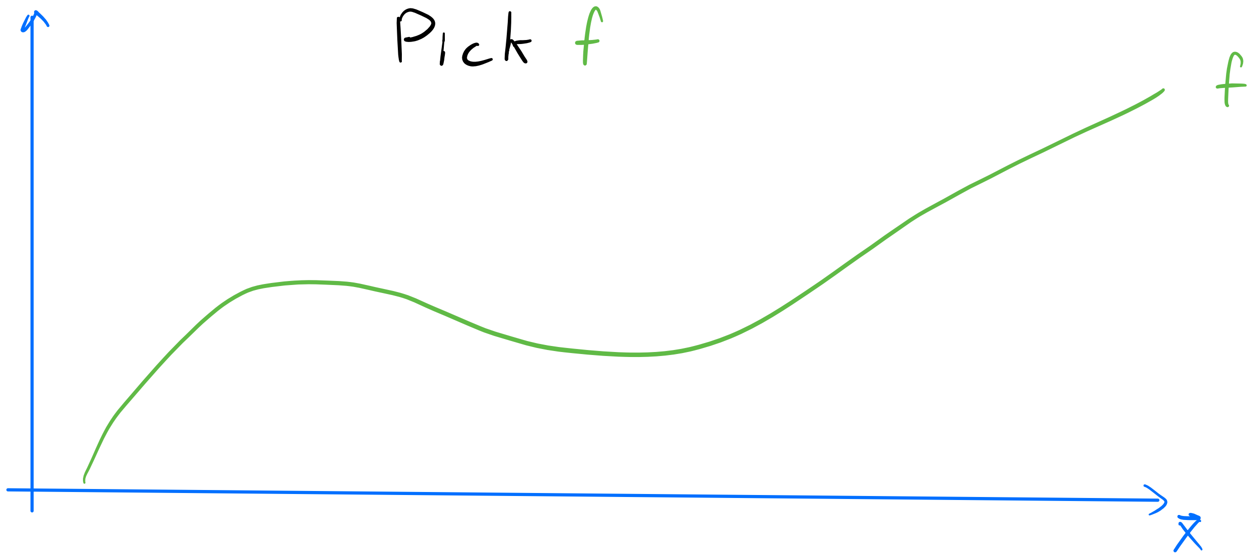
$$\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(\vec{X}_i), Y_i) \qquad L(f) = \mathbb{E}[\ell(f(\vec{X}), Y)]$$

(estimate of $L(f)$) $\qquad$ (expected loss)

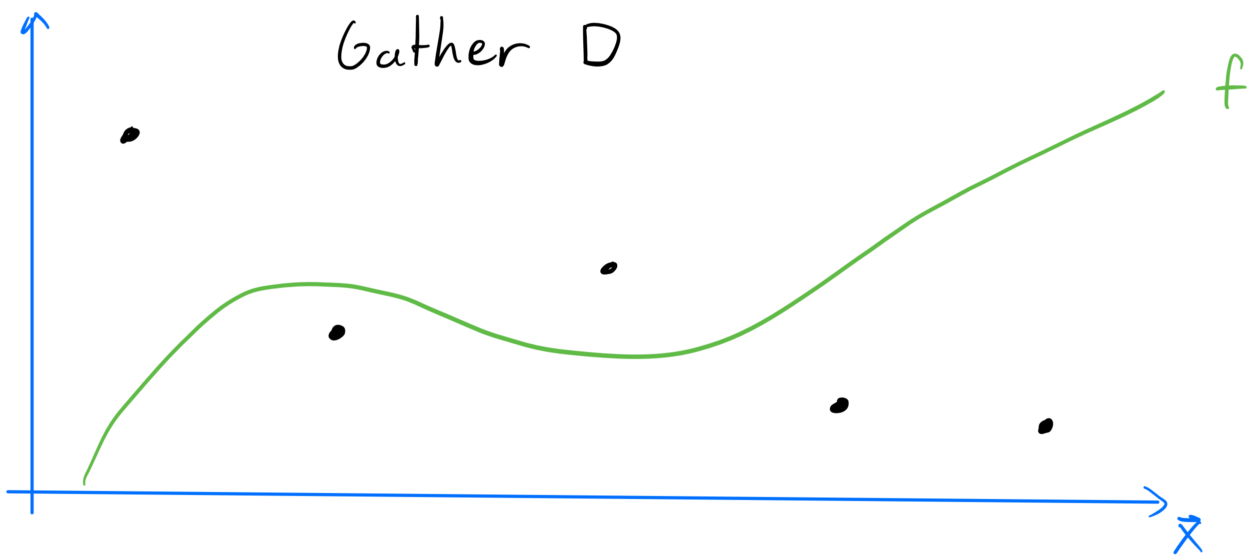
If we pick $f \in \mathcal{F}$ and then gather D

