

# GML Fall 25, Interactive Session on Multi-objective Learning

## Recap of presentation & goal

**Setting.** We consider  $K \in \mathbb{N}$  binary classification problems, composed of distributions  $P^1, \dots, P^K$  on  $\mathcal{X} \times \{0, 1\}$  where  $\mathcal{X} = \{x_1, \dots, x_m\}$ . We use zero-one loss  $\ell(y, \hat{y}) = \mathbf{1}\{y \neq \hat{y}\}$  and risks of a classifier  $f : \mathcal{X} \rightarrow \{0, 1\}$  defined as

$$\mathcal{R}_k(f) = \mathbb{E}_{P^k} \ell(Y, f(X)) \quad \text{where} \quad f_k^* \in \arg \min_{f: \mathcal{X} \rightarrow \{0, 1\}} \mathcal{R}_k(f)$$

achieves optimal risk. We consider the set  $\mathcal{G}$  of all  $2^m$  possible functions  $\mathcal{X} \rightarrow \{0, 1\}$ . Our goal is to learn a subset of functions  $\hat{g}_\lambda \in \mathcal{G}$  with  $\lambda \in \Delta^{K-1} = \left\{ \lambda \in \mathbb{R}^K : \sum_{k=1}^K \lambda_k = 1, \lambda_k \geq 0 \right\}$  that achieves

$$\mathbb{P} \left( \forall \lambda \in \Delta^{K-1} : \sum_{k=1}^K \lambda_k \mathcal{R}_k(\hat{g}_\lambda) \leq \inf_{g \in \mathcal{G}} \sum_{k=1}^K \lambda_k \mathcal{R}_k(g) + \varepsilon \right) \geq 1 - \delta, \quad (1)$$

where we take the probability with respect to  $K$  i.i.d. datasets of size  $n$  sampled from  $P^k$ .

Recall from class that  $\mathcal{G}$  has VC dimension  $\text{VC}(\mathcal{G}) = m$ . We know that learning  $\mathcal{G}$  to error  $\varepsilon$  on one distribution  $P^k$  requires  $\Theta(\text{VC}(\mathcal{G})/\varepsilon^2)$  i.i.d. samples. So it is not surprising that achieving Eq. (1) requires at least  $\Theta(K \cdot \text{VC}(\mathcal{G})/\varepsilon^2)$  samples. After all, Eq. (1) includes all individual learning tasks. But does the hardness of Eq. (1) really originate in the fact that it includes the individual learning tasks?

**Goal of the session.** To investigate this, we assume that the learner has additional access to  $\{f_k^*, P_X^k\}_{k=1}^K$ : it can perfectly solve all tasks individually, and even has access to the marginal distributions over  $\mathcal{X}$ . If we can show a similar lower bound of order  $\Omega(K \cdot \text{VC}(\mathcal{G})/\varepsilon^2)$ , we know that achieving good trade-offs can be inherently hard, even when we know how to solve the tasks separately! In particular, we will prove such a lower bound for the simpler case of  $K = 2$  objectives.

To that end, we study the specific family of distributions  $P_\sigma^1, P_\sigma^2$  indexed by  $\sigma \in \{0, 1\}^m$ , which are defined through

$$X_\sigma^1, X_\sigma^2 \sim \text{Uniform}(\mathcal{X}), \quad (Y_\sigma^1 | X_\sigma^1 = x_i) \sim \text{Ber} \left( \frac{1}{2} + 4\varepsilon \sigma_i \right), \quad (Y_\sigma^2 | X_\sigma^2 = x_i) \sim \text{Ber} \left( \frac{1}{2} - 4\varepsilon (1 - \sigma_i) \right). \quad (2)$$

Hence, when we see  $n$  samples from both  $P_\sigma^1$  and  $P_\sigma^2$ , this is equivalent to seeing one sample  $Z$  from  $Q_\sigma := (P_\sigma^1 \otimes P_\sigma^2)^{\otimes n}$ , where  $Q_\sigma$  is now a distribution on  $(\mathcal{X} \times \{0, 1\})^{2n}$ . We denote  $\hat{g} \equiv \hat{g}(Z) \in \mathcal{G}$ .

