Leave-one-out approach to analyze high-d ERMs

Setting: Let $\mathcal{D} = \{x_i, y_i\}_{i \in [n]}$ n iid samples.

We train a linear model:

$$\hat{w} = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i \in [n]} \ell(\langle w, x_i \rangle, y_i) + \frac{\lambda}{2} \|w\|^2$$

(Ex) for today: Binary classification, Gaussian separated case

$$x_i \sim \mathcal{N}(0, Id), \quad y_i = \operatorname{sign}(\langle \beta, x_i \rangle)$$
$$\|\beta\| = 1.$$

The test error is $E_g(\hat{w}) = \mathbb{P}_{x_{test}, y_{test}}\left[y_{test} \neq \operatorname{sign}(\langle \hat{w}, x_{test} \rangle) \right]$

Claim (e.g. El Karoui, 2013)

In the limit $\begin{cases} n, d \to \infty \\ n/d = \alpha = \Theta_n(1) \end{cases}$, $E_g(\hat{w}) \xrightarrow{P} E(\alpha, \lambda)$,

where E is some complicated function.

$\longrightarrow$ Gives a typical case characterization in high-d.

In the following, we highlight (heuristically, a proof being typically pretty lengthy) how to reach this claim.

Leave-one-out

Idea: $E_g(\hat{w}) = \frac{1}{\pi} \arccos\left[\frac{\langle \hat{w}, \beta \rangle}{\|\hat{w}\|} \right]$.

Remember that since $\hat{w}$ minimizes the empirical risk,

$$\hat{w} = -\frac{1}{\lambda n} \sum_i \ell_i'(\underbrace{\langle w, x_i \rangle}_{\equiv r_i}) x_i.$$

Thus:

$$\|\hat{w}\|^2 = -\frac{1}{\lambda n} \sum_i \ell_i'(r_i) r_i.$$

Ideally, we would like
① Concentration
② Distribution of r_i.
$\Rightarrow$ what follows

We would like to understand the distribution of

$$r_i = \langle \hat{w}, x_i \rangle$$

↑ complicated correlations with x_i !

Idea: define the "leave-one-out" empirical risk

$$\text{L.O.O.} \begin{cases} \hat{R}_{1,i}(w) = \frac{1}{n} \sum_{j \neq i}' \ell_j(\langle w, x_j \rangle) + \frac{\lambda}{2}\|w\|^2 \\ \hat{w}_{1,i} = \operatorname{argmin}_w \hat{R}_{1,i}(w) \end{cases}$$

Then the estimator

$$\hat{w} = \operatorname{argmin}_w \left[\hat{R}_{1,i}(w) + \frac{1}{n}\ell_i(\langle w, x_i \rangle) \right]$$

Taylor exp.

should intuitively be close to

$$\tilde{w}_i = \operatorname{argmin}_w \left[\hat{R}_{1,i}(\hat{w}_{1,i}) + \frac{1}{2}\langle w - \hat{w}_{1,i}, H_{1,i}(w - \hat{w}_{1,i}) \rangle + \frac{1}{n}\ell_i(\langle w, x_i \rangle) \right]$$

where we introduced the Hessian $H_{1,i} = \frac{1}{n}\sum_{j \neq i}' \ell_j''(r_{j,1,i}) x_j x_j^\top + \lambda \mathrm{Id}$

Thus,

$$\hat{w} \approx \tilde{w}_i = \hat{w}_{1,i} - \frac{1}{n}\ell_i'(\underbrace{\tilde{r}_i}_{\langle \tilde{w}_i, x_i \rangle}) H_{1,i}^{-1} x_i \quad (*)$$

(Lemma $\sup_{i \in [n]} \|\hat{w} - \tilde{w}_i\| = O_{L^2}\left(\frac{\text{polylog}(n)}{n}\right)$).