(i.e. f is chosen independently of D)

Pick f



Gather D



Then: $\mathbb{E}[\hat{L}(f)] = L(f)$

$$\mathrm{Var}[\hat{L}(f)] = \mathrm{Var}\Big[\frac{1}{n}\sum_{i=1}^{n} \ell(f(\vec{X}_i), Y_i)\Big]$$

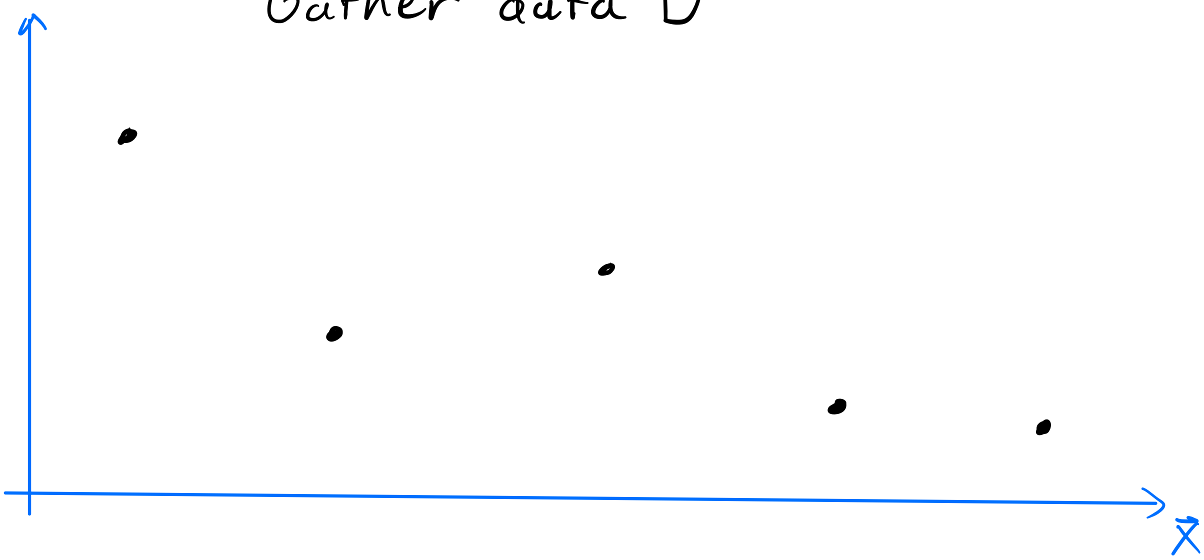
$$= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[\ell(f(\vec{X}_i), Y_i)]$$