**Theorem 1.** *Let  $K = 2$  and  $\varepsilon \in (0, 1/12)$ . Even when the learner  $\hat{g} : (\mathcal{X} \times \{0, 1\})^{2n} \rightarrow \mathcal{G}$  has access to solutions  $f_1^* \equiv 1, f_2^* \equiv 0$  and the marginals  $P_X^k$ , it still requires  $n \geq \text{VC}(\mathcal{G})/1024\varepsilon^2$  samples from  $P^1$  and  $P^2$  to achieve*

$$\min_{\sigma \in \{0, 1\}^m} \mathbb{P}_{Z \sim Q_\sigma} \left( \frac{\mathcal{R}_1(\hat{g}) + \mathcal{R}_2(\hat{g})}{2} - \inf_{g \in \mathcal{G}} \left\{ \frac{\mathcal{R}_1(g) + \mathcal{R}_2(g)}{2} \right\} \leq \varepsilon \right) \geq 5/6,$$

even though it can trivially achieve minimal risks  $\mathcal{R}_1(\hat{g}) = \mathcal{R}_1(f_1^*)$  or  $\mathcal{R}_2(\hat{g}) = \mathcal{R}_2(f_2^*)$  by returning  $f_1^*$  and  $f_2^*$ , respectively.

This demonstrates that for zero-one loss, even if we can perfectly solve each task separately, that does not imply any benefit for solving both tasks jointly (Eq. (1))!

**Bonus if you are done early:** From Theorem 1, derive the lower bound  $n \geq cK \cdot \text{VC}(\mathcal{G})/\varepsilon^2$  in the general case  $K \geq 2$ , where  $c > 0$  is some universal constant.

## Group 1: Gaining some intuition

Consider the following game: There are two biased coins,  $C_1$  and  $C_2$ , that have probabilities of landing heads up  $\mathbb{P}(C_1 = H) = p_1$  and  $\mathbb{P}(C_2 = H) = p_2$ . Before both coins are tossed, you have to make exactly one prediction  $\hat{y} \in \{H, T\}$ . After the coins are tossed, you then get points according to the following rule:

$$s(\hat{y}) := \text{score from betting } \hat{y} = \mathbf{1}\{C_1 = \hat{y}\} + \mathbf{1}\{C_2 = \hat{y}\} \in \{0, 1, 2\}.$$

Of course, without any prior knowledge, there is not much you can do beyond random guessing. Instead, you can choose to get prior knowledge in one of two ways:

1. (*sign*) You get to know  $y_1 \in \arg \max_{y \in \{H, T\}} \mathbb{P}(C_1 = y)$  and  $y_2 \in \arg \max_{y \in \{H, T\}} \mathbb{P}(C_2 = y)$ , but if there are multiple maximizers, you don't know which one you are getting.
2. (*toss*) You can toss the two coins once before the game and observe the outcome (i.e., observe i.i.d. copies  $\tilde{C}_1, \tilde{C}_2$  of  $C_1, C_2$ ).

In either case, you then play these strategies, respectively:

$$\hat{y}_{\text{sign}} = \begin{cases} H & \text{if } y_1 = y_2 = H, \\ T & \text{if } y_1 = y_2 = T, \\ \text{Uniform}(\{H, T\}) & \text{else,} \end{cases} \quad \text{and} \quad \hat{y}_{\text{toss}} = \begin{cases} H & \text{if } \tilde{C}_1 = \tilde{C}_2 = H, \\ T & \text{if } \tilde{C}_1 = \tilde{C}_2 = T, \\ \text{Uniform}(\{H, T\}) & \text{else.} \end{cases}$$

Answer the following questions as a function of  $p_1, p_2$ :

