Conditional mutual information and generalization

Last time, we saw (JLNRSS 2019), which showed one technique of proving generalization in the adaptive data analysis setting. There, given a dataset $z \in \mathcal{Z}^n$, one interacts by providing a sequence of statistical queries, $q : \mathcal{Z} \to [0,1]$. There, we were concerned with upper bounding with high probability:

$$\sup_{t \in [T]} |q_t(\mathcal{P}) - a_t|,$$

where $q_t(\mathcal{P})$ is the true answer to the statistical query (i.e. $q_t(\mathcal{P}) = E[q(x)]$ is the expected value) and a_t is the answer that our analysis obtained. The main technique used was to define $\mathcal{Q}_{\mathcal{A}(z)}$ to be the posterior distribution on $\mathcal{Z}^n$ when given the outcomes of the analysis. Then:

$$\sup_{t \in [T]} |q_t(\mathcal{P}^n) - a_t| \leq \underbrace{\sup_{t \in [T]} \left|q_t(\mathcal{P}^n) - q_t(\mathcal{Q}_{\mathcal{A}(z)})\right|}_{\text{posterior insensitivity}} + \underbrace{\sup_{t \in [T]} \left|q_t(\mathcal{Q}_{\mathcal{A}(z)}) - a_t\right|}_{\text{in-sample accuracy}}.$$

Recall that in-sample accuracy could be proved using the Bayesian resampling lemma. One way we can obtain bounds on the posterior insensitivity is through total variation; in particular:

$$\sup_{t \in [T]} \left|q_t(\mathcal{P}^n) - q_t(\mathcal{Q}_{\mathcal{A}(z)})\right| \leq d_{\text{TV}}\left(\mathcal{P}^n, \mathcal{Q}_{\mathcal{A}(z)}\right).$$

Note that if $\mathcal{A}$ is a differentially private mechanism, then we can obtain bounds on the total variation. Intuitively, if $\mathcal{P}^n$ and $\mathcal{Q}_{\mathcal{A}(z)}$ are hard to distinguish, then we learned very little about the particular dataset $z \sim \mathcal{P}^n$ we performed our analysis on from the answers $\mathcal{A}(z)$.

Motivated by this, we'll take a look at Steinke and Zakynthionou's *Reasoning about generalization via conditional mutual information* (SZ 2020). One quantity we may wish to look at in relationship to generalization is then the mutual information between z and $\mathcal{A}(z)$. Recall the definition of mutual information:

Definition 1 (Mutual information)**.** Let X and Y be two random variables jointly distributed according to $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$. The *mutual information* of X and Y is:

$$I(X;Y) = \text{KL}(\mathcal{P}(x,y) \,\|\, \mathcal{P}(x) \times \mathcal{P}(y)).$$

Then, we have the following bound on generalization:

Proposition 2 (Bounded mutual information implies generalization (RZ 2016))**.** *Let* $\ell : \mathcal{W} \times \mathcal{Z} \to [0,1]$*,* $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ *and* $Z \leftarrow \mathcal{P}^n$*. Then:*

$$\left| \mathbb{E}\left[\ell(\mathcal{A}(Z), Z) - \ell(\mathcal{A}(Z), \mathcal{P})\right] \right| \leq \sqrt{\frac{2}{n} \cdot I(\mathcal{A}(Z); Z)}$$

However, mutual information is often unbounded if the domain $\mathcal{Z}$ is infinite, even if generalization is easy to show—as (SZ 2020) write, "the fundamental issue with the mutual information approach is that even a single data point has infinite information content if the distribution is continuous". And so, in their paper, they "normalize" the information content by fixing a sample of size $2n$ beforehand, a procedure not unlike *double/ghost sampling* or *symmetrization.* But then, because we will need to take an expectation over this sample, we define the conditional mutual information:

Definition 3 (Conditional mutual information)**.** For random variables X, Y, Z, the *mutual information of X and Y conditioned on Z* is:

$$I(X;Y|Z) = \mathop{\mathbb{E}}_{z \leftarrow \mathcal{P}_Z} \left[I\left(X|Z=z; Y|Z=z\right) \right].$$

Definition 4 (Conditional mutual information of an algorithm)**.** Let $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ be a randomized or deterministic algorithm. Let $\mathcal{P}$ be a distribution on $\mathcal{Z}$ and let $\tilde{Z} \in \mathcal{Z}^{n\times 2}$ be $2n$ samples drawn independently from $\mathcal{P}$. Let $S_i \in \{0,1\}$ for $i \in [n]$ be i.i.d. uniform at random. Let $\tilde{Z}_S$ be the subset of $\tilde{Z}$ indexed by S. Then, the *conditional mutual information* of $\mathcal{A}$ with respect to $\mathcal{P}$ is:

$$\mathsf{CMI}_{\mathcal{P}}(\mathcal{A}) := I\big(\mathcal{A}(\tilde{Z}_S); S \,|\, \tilde{Z}\big).$$

In short, sample n pairs $(z_0^i, z_1^i) \in \mathcal{Z}^2$ of data and randomly select one sample from each pair to make up Z_S the sample upon which the algorithm learns.

