Active Statistical Inference

By Tijana Zrnic & Emmanuel J. Candès

Problem Setting

- We have a collection of n values $X_1, X_2, \ldots, X_n \sim_{\text{iid}} \mathbb{P}_X$. Each X_i has a corresponding *label* Y_i which is unobserved, with $Y_i \sim \mathbb{P}_{Y|X}$.
- We want to perform inference on some parameter $\theta^\star$ (e.g. the mean label $\mathbb{E}\left[Y_i\right]$) which is a functional of $\mathbb{P}_{(X,Y)} := \mathbb{P}_X \times \mathbb{P}_{Y|X}$.
  - $\theta^\star$ is an element of some parameter space Θ.
  - In particular, the authors focus on hypothesis tests and/or constructing confidence intervals for the parameter $\theta^\star$.
- We have a budget of $n_b \ll n$ labels which we can collect.
  - The goal is to have the *expected number* of collected labels (n_{lab}) be less than n_b
- We have some (often black-box) model f for predicting Y_i given X_i.

Loss Function

- To perform inference on the parameter $\theta^\star$, we utilize a loss function $\ell_\theta(X, Y)$ which is **convex** w.r.t θ.
- Possible examples of this loss are:
  - $\ell_\theta(x, y) = \frac{1}{2}(y - \theta)^2$, target is $\theta^\star = \mathbb{E}\left[Y\right]$
  - $\ell_\theta(x, y) = \frac{1}{2}(y - x^\top\theta)^2$, target is linear regression coefficients
  - $\ell_\theta(x, y) = (1 - q)(\theta - y)\mathbf{1}\{y \leq \theta\} + q(y - \theta)\mathbf{1}\{y > \theta\}$, target is q^{th} quantile of Y for $0 < q < 1$
- For Bayesian Optimization, we want to find $x^\star$ which maximizes $g(x^\star)$ for the unknown function g. In a scenario with zero observational noise, this is effectively the 1.00^{th} quantile of Y.

Batch and Sequential Settings

- The authors propose two versions of their algorithm; the *batch* and *sequential* setting.
- In the *batch* setting, we have a pre-existing predictive model f and simultaneously decide whether or not to acquire Y_i.
  - This version is conceptually easier but reliant on a good model f.
- In the *sequential* setting, we instead acquire Y_i one point at a time and update the predictive model f in accordance with the new data.
  - This is more in line with Bayesian optimization, but from a more frequentist angle.
  - The *sequential* setting also allows us to train a model f "from scratch", which is similar to starting with an uninformative prior for $Y_i \mid X_i$ and doing posterior updates.

Sequential Sampling

- To choose the next label to collect in the *sequential* setting, the authors create a sampling rule $\pi : \mathcal{X} \to [0, 1]$ and collect the label Y_i with probability $\pi(X_i)$.
  - The value of $\pi(X_i)$ is based on the uncertainty of $f(X_i)$, with $\pi(X_i) \approx 1$ when $f(X_i)$ is uncertain and $\pi(X_i) \approx 0$ when $f(X_i)$ has high certainty.
  - The uncertainty is measured as a function $u(\cdot)$, where $\pi(\cdot) \propto u(\cdot)$ is scaled to ensure $\mathbb{E}\left[n_{\mathsf{lab}}\right] \leq n_b$.
- At each step $t \in \{1, 2, \ldots, n\}$, we observe $X_t \in \mathcal{X}$ and collect the label Y_t with probability $\pi_t(X_t)$.
  - $\pi_t(\cdot)$ is a **scaled** measure of the uncertainty of f after collecting the labelled dataset $\{(X_i, Y_i)\}_{i=1}^{n}$.

Sequential Sampling (cont.)

- For general convex M-estimation problems, the sequential estimator is
$$\hat{\theta}^{\pi_n} = \arg\min_{\theta \in \Theta} L^{\pi_n}(\theta)$$
- In the equation above, we define
$$L^{\pi_n}(\theta) = \frac{1}{n} \sum_{t=1}^{n} \left[\ell_{\theta,t}^{f_t} + (\ell_{\theta,t} - \ell_{\theta,t}^{f_t}) \frac{\xi_t}{\pi_t(X_t)} \right]$$
- Under the **Lindeberg condition**, we find that $\hat{\theta}^{\pi_n}$ is asymptotically Normal, and this can be used to construct $1 - \alpha$ confidence intervals for $\theta^\star$.

Similarities to `BayesOpt`

- The *batch* setting is fairly reliant on the existence of a pre-trained model f, but the *sequential* setting has several similarities to `BayesOpt`.
- The overall methodology of the *sequential* setting is quite similar to Bayesian Optimization as the method involves sequential data collection and updating our "beliefs" (the function f_t) in accordance with new data.
- The uncertainty function $u(\cdot)$ **acts** similarly to a utility function used for guiding data collection.
  - The primary difference is that $u(\cdot)$ is mostly reliant on uncertainty regarding the predicted value of $Y_i \mid X_i$ instead of depending on the loss function and/or the target $\theta^\star$.

Differences from `BayesOpt`

- The proposed active inference method is suited for M-estimation problems, and Bayesian optimization does not (really) fall under this category.
- The active inference method involves an initial sample of points $X_1, X_2, \ldots, X_n \sim_{\mathrm{iid}} \mathbb{P}_X$ where we can choose to acquire the corresponding labels, as opposed to `BayesOpt`, which does not need to fix the data points *a priori*.
- The *sequential* setting seems particularly useful when the features of X are categorical/discrete because `BayesOpt` requires a compact space $\mathcal{X}$.