Problem Set 3

[You may discuss these problems with classmates. Please do not troll the internet looking for solutions to these problems. All writing must be done independently, and you must fully understand everything you write.]

1. Chernoff Bound applications. For each question, be sure to specify which Chernoff bound you use:

(a) Suppose there are 80 students enrolled, and the event that each student attends a given lecture are independent, with the expected number of students who attend a given class equal to 50. What is a bound on the probability that fewer than 30 students attend a lecture?

Solution: Let $X_i$ be the event that student $i$ attends the lecture, and let $X = \sum_{i=1}^{n} X_i$. Then

$$
P[X < 30] = P[X < (1 - 2/5)\mathbb{E}[X]] 
\leq \exp\left(-\left(\frac{2/5}{50}\right)^2\cdot \frac{50}{2}\right) 
= 1.8 \times 10^{-2}.
$$

(b) Suppose you want to estimate $\pi$ as follows: draw a (perfectly accurate) circle of radius 1, inside a square of side length 2; choose a set of $n$ independent uniformly random points from within the square, and let $k$ denote the number of points that end up inside the circle. Output “$\pi \approx 4\frac{k}{n}$”. What does $n$ need to be to guarantee that with probability at least 1/2, this estimate is within $\pm0.001$ (additive error)?

Solution: Let $X_i$ be the event that the $i$th point is inside the circle and let $X = \sum_{i=1}^{n} X_i$. Then

$$
P[|4X/n - \pi| > 0.001] = P[|X - \pi n/4| > 0.00025 \cdot (\pi n/4)] 
= P[|X - \pi n/4| > 0.00025 \mathbb{E}[X]] 
\leq 2 \exp\left(-0.00025^2 n \pi/12\right).
$$

We thus need $\exp\left(-0.00025^2 n \pi/12\right) \leq 1/4$, or $n \geq 12 \log(4)/(0.00025^2 \pi) \approx 84,700,000$.

2. In class, we saw a sampling-based randomized algorithm for computing the median of a set $S$ of $n$ (distinct) numbers (see Section 3.4 of Prob. and Comp. to refresh your memory). The algorithm succeeds with high probability (probability $> 1 - \exp(-cn)$ for some constant $c$) and, provided it succeeds, will compare $\frac{3}{2}n + O(n^{3/4} \log n)$ pairs of numbers. There is a different randomized algorithm that many of you might have seen, that resembles quick-sort with a random pivot. In this problem, you will show that this algorithm is significantly worse than the sampling-based algorithm we saw in class. The algorithm is as follows:

- Choose a uniformly random element $x \leftarrow S$, and form the set $S_1 = \{y \in S : y < x\}$ and $S_2 = \{y \in S : y > x\}$ by comparing every element of $S$ to $x$. 


3. Suppose a class has $n$ students. On a given day, assume that if the $i$th student is asked “What is the temperature today”, her response is drawn from a Gaussian (Normal) distribution whose mean is the actual temperature, and variance is $\sigma_i^2$ (and her guess is independent from those of the other students).

(a) If $\sigma_i = 1$ for all $i$, how accurate will the average of the $n$ guesses be? [Feel free to use the fact that the sum of independent Gaussians is Gaussian: namely for $X_i$ distributed according to $N(\mu, \sigma_i^2)$, if the $X_i$ are independent, then $\sum_i X_i$ is distributed according to $N(\sum_i \mu_i, \sum_i \sigma_i^2)$.]  

Solution: We let $X_i$ be the $i$th guess, then $\frac{\sum_{i=1}^n X_i}{n}$ will be Gaussian with variance $1/n$, so the guess will be correct to about $\pm 1/\sqrt{n}$.
(b) If $\sigma_i = 1$ for all $i$, how accurate can we expect the median of the guesses to be? Specifically, find some function $d(n)$ for which with probability at least 0.9, the median will be within distance $d(n)$ from the true temperature.

**Solution:** We claim again that the guess will be within $O(1/\sqrt{n})$. Indeed, let $X_i$ be the event that the $i$th guess is more than $a/\sqrt{n}$ too large, and let $Y_i$ be the event that the $i$th guess is more than $a/\sqrt{n}$ too small. Then $X_i$ and $Y_i$ both have expectation equal to

$$
\int_{a/\sqrt{n}}^{\infty} \exp(-x^2/2\sqrt{2\pi}) dx \leq \frac{1}{2} - \frac{a}{\sqrt{2\pi n}} \cdot \exp(-a^2/2n).
$$

Since we treat $a$ as constant, $\exp(-a^2/2n)$ is close to 1 when $n$ grows, we thus will ignore this term in the future. Let $X = \sum_{i=1}^{n} X_i$ and $Y = \sum_{i=1}^{n} Y_i$. The median is off by more than $a/\sqrt{n}$ if and only if $\max(X,Y) \geq n/2$. But we can bound each of these by the Chernoff bound:

$$
\mathbb{P}[X \geq n/2] = \mathbb{P}[X \geq \frac{n}{n - a/\sqrt{n}} \mathbb{E}[X]] \\
\leq \mathbb{P}[X \geq (1 + a/\sqrt{n})\mathbb{E}[X]] \\
\leq \exp\left(-\frac{a^2/2 \cdot (n/2 - a/\sqrt{n})}{2}\right) \\
= \exp(-a^2/2 + a^3/2\sqrt{n}).
$$

For $n > 16$ and $a = 2$ this is bounded by $1/e$. So $\mathbb{P}[X \geq n/2] \leq \frac{1}{e}$, and hence also $\mathbb{P}[Y \geq n/2] \leq \frac{1}{e}$. Finally, $\mathbb{P}[\max(X,Y) \geq n/2] \leq \frac{2}{e}$, so with constant probability the median is within $2/\sqrt{n}$ of the answer.

