Sample Complexity of Forecast Aggregation

Tao Lin, Yiling Chen  Harvard University

Main Question: Sample Complexity

- Oftentimes in practice we have samples from \( P \) (samples of experts’ predictions and the realization of the event):
  \[
  S_T = \{(r_1^{(t)}, \ldots, r_n^{(t)}), \omega^{(t)}\}, \ldots, \{(r_1^{(T)}, \ldots, r_n^{(T)}), \omega^{(T)}\}
  \]
- Can we learn a good aggregator \( f = \hat{f}_S \) from \( S_T \)?
- More specifically,

How many samples do we need to learn an \( \varepsilon \)-optimal aggregator \( f \) with probability at least \( 1 - \delta \)?

Theorem 1 (General Case)

Assume \( |S_T| = m \). The sample complexity of forecast aggregation is:

\[
O \left( \frac{m^n + \log(1/\delta)}{\varepsilon^2} \right) \geq T(\varepsilon, \delta) \geq \Omega \left( \frac{m^{n-2} + \log(1/\delta)}{\varepsilon} \right)
\]

Proof idea 1: Reduction to Distribution Learning

We reduce forecast aggregation to/from the distribution learning problem:
  - given samples from an unknown discrete distribution \( D \), estimate \( D \) within total variation distance \( \varepsilon_{TV} \)
  - has sample complexity \( \Omega \left( \frac{|X| + \log(1/\delta)}{\varepsilon_{TV}^2} \right) \)

Lemma 1 (informal):

\[
\mathbb{E} \left[ \|\hat{f}(r) - f^*(r)\|^2 \right] \leq \varepsilon \Rightarrow \|D - \hat{D}\|_1 \leq O(\sqrt{\varepsilon}) =: \varepsilon_{TV}
\]

Take-Away Message

Forecast aggregation in general is as difficult as distribution learning.

Theorem 2 (Conditional Independence)

If experts’ signals \( s_1, \ldots, s_n \) are independent conditioned on \( \omega \), then:

\[
\hat{f}(\frac{1}{\varepsilon}) \geq T_{\text{cond-ind}}(\varepsilon, \delta) \geq \tilde{O}(\frac{1}{\varepsilon^3})
\]

This is independent of \# of experts and signals!

Proof idea 2: Pseudo-Dimension

- In the cond. ind. case, the optimal aggregator has a simple form: Let \( p = P(\omega = 1) \),
  \[
  f^*(r_1, \ldots, r_n) = \frac{1}{1 + \left( \frac{p}{1-p} \right)^{n-1} \prod_{i=1}^{n} \frac{1-r_i}{r_i}}
  \]
- We prove that the pseudo-dimension of the class of loss functions associated with the aggregators of the form
  \[
  f^*(r_1, \ldots, r_n) = \frac{1}{1 + \theta^{n-1} \prod_{i=1}^{n} \frac{1-r_i}{r_i}}
  \]
  is bounded by \( d = O(1) \).
- This means that the empirically optimal aggregator is \( \varepsilon \)-optimal, if the number of samples is at least
  \[
  O \left( \frac{1}{\varepsilon^2} \right) \geq \frac{d \cdot \log \frac{1}{\varepsilon}}{\varepsilon} = \hat{O}(\frac{1}{\varepsilon^2})
  \]

Future Work

- Close the gap between \( \varepsilon^2 \) and \( \varepsilon \):
  - Conjecture: should be \( \varepsilon \)
  - The case between general distributions and cond. Ind. distributions?
- Recruiting more experts? (Obtaining samples is difficult. Finding more people is easy. Can that help with aggregation?)
- Continuous distributions, other loss functions, etc.