Theorem 5. *Let $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ and $\ell : \mathcal{W} \times \mathcal{Z} \to [0,1]$. Let $\mathcal{P}$ be a distribution on $\mathcal{Z}$ and define $\ell(w, \mathcal{P}) = \mathbb{E}_{Z\leftarrow\mathcal{P}}[\ell(w, Z)]$ and $\ell(w, z) = \frac{1}{n}\sum_{i=1}^n \ell(w, z_i)$. Then:*

$$\left| \mathop{\mathbb{E}}_{Z \leftarrow \mathcal{P}^n} \left[\ell(\mathcal{A}(Z), Z) - \ell(\mathcal{A}(Z) - \mathcal{P}) \right] \right| \le \sqrt{\frac{2}{n} \cdot \mathsf{CMI}_{\mathcal{P}}(\mathcal{A})}.$$

In particular, because the remaining samples $Z_{\overline{S}}$ were independent from Z_S:

$$\mathbb{E}[\ell(\mathcal{A}(Z_S), Z_S) - \ell(\mathcal{A}(Z_S), \mathcal{P})] = \mathop{\mathbb{E}}_{Z,S} \left[\sum_{i=1}^n \ell(\mathcal{A}_S, z_{S_i}) - \ell(\mathcal{A}_S, z_{\overline{S}_i}) \right].$$

Suggestively, for a sample $S' \in \mathcal{Z}^n$, let us write $f_S(S')$ for:

$$f_S(S') = \sum_{i=1}^n \ell(\mathcal{A}_S, z_{S'_i}) - \ell(\mathcal{A}_S, z_{\overline{S}'_i}).$$

Then notice that if S' is independent of S, then $E[f_S(S')] = 0$. In particular, we'd really like to bound:

$$\mathbb{E}[\ell(\mathcal{A}(Z_S), Z_S) - \ell(\mathcal{A}(Z_S), \mathcal{P})] = \mathop{\mathbb{E}}_{Z,S} \left[\mathop{\mathbb{E}}_{S}[f_S(S)] - \mathop{\mathbb{E}}_{S'}[f_S(S')] \right].$$

Consider more closely the expression inside the expectation. Last time, we made use of the following characterization of total variation:

$$d_{\mathrm{TV}}(Q, P) = \sup_{f:\mathcal{X}\to[0,1]} \mathop{\mathbb{E}}_{Q}[f(x)] - \mathop{\mathbb{E}}_{P}[f(x)].$$

This time, *Donsker-Varadhan dual characterization of KL divergence* or the *Gibbs variational principle*:

Theorem 6 (Characterization of KL divergence)**.** *Let P and Q be distributions on Ω with $P \ll Q$ and let $f : \Omega \to \mathbb{R}$ be measurable. Then:*

$$\mathrm{KL}(Q \,\|\, P) = \sup_f \mathop{\mathbb{E}}_{Q}[f(x)] - \log \mathop{\mathbb{E}}_{P}[\exp(f(x))].$$

As a corollary, for any measurable f, we have:

$$\mathop{\mathbb{E}}_{Q}[f(x)] \le \inf_{t>0} \frac{\mathrm{KL}(Q \,\|\, P) + \log \mathbb{E}_P\left[t \exp(f(x))\right]}{t},$$

which just follows from applying the inequality with tf and optimizing over t. Naturally, we will want to convert the $\mathbb{E}_Q[t\exp(fx)]$ term into something easier to work with, so we'll use:

Lemma 7 (Hoeffding)**.** *Let $X \in [a,b]$ be a random variable with mean μ. Then, for all $t \in \mathbb{R}$,*

$$\mathbb{E}[e^{tX}] \leq e^{t\mu + t^2(b-a)^2/8}.$$

Letting Z be fixed, it follows from the definition of mutual information that if Q is a distribution over $(\mathcal{A}(Z_S), S)$ and P is a distribution over $(\mathcal{A}(Z_S), S')$, then:

$$\mathrm{KL}(Q \,\|\, P) = I\big((\mathcal{A}(Z_S), S)\,;\,(\mathcal{A}(Z_S), S')\big).$$