(c) Suppose $\sigma_i^2 = 1$ for $i \leq n/2$, and $\sigma_i^2 = n$ for $i > n/2$. Are we better-off returning the mean of the guesses, or the median? Support your claim with a Chernoff bound or two.

**Solution:** It is better to return the median. To see this, first note that the mean is distributed according to a normal distribution with variance $\sum_{i=1}^{n} \frac{\sigma_i^2}{n}$, and so will have error approximately $\sqrt{\sum_{i=1}^{n} \frac{\sigma_i^2}{n^2}}$. On the other hand, if we again let $X_i$ be the event that the $i$th sample exceeds the mean by $a/\sqrt{n}$, then

$$
\mathbb{E}[X_i] = \int_{a/\sqrt{n}}^{\infty} \exp\left(-\frac{x^2}{2\sigma_i^2}\right) \sqrt{2\pi \sigma_i^2} dx \\
\leq \frac{1}{2} - \frac{a}{\sqrt{2\pi n \sigma_i^2}}.
$$

If we let $X = \sum_{i=1}^{n} X_i$ then

$$
\mathbb{E}[X] = \frac{n}{2} - \frac{a}{\sqrt{2\pi n}} \cdot \sum_{i=1}^{n} \frac{1}{\sigma_i} \\
= \frac{n}{2} \left(1 - \frac{a}{\sqrt{8\pi n}} \cdot \sum_{i=1}^{n} \frac{1}{\sigma_i}\right).
$$
Thus, using the same Chernoff bound as before, we see that $a$ will need to grow as
\[\sum_{i=1}^{n} \frac{1}{\sigma_i},\]
and so the error will be approximately
\[\frac{1}{\sqrt{n}} \times \sum_{i=1}^{n} \frac{1}{\sigma_i},\]
which, by the RMS-HM\(^1\) inequality, is always at least as small, and often much smaller than, the error of
\[\frac{1}{\sqrt{n}} \times \sqrt{\sum_{i=1}^{n} \frac{\sigma_i^2}{n}}.\]

Note that one can also get a much tighter bound, showing that the median is within $O(1/\sqrt{n})$ of the true value. The approach to prove this is via a 2-step argument, first showing that the number of large variance guesses that are above/below the true value will be $\pm O(\sqrt{n})$, and given that, with high probability there will be enough of the small variance guesses close to the truth to ‘clean up’ the imbalance in the number of large variance guesses. The intuition is that no matter what the high-variance is, the quality of the median will be determined by the small variance.

(d) BONUS: Consider weighting the $i$th guess by $\frac{1}{\sigma_i} \sum_{j} \frac{1}{\sigma_j}$. Among linear combinations of the guess, this minimizes the variance, and will make the returned estimate distributed according to the gaussian of zero mean and variance $\frac{1}{\sum_{j} 1/\sigma_j}$, and hence one can expect to be within $\pm \frac{1}{\sqrt{\sum_{j} 1/\sigma_j}}$. There are a few ways to see that this is near-optimal. Perhaps the easiest is to argue that information theoretically, letting $G_t$ be the multivariate gaussian whose $i$th coordinate is chosen according to $N(t, \sigma^2_i)$ we have that $G_t$ is the distribution of the $n$ guesses in the case that the true temperature is $t$. Now, we will show that $||G_0 - G_t||_1 \leq c$ for some constant $c < 1$, from which it follows that information theoretically, a sample from these two distributions can not be distinguished with probability greater than $1/2 + c/2$, and hence no estimator can estimate the temperature to within $\pm t/2$ with probability of success greater than this. Letting $t = \frac{1}{\sqrt{\sum_{j} 1/\sigma_j}}$, consider the linear transformation that puts $G_0, G_t$ into isotropic position—namely scaling the $i$th coordinate by a factor of $1/\sigma_i$. Let $G'_0, G'_t$ be the transformed distributions, note that $||G_0 - G_t||_1 = ||G'_0 - G'_t||_1$. To conclude, $G'_0$ and $G'_t$ are $n$ dimensional Gaussians with identity covariance, whose means are separated by $t\sqrt{\sum_{i} \frac{1}{\sigma_i^2}} = 1$, and hence $||G'_0 - G'_t||_1$ is just the $\ell_1$ distance between $N(0, 1)$ and $N(1, 1)$ which is a constant less than 1 ($\approx 0.4$), and no estimator can return a guess of the true temperature that is accurate to within $t/2 = \frac{1}{2\sqrt{\sum_{j} 1/\sigma_j}}$ with probability more than some constant bounded below 1 ($\approx 0.7$).

\(^1\)Root Mean Square-Harmonic Mean inequality.