The Statistical Tests

We can cast the second plank in the general context of predictive modeling. Consider a forecaster who places estimates $\{q_i\}$ on the true probabilities $\{p_i\}$ of various events. In the quantum case, the p_i come from distributions in $\mathcal{D}_{1+F}$, where the F that applies to the latter sampling stage can be estimated based on the size and depth of C. The q_i come from the physical quantum device—that is to say, from the strings z that it outputs. What's needed is to compute the corresponding outcome probability q_z analytically based on the given circuit C. This must be done *classically*, and incurs the "T_0-versus-T" issue discussed above. **[See Addendum below.]**

But before we get to that issue, let's say more from the viewpoint of predictive modeling. We measure how well the forecasts q_i conform to the true p_i by applying a prediction scoring **rule**. If outcome i happens, then the *log-likelihood rule* assesses a penalty of

$$L_i = \log(\frac{1}{q_i}).$$

This is zero if the outcome was predicted with certainty but goes to infinity if the individual q_i is very low—which is an issue in the quantum case. The expected score based on the true probabilities is

$$E[L_i] = \sum_i p_i \log(\frac{1}{q_i}). \qquad (2)$$

The log-likelihood rule is **strictly proper** insofar as the unique way to minimize $E[L_i]$ is to set $q_i = p_i$ for each i. In human contexts this means the model has incentive to be as accurate as possible. For the quantum device, knowing the F that applies to its running of circuits C suffices to calculate $E[L_i]$ as "E_{1+F}," and hence to benchmark how accurately the device is conforming to the target.

The formula (**2**) is the **cross-entropy** between the $\vec{p}$ and $\vec{q}$ distributions. It is advocated in several predecessor papers on quantum supremacy experiments, but in fact the team shifted to something simpler they call "linear cross-entropy." They simply show that the q_z from their samples collectively beat the "E_1" that applies to $\mathcal{D}_1$—more simply put, that when summed over T-many trials z_t,

$$\frac{1}{T}\sum_{t=1}^{T} q_{z_t} > \frac{1}{N} + \delta. \qquad (3)$$

This just boils down to giving a **z-score** based on the modeling for $\mathcal{D}_1$. It is analogous to how I (Ken writing this) test for cheating at chess. We are flagging the physical device as getting surreptitious input from quantum to achieve a strength of $1 + \delta$ compared to a "classical player" who is "rated" as having strength 1.

The difference from showing that the device's score from (2) is within a hair of E_{1+F} is that this is based on E_1. To be sure, the paper shows that their z-scores conform to those one would expect an "E_{1+F}-rated" device to achieve. But this is still not the same as (2). Whether it is tantamount for enough purposes—including the theorem about AM—is where we're most unsure, and we note distinctions between fully (classically) sampling and "spoofing" the statistical tests(s) raised by Scott (including directly in reply to me **here**) and others. The authors say that using "linear cross-entropy" gave sharper results and that they tried other (unspecified) measures. We wonder how much of the space of scoring rules familiar in predictive modeling has been tried, and whether rules having more gentle tail behavior for tiny q_i than L_i might do better.

Finally, there is the issue that the team were able to verify q_z exactly only for circuits up to 43 qubits and/or with 14 levels, not 53 with 20 levels. This creates a dilemma in that IBM's paper may push them toward $n = 60$ or 70, but that increases the gap from instance sizes they can verify. This also pushes away from the possibly of observing the $\mathcal{D}_{1+F}$ nature of D_C more directly by finding repeated strings z in the second-stage sampling of a fixed C. The "birthday paradox" threshold for repeats is roughly $2^{n/2}$ samples, which might be feasible for n around 50 (given the classical work needed for each z, which IBM's cleverness might speed) but not above 60. The distinguishing power of repeats drops further with F. We intend to say more about these last few points, and we are sure there are many chapters still to write about supremacy experiments.

…

Addendum 10/28: On further review, the "outcome probability" of a string z comes from first exhaustively computing the probability r_z that would result from error-free operation of C and plugging that in to make $Fr_z + (1-F)\frac{1}{N}$. Although derived from the estimate of F and taking z from the device, this seems better to regard as the "true probability" p_z, rather than "q_z" as stated above. The actual quantity to regard as "q_z" is not calculable and estimating it would require observing repeats from the physical device. Equation (2) remains correct on principle, but as explained in these **notes** by Ryan O'Donnell, the reversed equation is used instead:

$$E[L_i'] = \sum_z q_z \log(\frac{1}{p_z}). \qquad (2')$$

The difference is that $\log(\frac{1}{p_z})$ can be calculated, and while q_z still cannot be, the act of sampling from the physical device estimates the idealized sum $\sum_i q_i \log(\frac{1}{p_i})$ closely enough. This switches the roles of "forecaster" and "forecastee," but the optimality of $q_z = p_z$ remains valid and the target value is the same as before. O'Donnell calls this inversion "slightly dicey" but (i) it was ultimately not used anyway, (ii) has an interpretation that regards the physical device as the ground truth, and (iii) may be equally amenable to asymptotic conditional hardness results. Likewise "q_{z_t}" should be re-named as "p_{z_t}" in (3).]