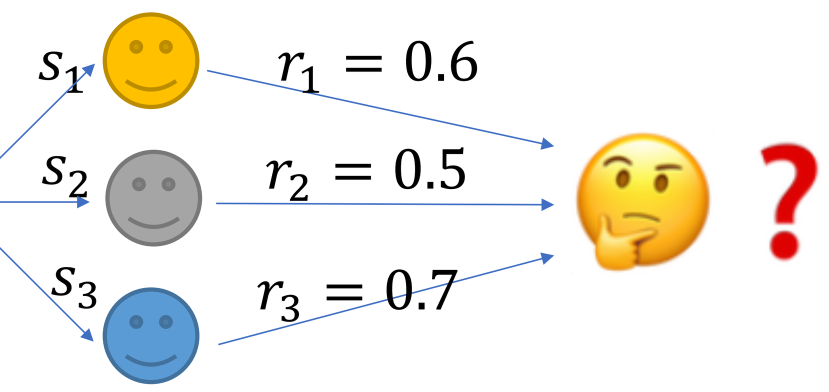
Sample Complexity of Forecast Aggregation

Tao Lin, Yiling Chen   Harvard University

Published at NeurIPS 2023 (spotlight)

Will GDP grow next season?
Will it rain tomorrow?
Will stock price ↗?
Is this image a cat or a dog?

$\omega \in \{$ yes, no $\}$



Background: Forecast Aggregation

- A principal wants to predict an unknown event $\omega \in \{0,1\}$
- He/she collects (probabilistic) predictions from $n \geq 2$ experts: $r_1, \dots, r_n \in [0,1]$
- Q: *How to aggregate these predictions into a single one?*
- $p = f(r_1, \dots, r_n) \in [0,1]$

A common approach in the literature -- **Bayesian model**:

Assume: $(\omega, s_1, \dots, s_n) \sim P$

- $s_i \in S_i$ is a private signal observed by expert i
- Predictions are posterior: $r_i = P(\omega = 1 | s_i)$

Then, the theoretically "optimal" way to aggregate the predictions is the Bayes rule:

$$p^* = f^*(r_1, \dots, r_n) = P(\omega = 1 \,|r_1, \dots, r_n)$$

"optimal": minimizing the squared error $\mathbb{E}[\,|f(\boldsymbol{r}) - \omega|^2]$

But in practice we hardly know P!
(instead, we have samples)

Paper:

Main Question: Sample Complexity

- Oftentimes in practice we have samples from P (samples of experts' predictions and the realization of the event):

$$S_T = \left\{\left(r_1^{(1)}, \dots, r_n^{(1)}, \omega^{(1)}\right), \dots, \left(r_1^{(T)}, \dots, r_n^{(T)}, \omega^{(T)}\right)\right\}$$

- Can we *learn* a good aggregator $\hat{f} = \hat{f}_{S_T}$ from S_T ?
- More specifically,

How many samples do we need to learn an ε-optimal aggregator $\hat{f}$ with probability at least $1 - \delta$?

Theorem 1 (General Case)

Assume $|S_i| = m$. 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 with support X, estimate D within total variation distance $\varepsilon_{\mathrm{TV}}$.
- has sample complexity $\Theta\left(\frac{|X| + \log(1/\delta)}{\varepsilon_{\mathrm{TV}}^2}\right)$

Lemma 1 (informal):

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

Take-Away Message

Forecast aggregation in general is *as difficult as* **distribution learning**.

Theorem 2 (Conditional Independence)

If experts' signals $s_1, \dots, s_n$ are independent conditioned on ω, then:

$$\tilde{O}\left(\frac{1}{\varepsilon^2}\right) \geq T_{\mathrm{cond-ind}}(\varepsilon, \delta) \geq \tilde{\Omega}\left(\frac{1}{\varepsilon}\right)$$

This is independent of the number of experts and signals!

Proof idea 2: Pseudo-Dimension

- In the cond. ind. case, the optimal aggregator has a simple form: Let $p = P(\omega = 1)$,

$$f^*(r_1, \dots, r_n) = \frac{1}{1 + \left(\frac{p}{1-p}\right)^{n-1} \prod_{i=1}^n \frac{1 - r_i}{r_i}}$$

- We prove that the *pseudo-dimension* of the class of loss functions associated with the aggregators of the form $f^\theta(r_1, \dots, r_n) = \frac{1}{1 + \theta^{n-1} \prod_{i=1}^n \frac{1-r}{r_i}}$ is bounded by $d = O(1)$.
- This means that the empirically optimal aggregator is ε-optimal, if the number of samples is at least

$$O\left(\frac{1}{\varepsilon^2}\left(d \cdot \log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right)\right) = \tilde{O}\left(\frac{1}{\varepsilon^2}\right)$$

Future Work

- Close the gap between ε^2 and ε:
  - *Conjecture*: should be ε
- 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.