$$= \frac{1}{n}\mathrm{Var}[\ell(f(\vec{X}_1), Y_1)]$$

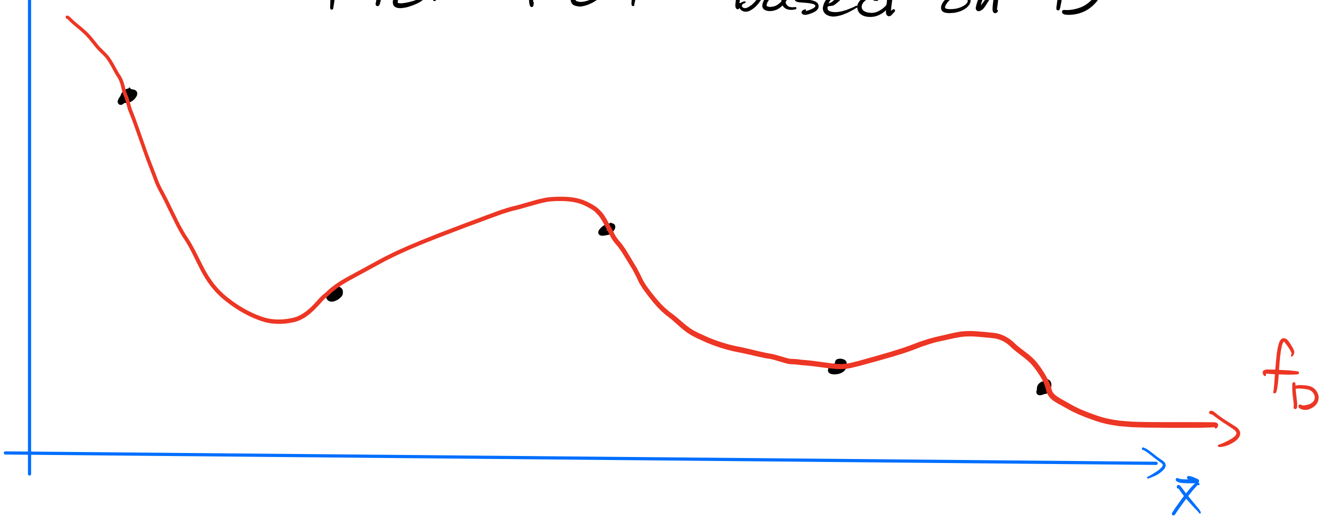
Since $\ell(f(\vec{X}_i), Y_i)$ are independent for all $i \in \{1, \ldots, n\}$

But we are gathering data D and then picking $\hat{f}_D \in \mathcal{F}$! (i.e. $\hat{f}_D$ depends on D)

Gather data D



Pick $f \in \mathcal{F}$ based on D



Then:

$$\mathbb{E}[\hat{L}(\hat{f}_D)] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(f(\vec{X}_i), Y_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(f(\vec{X}_i), Y_i)]$$

$$\neq \mathbb{E}[\ell(f(\vec{X}_1), Y_1)]$$

$$\mathrm{Var}[\hat{L}(\hat{f}_D)] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(\hat{f}_D(\vec{X}_i, Y_i))\right]$$

$$\neq \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[\ell(\hat{f}_D(\vec{X}_i), Y_i)]$$

$\ell(\hat{f}_D(\vec{X}_i), Y_i)$ are not i.i.d.

$\hat{f}_D$ depends on $(\vec{X}_1, Y_1), \ldots, (\vec{X}_n, Y_n)$!

Instead:

$$\mathbb{E}[\hat{L}(\hat{f}_D)] \leq \mathbb{E}[L(\hat{f}_D)]$$

$$\mathrm{Var}[\hat{L}(\hat{f}_D)] \leq \frac{1}{n} \max_{f \in \mathcal{F}} \mathrm{Var}[\ell(f(\vec{X}_1), Y_1)] + \mathrm{Var}[L(\hat{f}_D)]$$

As $\mathcal{F}$ gets larger, $\mathrm{Var}[\hat{L}(\hat{f}_D)]$ increases

$\Rightarrow$ i.e. $\hat{L}(\hat{f}_D)$ becomes a worse estimate of $L(\hat{f}_D)$

As n gets larger, $\mathrm{Var}[\hat{L}(\hat{f}_D)]$ decreases

$\Rightarrow$ $\hat{L}(\hat{f}_D)$ becomes a better estimate of $L(\hat{f}_D)$