Ivan Miller - Problem Set 1 for Adversarial AI. Tuesday, Sept. 13, 2022

1. Consider the adversarial game on slide [15] (of lecture 2). What would happen if we set ε greater than 1⁄2? Is there any hope of controlling how far away the adversary can pull the mean estimate of a normal distribution with standard deviation = 1 from its true mean of μ? Explain your reasoning.

The bound on how far away the estimate could get depends on the percentage of the data the adversary gets to corrupt, the distribution of the sample, number of samples, dimension, the way we compute the estimate. We would use median as a robust mean estimate, but even in one dimension, when an adversary gains control of more than half of the sample, there's very little we could do to prevent them from pulling the mean away.

If an adversary is not limited by any bounds of the perturbed values they could choose to replace the original distribution then there's virtually no control of the median. The higher the difference between the replacement and the original values the farther away the median would be pulled. In fact, by selecting the values far away from the original (either larger or smaller) an adversary could change the original distribution so much that an empirical mean would become a more accurate estimator compared to the median.

2. Consider a variant of the game on slide [15] where the adversary can perturb *every data point,* but can only choose a new value that is within plus or minus 1 from the original value. How far do you think this adversary can move the estimated mean of an $N(\mu, 1)$ distribution in this case? Explain your reasoning. (you should think of the number of samples n as being quite large).

An adversary would get more control if in addition to their ability to change every data point by plus or minus one they could sort the data before doing so. Since the data is normally distributed (standard deviation is 1), we should expect half of the distribution to be below average or median and the other half above. In that case an adversary could subtract 1 from each of the original values below the median and add 1 to the original values above the median. By doing so they could theoretically pull the median away by up to 2 (max distance between 1 and -1) divided by the number of data points in the sample. Since the number of samples is assumed to be quite large, as the sample size is growing the effect of the pull is going to be diminishing and could be negligible as the size of the sample grows.

3. Imagine someone gives you a data set of N real numbers and claims it is "Gaussian" distributed. Describe 3 tests you could perform to try to see if this is a plausible claim.

I would start with making a couple of plots of the data since it's probably the easiest way to check for normality (or rule it out): 1) Plot a histogram to check if it is bell-shaped and resembles normal distribution. 2) Create a QQ-Plot to plot the quantiles of our data against theoretical normal quantiles to check if they are aligned.
If the graphical methods are inconclusive, we could try hypothesis testing: Shapiro-Wilk test for samples up to 2,000 values or Kolmogorov-Smirnov for larger samples.