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 $\ell(y, \widehat{y}) = \mathbf{1}\{y \neq \widehat{y}\}$ and risks of a classifier $f : \mathcal{X} \to \{0,1\}$ defined as

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

achieves optimal risk. We consider the set $\mathcal{G}$ of all 2^m possible functions $\mathcal{X} \to \{0,1\}$. Our goal is to learn a subset of functions $\widehat{g}_\lambda \in \mathcal{G}$ 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

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

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 $\mathrm{VC}(\mathcal{G}) = m$. We know that learning $\mathcal{G}$ to error ε on one distribution P^k requires $\Theta(\mathrm{VC}(\mathcal{G})/\varepsilon^2)$ i.i.d. samples. So it is not surprising that achieving Eq. (1) requires at least $\Theta(K \cdot \mathrm{VC}(\mathcal{G})/\varepsilon^2)$ 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 $\left\{f_k^\star, P_X^k\right\}_{k=1}^K$: 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 $\Omega(K \cdot \mathrm{VC}(\mathcal{G})/\varepsilon^2)$, 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_σ^1, P_σ^2 indexed by $\sigma \in \{0,1\}^m$, which are defined through

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

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

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

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

even though it can trivially achieve minimal risks $\mathcal{R}_1(\widehat{g}) = \mathcal{R}_1(f_1^\star)$ or $\mathcal{R}_2(\widehat{g}) = \mathcal{R}_2(f_2^\star)$ by returning $f_1^\star$ and $f_2^\star$, 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 $n \geq cK \cdot \mathrm{VC}(\mathcal{G})/\varepsilon^2$ 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 $\widehat{y} \in \{H, T\}$. After the coins are tossed, you then get points according to the following rule:

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

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 $\widetilde{C}_1, \widetilde{C}_2$ of C_1, C_2).

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

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

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

(a) **Compute** $\max_{y \in \{H,T\}} \mathbb{E}[s(y)]$ and $\mathbb{E}_{\widehat{y}_{\text{rand}} \sim \text{Uniform}(\{H,T\})} \mathbb{E}[s(\widehat{y}_{\text{rand}})]$.

(b) **Compute** the expected scores $\mathbb{E}[s(\widehat{y}_{\text{toss}})]]$ and $\min_{y_1, y_2} \mathbb{E}[s(\widehat{y}_{\text{sign}})]]$ (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:

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

(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?

Solution. The expected score for playing heads and tails is

$$\mathbb{E}[s(y)] = \mathbb{P}(C_1 = y) + \mathbb{P}(C_2 = y) = \begin{cases} p_1 + p_2 & y = H, \\ 2 - p_1 - p_2 & y = T. \end{cases}$$

(a) So, under perfect knowledge, the optimal strategy is to play $y = H$ if $p_1 + p_2 \geq 1$, otherwise play $y = T$, which then yields $\max_{y \in \{H,T\}} \mathbb{E}[s(y)] = \max\{p_1 + p_2, 2 - p_1 - p_2\}$. On the other hand, a random guesser that chooses $\widehat{y}_{\text{rand}} = \text{Uniform}(\{H, T\})$ will achieve expected score $\mathbb{E}[s(\widehat{y}_{\text{rand}})] = \frac{1}{2}(p_1 + p_2) + \frac{1}{2}(2 - p_1 - p_2) = 1$, which serves as a baseline for any strategy.

(b) A direct calculation shows that the first strategy has the expected score

$$\min_{y_1, y_2} \mathbb{E}[s(\widehat{y}_{\text{sign}})] = \min_{y_1, y_2} \begin{cases} p_1 + p_2 & y_1 = y_2 = H \\ 2 - p_1 - p_2 & y_1 = y_2 = T \\ 1 & \text{else} \end{cases} = \begin{cases} p_1 + p_2 & p_1, p_2 > 1/2 \\ 2 - p_1 - p_2 & p_1, p_2 < 1/2 \\ 1 & \text{else,} \end{cases}$$

where minimizing over y_1, y_2 yields the second equality: If $p_1 = 1/2$ and $p_2 > 1/2$, choose y_1 adversarially as $y_1 = T$, and so on.

To compute the score for the second strategy, define the events $A_H = \{\widetilde{C}_1 = \widetilde{C}_2 = H\}$, $A_T = \{\widetilde{C}_1 = \widetilde{C}_2 = T\}$ and $A_R = (A_H \cup A_T)^c$ and note that $\mathbb{P}(A_H) = p_1 p_2$, $\mathbb{P}(A_T) = (1 - p_1)(1 - p_2)$ and $\mathbb{P}(A_R) = 1 - p_1 p_2 - (1 - p_1)(1 - p_2) = p_1 + p_2 - 2p_1 p_2$. We can see that the expected score is

$$\begin{aligned} \mathbb{E}[s(\widehat{y}_{\text{toss}})] &= (\mathbb{P}(A_H) + \tfrac{1}{2}\mathbb{P}(A_R))(p_1 + p_2) + (\mathbb{P}(A_T) + \tfrac{1}{2}\mathbb{P}(A_R)))(2 - (p_1 + p_2)) \\ &= \tfrac{1}{2}(p_1 + p_2)(p_1 + p_2) + ((1 - p_1)(1 - p_2) + \tfrac{1}{2}(p_1 + p_2 - 2p_1 p_2))(2 - (p_1 + p_2)) \\ &= 1 + (p_1 + p_2 - 1)^2. \end{aligned}$$

See also Fig. 1.

