Achieving Minimax Rates in Pool-Based Batch Active Learning

Zhilei Wang
Citadel Securities, NY
zhileiwang92@gmail.com

International Conference on Machine Learning 2022

July 17-23rd, 2022

Collaborators: Claudio Gentile[1], Tong Zhang[1,2]

[1]Google Research, [2]HKUST

Active Learning (AL) and Contribution

Aims at reducing data requirement in (statistical) inference
by designing algorithms that can learn and generalize from
small subset of training data

→ Pool-based AL:
Algorithm has prior access to large unlabeled pool of data points

→ Batch Pool-based AL:
Pool-based AL where labels are requested in sequence of batches

Contribution:

First theoretical results that carefully tradeoff

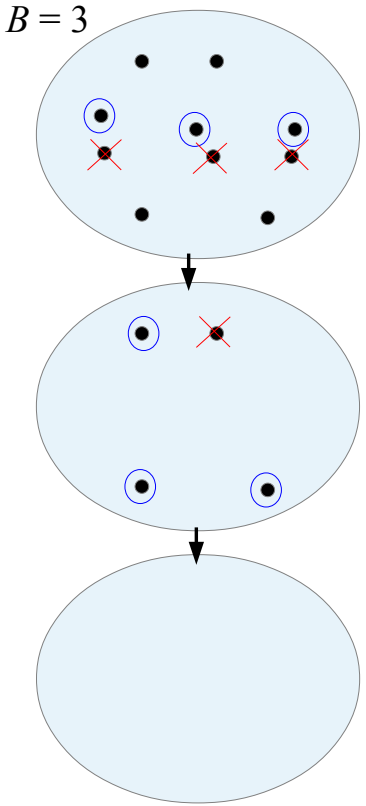
informativeness and diversity to rigorously quantify

statistical performance in batch pool-based AL

Batch Pool-based AL

Batch pool-based AL:

- Given unlabeled $Pool \subseteq \mathcal{X}$ of points and batch size B, select subset of points (circles) to query labels of
- Train model using the current batch, delete non-informative points (crosses)
- Repeat until stopping criterion fulfilled (e.g. $Pool$ is empty)
- Batch AL: selection of points within every batch cannot depend on labels gathered on that batch but only on past batches
- Alg to tradeoff informativeness vs. diversity
- The larger B the harder the problem (smaller adaptivity power)



Problem settings

- Binary classification $y_t \in \mathcal{Y} = \{\pm 1\}$
- Data $\{(x_t, y_t)\}_{t=1}^T$ drawn i.i.d. according to unknown distribution on $\mathcal{X} \times \mathcal{Y}$
- Parametric realizable scenario

  $$\mathbb{P}(y = 1 \,|\, x) = h^*(x), \qquad h^* \in \mathcal{H}$$

  for some (large) function space $\mathcal{H}$

  (reasonable in DNN-based overparametrized regimes)
- Work on noisy (generalized) linear model
- Low noise condition

  $$\Pr\left(|h^*(x) - \frac{1}{2}| < \epsilon\right) \leq \epsilon^\alpha, \qquad \epsilon \in (0, \epsilon_0)$$

Performance measure

- Excess risk of model $\widehat{h} : \mathcal{X} \to [0,1]$ w.r.t. 0/1 loss:

$$R_T(\widehat{h}) = \mathbb{P}\left(\{y \neq \mathrm{sgn}(\widehat{h}(x) - 1/2)\}\right) - \mathbb{P}\left(\{y \neq \mathrm{sgn}(h^*(x) - 1/2)\}\right)$$

- Want to compute model $\widehat{h}$ s.t.
  - small $R_T(\widehat{h})$
  - small N_T (total no. of queried labels), and
  - small adaptivity
    (as large B as possible, still retaining good performance)
    $\to$ twofold notion:
    1. no. of interactions with labelers
    2. no. of times we retrain model

Diversity measure

Model-based diversity $D(x, S)$:
Quantifies how diverse unlabeled point x is w.r.t. unlabeled set S

- (Generalized) Linear Model

$$D^2(x, S) = x^\top \big(I + \sum_{z \in S} zz^\top\big)^{-1} x$$

- $D(x, S)$ is maximized when $\boldsymbol{x}$ is perpendicular to all $z \in S$

- Generalization to non-linear setting

$$D^2(x, S) = \sup_{f,g \in \mathcal{H}} \frac{(f(x) - g(x))^2}{\sum_{z \in S}(f(z) - g(z))^2 + 1}$$

  c.f. "Fast Rates in Pool-Based Batch Active Learning"
  extended version of this paper:
  `https://arxiv.org/abs/2202.05448`

Generic Algorithm

Stage ℓ:

- $Q_\ell \leftarrow \emptyset$
- Pick $x \in \text{argmax}_{x \in Pool_{\ell-1}} D(x, Q_\ell)$
- $Q_\ell \leftarrow Q_\ell \cup \{x\}$
- repeat until all remaining $x \in Pool_{\ell-1}$ are s.t. $D(x, Q_\ell) \leq 1/2^\ell$
- Query all labels in Q_ℓ and compute predictor $\widehat{h}_\ell$ based on those labels
- Eliminate from $Pool_{\ell-1}$ both queried points and points on which we can confidently predict sign of $h^* - 1/2$:

$$Pool_\ell \leftarrow Pool_{\ell-1} \setminus \{x \in Q_\ell \vee x \in Pool_{\ell-1} : |\widehat{h}_\ell(x) - 1/2| \geq 1/2^\ell\}$$

Stop: when $2^\ell > |Pool_\ell|$

Output: single model $\widehat{h}$ **only** trained on pseudo-labels at each stage

Connection to generalization: Diversity measure $D(x, S)$ should be s.t

$$|\widehat{h}_\ell(x) - h^*(x)| \leq D(x, Q_\ell) \quad \forall x$$

so pseudo-labels are (w.h.p.) accurate

Theoretical guarantees

- Easily adapt to constant batch size B
- W.h.p. excess risk $R_T \lesssim T^{-\frac{\alpha+1}{\alpha+2}}$
- Total no. of labels N_T bounded by $B + T^{\frac{2}{\alpha+2}}$
- Rate of convergence:

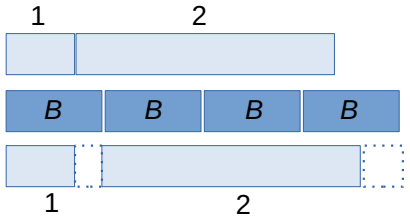
$$R_T(N_T) \approx N_T^{-\frac{\alpha+1}{2}}$$

  minimax rate under VC classes

- Can afford batch size $B \approx T^{\frac{2}{\alpha+2}}$
- No. of stages (i.e., no. of retrainings):

$$\frac{\log T}{\alpha+2} + \log\log T$$

  (w.h.p.)



See you at the poster session!!