Precise Regret Bounds for Log-loss via a Truncated Bayesian Algorithm

Changlong Wu (CSoI, Purdue University)

Joint work with

Mohsen Heidari[1,2], Ananth Grama[1], Wojciech Szpankowski[1]
[1]CSoI, Purdue University [2]Indiana University

NeurIPS2022 Online Talk

Introduction

Let $\mathcal{X}$ be a set of features, $\mathcal{Y} = \{0, 1\}$ be the true label space and $\hat{\mathcal{Y}} = [0, 1]$ be the prediction space.

- **Online learning** is game between Nature and Predictor.
- At each time step t, Nature selects $\mathbf{x}_t \in \mathcal{X}$ and reveals to Predictor.
- Predictor makes prediction $\hat{y}_t \in \hat{\mathcal{Y}}$, based on $\mathbf{x}^t = \{\mathbf{x}_1 \cdots, \mathbf{x}_t\}$ and $y^{t-1} = \{y_1, \cdots, y_{t-1}\}$.
- Nature reveals the true label $y_t \in \mathcal{Y}$, and the Predictor incurs a loss $\ell(\hat{y}_t, y_t)$ where ℓ is the logarithmic loss:

$$\ell(\hat{y}_t, y_t) = -y_t \log(\hat{y}_t) - (1 - y_t) \log(1 - \hat{y}_t).$$

- The game continues upto time T

Introduction

Let $\mathcal{X}$ be a set of features, $\mathcal{Y} = \{0, 1\}$ be the true label space and $\hat{\mathcal{Y}} = [0, 1]$ be the prediction space.

- **Online learning** is game between Nature and Predictor.
- At each time step t, Nature selects $\mathbf{x}_t \in \mathcal{X}$ and reveals to Predictor.
- Predictor makes prediction $\hat{y}_t \in \hat{\mathcal{Y}}$, based on $\mathbf{x}^t = \{\mathbf{x}_1 \cdots, \mathbf{x}_t\}$ and $y^{t-1} = \{y_1, \cdots, y_{t-1}\}$.
- Nature reveals the true label $y_t \in \mathcal{Y}$, and the Predictor incurs a loss $\ell(\hat{y}_t, y_t)$ where ℓ is the logarithmic loss:

$$\ell(\hat{y}_t, y_t) = -y_t \log(\hat{y}_t) - (1 - y_t) \log(1 - \hat{y}_t).$$

- The game continues upto time T

Goal: Minimize the accumulative loss $\sum_{t=1}^{T} \ell(\hat{y}_t, y_t)$.

Introduction

To avoid trivial impossibility result, an additional expert class $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ is introduced.

Introduction

To avoid trivial impossibility result, an additional expert class $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ is introduced.

- The quality of prediction is measured trough (pointwise) regret:

$$R(\hat{y}^T, y^T, \mathcal{H} \mid \mathbf{x}^T) = \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t).$$

Introduction

To avoid trivial impossibility result, an additional expert class $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ is introduced.

- The quality of prediction is measured trough (pointwise) regret:

$$R(\hat{y}^T, y^T, \mathcal{H} \mid \mathbf{x}^T) = \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t).$$

- We are interested in analyzing the sequential minimax regret:

$$r_T^a(\mathcal{H}) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} \sup_{y_1} \cdots \sup_{\mathbf{x}_T} \inf_{\hat{y}_T} \sup_{y_T} R(\hat{y}^T, y^T, \mathcal{H} \mid \mathbf{x}^T).$$

Introduction

To avoid trivial impossibility result, an additional expert class $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ is introduced.

- The quality of prediction is measured trough (pointwise) regret:

$$R(\hat{y}^T, y^T, \mathcal{H} \mid \mathbf{x}^T) = \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t).$$

- We are interested in analyzing the sequential minimax regret:

$$r_T^a(\mathcal{H}) = \sup_{\mathbf{x}_1} \inf_{\hat{y}_1} \sup_{y_1} \cdots \sup_{\mathbf{x}_T} \inf_{\hat{y}_T} \sup_{y_T} R(\hat{y}^T, y^T, \mathcal{H} \mid \mathbf{x}^T).$$

Example: The parametric linear class is defined to be

$$\mathcal{H}^{\text{lin}} = \{h_{\mathbf{w}}(\mathbf{x}) = |\langle \mathbf{w}, \mathbf{x} \rangle| : \mathbf{w}, \mathbf{x} \in \mathbf{R}^d \text{ and } ||\mathbf{w}||_2, ||\mathbf{x}||_2 \leq 1\}.$$

Prior work

- A large body of work in information theory that assumes $\mathbf{x}^T$ is given in advance (a.k.a. *simulatable* case). Completely characterized by the *Shtarkov sum*. [Sht87, Ris84, BRY98, CL01, DS04]
- For *finite* $\mathcal{H}$, we have $r_T^a(\mathcal{H}) \leq \log |\mathcal{H}|$ by *Aggregating Algorithm* [Vol01] (i.e., Bayesian algorithm).
- For *infinite* $\mathcal{H}$, [RS15] showed $r_T^a(\mathcal{H}) = o(T)$ if and only if the *sequential fat shattering* number of $\mathcal{H}$ is finite. But provide only suboptimal bounds, e.g., it gives $r_T^a(\mathcal{H}^{\text{lin}}) \leq \tilde{O}(T^{3/4})$.
- Tighter bound was provided in [BFR20] that improves universally [RS15], e.g., it gives $r_T^a(\mathcal{H}^{\text{lin}}) \leq \tilde{O}(T^{2/3})$. For non-parametric Lipschitz functions, they also provide a matching lower bound. However, the approach is non-constructive.

Our contributions

1. We provide an explicit algorithmic approach that achieves the bound as in [BFR20] with better (optimal) constants.
2. We provide a general approach for deriving lower bounds through the concept of fixed design regret:

$$r^*_T(\mathcal{H} \mid \mathbf{x}^T) = \inf_{\phi^T} \sup_{y^T} R(\phi^T, y^T, \mathcal{H} \mid \mathbf{x}^T).$$

3. Establishes precise regret bounds for specific classes that either improves or provide best bound compare to prior known results, e.g., we have (for $d \geq T$):

$$\Omega(T^{2/3}) \leq r^a_T(\mathcal{H}^{\text{lin}}) \leq \tilde{O}(T^{2/3}).$$

Main Techniques

- **Upper Bounds**: applying Bayesian Averaging over a *Smooth Truncated Sequential* covering set, based on the sequential converging construction as in [RST10] together with the following smooth truncation approach

$$\text{trunc}(g(\mathbf{x})) = \frac{g(\mathbf{x}) + \alpha}{1 + 2\alpha}.$$

- **Lower Bounds**: analyzing the fixed design regret $r_T^*(\mathcal{H} \mid \mathbf{x}^T)$ via the Shtarkov sum, by selecting some *hard* features $\mathbf{x}^T$ that maximize $r_T^*(\mathcal{H} \mid \mathbf{x}^T)$.

Thanks!