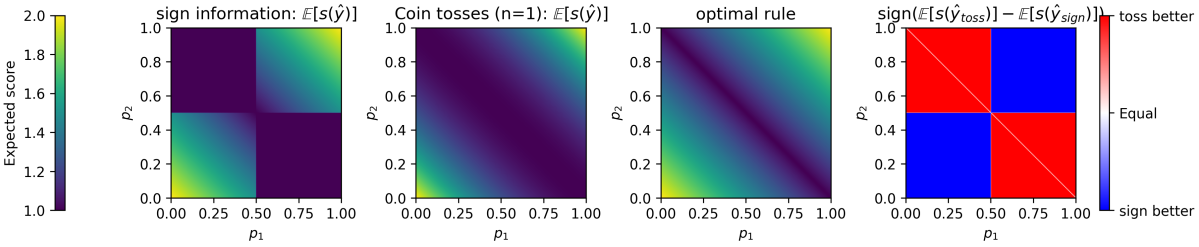

Figure 1: Expected scores.

(c) We see that $\mathcal{P} = \{(p_1, p_2) : p_1 \le 1/2 \le p_2 \text{ or } p_2 \le 1/2 \le p_1\}$. By symmetry, we can focus on the case $p_1 \le 1/2 \le p_2$ and $p_1 + p_2 \ge 1$. Then

$$\max_{(p_1,p_2)\in\mathcal{P}} \left\{ \max_{y\in\{H,T\}} \mathbb{E}[s(y)] - \mathbb{E}[s(\widehat{y}_{\text{toss}})]] \right\} = \max_{(p_1,p_2)\in\mathcal{P}} \left\{ p_1 + p_2 - 1 - (p_1 + p_2 - 1)^2 \right\}$$

which is clearly attained at $p_1 = 1/2, p_2 = 1$. By symmetry, we know the set of solution is

$$\{(0, 1/2), (1/2, 1), (1/2, 0), (1, 1/2)\}.$$

(d) Consider the distributions of Y_σ^1, Y_σ^2 conditioned on $X_\sigma^k = x_i$, and w.l.o.g. consider $\sigma_i = 1$: We have that $p_1 = 1/2 + 4\varepsilon > 1/2$ and $p_2 = 1/2$, that is, coin 1 is biased towards 1, while the other coin is unbiased. In this setting, we have seen that the strategy using sign information has expected score 1, and does not perform better than random guessing. However, seeing just one sample of labeled data, we know that the coin toss strategy improves upon random guessing. Naturally, the more often we toss the coins, the closer to optimal the estimates will be of an "adapted" coin tossing strategy will be:

$$\widehat{y}_n = \begin{cases} H & \frac{1}{n}\sum_{i=1}^n \widetilde{C}_{1,i} + \frac{1}{n}\sum_{i=1}^n \widetilde{C}_{2,i} > 1 \\ T & \frac{1}{n}\sum_{i=1}^n \widetilde{C}_{1,i} + \frac{1}{n}\sum_{i=1}^n \widetilde{C}_{2,i} < 1 \end{cases}$$

with the standard tie-break. In particular, without proving it here, the parameters maximizing the excess risk then become

$$\left\{(1/2 - 1/\sqrt{4n}, 1/2), (1/2, 1/2 + 1/\sqrt{4n}), (1/2, 1/2 - 1/\sqrt{4n}), (1/2 + 1/\sqrt{4n}, 1/2)\right\}.$$

Hence, we should choose $\varepsilon \asymp 1/\sqrt{n}$, or, $n \asymp 1/\varepsilon^2$. To solve m such coin tossing problems simultaneously, we hence need $n \asymp m/\varepsilon^2$. This is exactly the intuition we will use to establish the lower bound.

□

Group 2: Reduction to estimating σ

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_σ from Eq. (2), it holds that

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

Hints: 1. It may be useful to replace the risks by the excess risks $\mathcal{E}_k(g) = \mathcal{R}_k(g) - \mathcal{R}_k(f_k^\star)$.
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 $\widehat{g}$, consider the estimator $\widehat{\sigma}(Z) = \widehat{g}(x_1^m) \in \mathbb{R}^n$ where $g(x_1^m) = (g(x_1), \dots, g(x_m)) \in \{0,1\}^m$.

