Nuances in Margin Conditions Determine Gains in Active Learning

Samory Kpotufe, Gan Yuan, Yunfan Zhao

Columbia University

Active learning and margin conditions

Let $P_{X,Y}$ be a joint distribution on $[0,1]^d \times \{1,2\}$, with smooth regression functions $\eta_y(x) = \mathbb{P}(Y = y|X = x)$, $y = 1,2$. We consider an **active learner** $\hat{h}$ that interactively query the label at any point in Support(P_X).

- **Excess risk under 0-1 loss**:

$$\mathcal{E}(\hat{h}) = \mathbb{E}[\max_{y \in \{0,1\}} \eta_y(X) - \eta_{\hat{h}(X)}(X)].$$

- **Two notions of margin conditions:**

(MC1) $\forall \tau > 0$, $\mathbb{P}(0 < |\eta_1(X) - \eta_2(X)| < \tau) \lesssim \tau^\beta$;

(MC2) $\forall \tau > 0$, $\mathbb{P}(|\eta_1(X) - \eta_2(X)| < \tau) \lesssim \tau^\beta$.

The seemingly benign difference in (MC1) and (MC2) determines whether active learners can has faster excess risk rate than passive ones.

No gain under (MC1) + strong density

Let $\mathcal{P}(\alpha, \beta)$ denote the class of distributions such that:

- P_X satisfies a "strong density condition" (nearly uniform);
- The regression functions are α-Hölder with $0 < \alpha \leq 1$;
- $P_{X,Y}$ satisfies (MC1) with parameter $\beta > 0$.

Theorem 1

For $\alpha\beta \leq d$, $\exists C_1 > 0$, independent of n:

$$\inf_{\text{active learner } \hat{h}} \sup_{P_{X,Y} \in \mathcal{P}(\alpha,\beta)} \mathbb{E}\, \mathcal{E}(\hat{h}_n) \geq C_1 n^{-\frac{\alpha(\beta+1)}{2\alpha+d}}.$$

Remark 1

The rate in Theorem 1 matches the lower minimax passive rate.

Improved rate under (MC2) + strong density

Theorem 2

Let $\alpha\beta \leq d$. Assume that:

- *P_X satisfies a "strong density condition" (nearly uniform);*
- *The regression functions are α-Hölder with $0 < \alpha \leq 1$;*
- *$P_{X,Y}$ satisfies (MC2) with β.*

Then, $\exists$ an active learner $\hat{h}_n$ and $C_2 > 0$ independent of n, such that w.h.p,

$$\mathcal{E}\left(\hat{h}_n\right) \leq \; C_2 n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\alpha\beta}}$$

Remark 2

The resulting upper bound in Theorem 2 is faster than the lower minimax rate of passive learning.

Improved rate under (MC1) + general density

Theorem 3

Let $\alpha\beta \leq d$. Assume that:

- *The regression functions are α-Hölder with $0 < \alpha \leq 1$;*
- *$P_{X,Y}$ satisfies (MC1) with β.*

Then, $\exists$ an active learner $\hat{h}_n$ and $C_3 > 0$ independent of n, such that w.h.p,

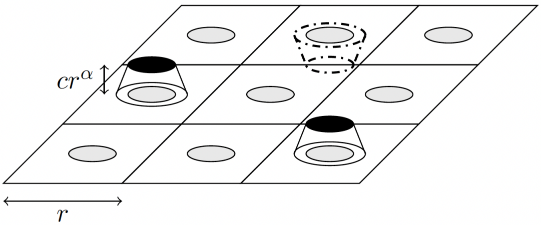
$$\mathcal{E}\left(\hat{h}_n\right) \leq \; C_3 n^{-\frac{\alpha(\beta+1)}{2\alpha+d}}.$$

Remark 3

The rate in Theorem 3 is faster than the lower minimax rate of passive learning under general density, i.e. $n^{-\frac{\alpha(\beta+1)}{2\alpha+d+\alpha\beta}}$.

Outline of Proof for Theorem 1

- **A randomized construction for $P_{X,Y}$**
  To make sure that the learner does not know the locations of the bumps, otherwise it can have savings by sampling only at bumps.



- **Such $P_{X,Y}$ is in the class $\mathcal{P}(\alpha, \beta)$ w.h.p.**

Outline of Proof for Theorem 1

- **Decouple the sampling mechanism and label prediction**
  - For fixed sampling mechanism, there is an optimal label prediction rule;
  - We name the active learners with such prediction rule **Conditional Neyman Pearson (CNP) Learners**.
- **No CNP learners enjoys a faster rate than** $n^{-\alpha(\beta+1)/(2\alpha+d)}$