Taking $\langle x_i, (*) \rangle$:

$$\hat{r}_i = \underbrace{\langle \hat{w}_{1,i}, x_i \rangle}_{\sim \mathcal{N}(0, \|\hat{w}\|)} - \ell_i'(\tilde{r}_i) \underbrace{\frac{x_i^\top H_{1,i}^{-1} x_i}{n}}_{\approx \frac{1}{n}\mathrm{tr}(H_{1,i}^{-1}) \equiv V}$$

Gaussian: we removed the concentrated-complicated stat. dependencies !

Thus $\boxed{\tilde{r}_i = \operatorname{prox}_{V\ell(\cdot, y_i)}(\|\hat{w}\| g)}$

This result allows to go back and "close" the equation on $\|\hat{w}\|$

$$\|\hat{w}\|^2 = -\frac{1}{\lambda n}\sum_i \ell_i'(r_i)\, r_i \approx -\frac{1}{\lambda n}\sum_i \ell_i'(\tilde{r}_i)\,\tilde{r}_i$$

$$\approx -\frac{1}{\delta}\mathbb{E}\left[\ell'\left(\operatorname{prox}_{V\ell(\cdot;y)}(\|\hat{w}\|_{G,q} \operatorname{prox}_{V\ell(\cdot,y)}(\|\hat{w}\|_G)\right)\right]$$

Iterated ERMs:

(ERM I) $\hat{w}_0 = \underset{w\in\mathbb{R}^d}{\operatorname{argmin}} \sum_{i\in[n]} \ell_0(\langle w, x_i\rangle, y_i) + \frac{\lambda}{2}\|w\|^2$

(ERM II) $\hat{w} = \underset{w\in\mathbb{R}^d}{\operatorname{argmin}} \sum_{i\in[n]} \ell\left(\langle w, x_i\rangle, \langle \hat{w}_0, x_i\rangle, y_i\right) + \frac{\lambda}{2}\|w\|^2$

$\hat{w}_0$ depends on all the $\{x_j\}_{j\in[n]}$ through $\hat{w}_0$ (see ERM I).

Now, removing the i-th term from the sum no longer suffices to unravel all stat.-dependences on x_i. → nested L.O.O.

Thm (very informal) with Yue Lu, '26.

$$\begin{cases} \tilde{r}_{0,i} = \operatorname{prox}_{V\ell_0(\cdot, y_i)}(\|\hat{w}_0\| g_1) \to \text{same} \\ \tilde{r}_i = \operatorname{prox}_{V\ell(\cdot, \tilde{r}_{0,i}, y_i)}\left(\|\hat{w}\| g_2 + \chi \underbrace{\partial_u \ell_0(\tilde{r}_{0,i})}_{\text{additional term}}\right) \end{cases}$$

Applications:

- Optimization: e.g. proximal descents.
- Sample reweighting, e.g. AdaBoost.

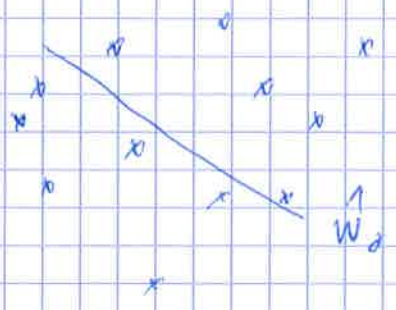
(Ex.) Active learning

You have n unlabeled data pts $\{x_i\}_{i\in[n]}$, but only enough budget to label a fraction $\gamma \in (0,1)$ of them. How to choose which points to label?