Solution. First, note that by linearity,

$$\frac{\mathcal{R}_1(\widehat{g}) + \mathcal{R}_2(\widehat{g})}{2} - \inf_{g\in\mathcal{G}} \left\{ \frac{\mathcal{R}_1(g) + \mathcal{R}_2(g)}{2} \right\} = \frac{\mathcal{E}_1(\widehat{g}) + \mathcal{E}_2(\widehat{g})}{2} - \inf_{g\in\mathcal{G}} \left\{ \frac{\mathcal{E}_1(g) + \mathcal{E}_2(g)}{2} \right\}.$$

Second, note that on the first distribution, $f_1^\star \equiv 1$ is an optimal predictor, and so a function g only incurs a worse error on x_i than $f_1^\star$ if $\sigma_i = 1$ and $g(x_i) = 0$, in which case the expected error is exactly $\frac{1}{2} + 4\varepsilon$

whereas $f^\star$ would have error $\frac{1}{2} - 4\varepsilon$, and so the excess on such a point is 8ε. The analogue also holds on the second distribution, and so averaging yields that the excess risks of any g are

$$\mathcal{E}_1(g) = \frac{8\varepsilon}{m}\sum_{j=1}^{m}\sigma_j \mathbf{1}\left\{g(x_i) = 0\right\} \qquad \text{and} \qquad \mathcal{E}_2(g) = \frac{8\varepsilon}{m}\sum_{j=1}^{m}(1-\sigma_j)\mathbf{1}\left\{g(x_i) = 1\right\}.$$

Combining the two yields

$$\frac{\mathcal{E}_1(g) + \mathcal{E}_2(g)}{2} = \frac{4\varepsilon}{m}\sum_{j=1}^{m}\mathbf{1}\left\{g(x_i) \neq \sigma_i\right\} = \frac{4\varepsilon}{m}\left\|g(x_1^m) - \sigma\right\|_1.$$

Since $\mathcal{G}$ contains one g so that $g(x_j) = \sigma_j$ for all j, the infimum of the excess risks is zero.

From the previous derivation, we get that

$$\frac{\mathcal{E}_1(g) + \mathcal{E}_2(g)}{2} = \frac{4\varepsilon}{m}\left\|g(x_1^m) - \sigma\right\|_1 \leq \varepsilon \iff \frac{\left\|g(x_1^m) - \sigma\right\|_1}{m} \leq \frac{1}{4}.$$

The equality then follows from the previous derivation by considering, for any $\widehat{g}$, the estimator $(\widehat{\sigma}(Z))_i = \widehat{g}(x_i)$. □

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 σ and σ' differ only in one coordinate, and let $Q_\sigma, \sigma \in \{0,1\}^m$ be any distributions. It holds that*

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

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 $\varepsilon \leq 1/12$, $\operatorname{KL}(Q_\sigma, Q_{\sigma'}) \leq \frac{128n\varepsilon^2}{m}$ where Q_σ is defined in Eq. (2).

Solution. We will show that if $n < m/1024\varepsilon^2$,

$$\sup_{\widehat{\sigma}}\min_{\sigma}\mathop{\mathbb{P}}_{Z\sim Q_\sigma}\left(\frac{\left\|\widehat{\sigma}(Z) - \sigma\right\|_1}{m} \leq \frac{1}{4}\right) < \frac{5}{6}.$$

Theorem 1 then follows by the reduction in Eq. (3).

Fix any estimator $\widehat{\sigma}$. From Markov's inequality, we get that

$$\mathop{\mathbb{P}}_{Z\sim Q_\sigma}\left(\frac{\left\|\widehat{\sigma}(Z) - \sigma\right\|_1}{m} \leq \frac{1}{4}\right) = \mathop{\mathbb{P}}_{Z\sim Q_\sigma}\left(1 - \frac{\left\|\widehat{\sigma}(Z) - \sigma\right\|_1}{m} \geq \frac{3}{4}\right) \leq \frac{4}{3}\left(1 - \mathop{\mathbb{E}}_{Z\sim Q_\sigma}\left[\frac{\left\|\widehat{\sigma}(Z) - \sigma\right\|_1}{m}\right]\right).$$

From Assouad's Lemma and the previous bound, we get

$$\max_{\sigma}\mathop{\mathbb{E}}_{Z\sim Q_\sigma}\left[\left\|\widehat{\sigma}(Z) - \sigma\right\|_1\right] \geq \frac{m}{2}\left(1 - \sqrt{\frac{64n\varepsilon^2}{m}}\right) > \frac{m}{2}\left(1 - \frac{1}{4}\right) = \frac{3m}{8}$$

where the last inequality follows from $n < m/1024\varepsilon^2$. Combining the two, we get that

$$\min_{\sigma}\mathop{\mathbb{P}}_{Z\sim Q_\sigma}\left(\frac{\left\|\widehat{\sigma}(Z) - \sigma\right\|_1}{d} \leq \frac{1}{4}\right) < \frac{4}{3}\left(1 - \frac{3}{8}\right) = \frac{5}{6}.$$

That concludes the proof by contradiction. □