- (a) **Compute**  $\max_{y \in \{H, T\}} \mathbb{E}[s(y)]$  and  $\mathbb{E}_{\hat{y}_{\text{rand}} \sim \text{Uniform}(\{H, T\})} \mathbb{E}[s(\hat{y}_{\text{rand}})]$ .
- (b) **Compute** the expected scores  $\mathbb{E}[s(\hat{y}_{\text{toss}})]$  and  $\min_{y_1, y_2} \mathbb{E}[s(\hat{y}_{\text{sign}})]$  (where the min is over the maximizer in  $\{H, T\}$ ) and try to visualize them as a function of  $(p_1, p_2) \in [0, 1]^2$ .
- (c) On the set  $\mathcal{P} \subset [0, 1]^2$  where the sign strategy is no better than random guessing, **solve** the following maximization problem:

$$\max_{(p_1, p_2) \in \mathcal{P}} \left\{ \max_{y \in \{H, T\}} \mathbb{E}[s(y)] - \mathbb{E}[s(\hat{y}_{\text{toss}})] \right\}.$$

- (d) How does it relate to the distributions in Eq. (2)? And what do you expect happens when we get to throw the dice multiple times before the game?

## Group 2: Reduction to estimating $\sigma$

In this exercise, we prove an equality that is reminiscent of the “estimation to testing” reduction we saw in the lecture on minimax lower bounds. This is a key ingredient for proving Theorem 1.

**Prove that** for the set of distributions  $Q_\sigma$  from Eq. (2), it holds that

$$\sup_{\hat{g}} \min_{\sigma} \mathbb{P}_{Z \sim Q_\sigma} \left( \frac{\mathcal{R}_1(\hat{g}) + \mathcal{R}_2(\hat{g})}{2} - \inf_{g \in \mathcal{G}} \left\{ \frac{\mathcal{R}_1(g) + \mathcal{R}_2(g)}{2} \right\} \leq \varepsilon \right) = \sup_{\hat{\sigma}} \min_{\sigma} \mathbb{P}_{Z \sim Q_\sigma} \left( \frac{\|\hat{\sigma}(Z) - \sigma\|_1}{d} \leq \frac{1}{4} \right). \quad (3)$$

*Hints:* 1. It may be useful to replace the risks by the excess risks  $\mathcal{E}_k(g) = \mathcal{R}_k(g) - \mathcal{R}_k(f_k^*)$ .

2. When you explicitly compute  $\frac{\mathcal{E}_1(g) + \mathcal{E}_2(g)}{2}$ , notice that the infimum over  $\mathcal{G}$  vanishes. Why?

3. Given an  $\hat{g}$ , consider the estimator  $\hat{\sigma}(Z) = \hat{g}(x_1^m) \in \mathbb{R}^n$  where  $g(x_1^m) = (g(x_1), \dots, g(x_m)) \in \{0, 1\}^m$ .

## Group 3: Application of Assouad's method

We are now going to prove Theorem 1 using an alternative to Fano's method, which is called *Assouad's method*, captured in the following Lemma:

**Lemma 1** (Assouad). *Write  $\sigma \sim \sigma'$  if  $\sigma$  and  $\sigma'$  differ only in one coordinate, and let  $Q_\sigma, \sigma \in \{0, 1\}^m$  be any distributions. It holds that*

$$\inf_{\hat{\sigma}} \max_{\sigma} \mathbb{E}_{Z \sim Q_\sigma} [\|\hat{\sigma}(Z) - \sigma\|_1] \geq \frac{m}{2} \min \left\{ 1 - \sqrt{\frac{1}{2} \text{KL}(Q_\sigma, Q_{\sigma'})} : \sigma \sim \sigma' \right\}.$$

Using Lemma 1, and the reduction from Eq. (3) from Group 2, **prove** Theorem 1.

*Hints:* 1. Assume that  $n < m/1024\varepsilon^2$  and find a contradiction.

2. You may use that if  $\sigma \sim \sigma'$  and  $\varepsilon \leq 1/12$ ,  $\text{KL}(Q_\sigma, Q_{\sigma'}) \leq \frac{128n\varepsilon^2}{m}$  where  $Q_\sigma$  is defined in Eq. (2).