Idea (margin-based / uncertainty based selection)

If a pre-estimate $\hat{w}_0$ is available, label points closest to decision boundary

$$\left(\operatorname*{argmin}_{i} |\langle \hat{w}_0, x_i \rangle| \right)$$

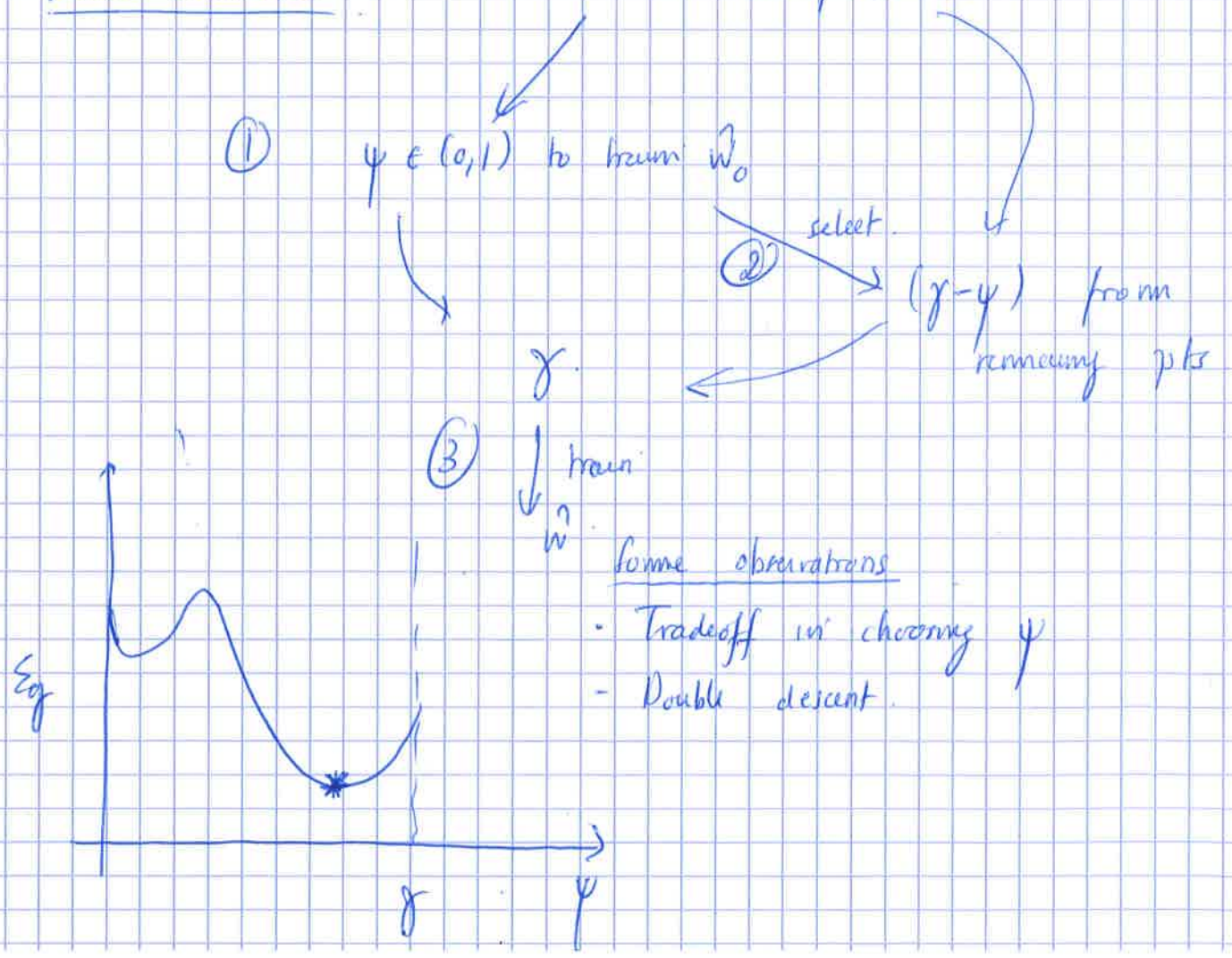


Many works in $n, d \to \infty$, $n/d = O(1)$ regime

[Kolesnov et al, '23, Sorscher et al, '22]

assume $\hat{w}_0 \perp\!\!\!\perp$ dataset (oracle, or held-out data).

More realistic:



Some observations

- Tradeoff in choosing ψ
- Double descent