Consecutive Totient Difference Bounds

> **Problem.** Can we derive a somewhat sharp bound in terms of n for $\varpi(n) := |\varphi(n+1) - \varphi(n)|$? In other words, consider functions $f(n)$ such that
>
> $$|\varphi(n+1) - \varphi(n)| \leq f(n)$$
>
> for all n (or perhaps for all sufficiently large n?). What can we say about such functions, and what related statistics can we measure?
>
> See the MSE link here.

Idea. An obvious trivial (somewhat loose) bound should occur in maximizing one of the totient terms while minimizing the other one. In particular, if we take $n+1$ to be prime, $\varphi(n+1) = n$. Clearly then $\varphi(n) \leq \varphi(n+1)$, so we can immediately recover that

$$|\varphi(n+1) - \varphi(n)| \leq n.$$

This isn't really a formal proof that this holds for all integers, but it should definitely hold and it's probably good enough for now while I'm still just playing around with things.

Idea. We can try considering a worse case scenario. Suppose n is a Mersenne prime, where $n = 2^p - 1$. We then get that

$$\varpi(n) = |2^{p-1} - (2^p - 2)| = 2^{p-1} - 2 = \frac{n}{2} - \frac{3}{2}.$$

Perhaps $\varpi(n)$ should be $O(n)$?

Idea. As the MSE commenter @lulu suggests, we could take a look the average and perhaps look at the statistics/distribution of the differences up to some limit L. We can likely do this in $O(L \log L)$ time, as we can create a lowest prime factor sieve in around $O(L \log L)$ time and then factor and evalute each $\varphi(n)$ in $O(\log n)$ time. This way, we can directly compute the mean, standard deviation, and other statistics in reasonably fast time.

Calculating

$$\frac{1}{L} \sum_{n=1}^{L} \frac{|\varphi(n+1) - \varphi(n)|}{n}$$

for $L = 10^8$ gives a mean of roughly 0.406364 (the standard deviation of the values is roughly 0.214707). Interestingly enough, the mean is close in value to $4/\pi^2$, but this could be a coincidence.

It should be noted that we could obtain the closed form (as a long string of $\phi(k)$'s) if we knew the sign of $\varphi(n+1) - \varphi(n)$, which is equivalent to knowing when $\varphi(n+1)$ is greater than or less than $\varphi(n)$. This allows us to rewrite the expression as:

$$\frac{1}{L} \sum_{n=1}^{L+1} a_n \varphi(n),$$

where $a_1 = -\operatorname{sgn}(\varphi(2) - \varphi(1))/1 = -1$, $a_{L+1} = \operatorname{sgn}(\varphi(L+1) - \varphi(L))/L$, and for $n \neq 1, L+1$,

$$a_n = \frac{\operatorname{sgn}(\varphi(n) - \varphi(n-1))}{n-1} - \frac{\operatorname{sgn}(\varphi(n+1) - \varphi(n))}{n}.$$

An interesting idea would be to estimate the true value by taking these coefficients to be random variables and then using linearity of expectation. To do this, we'd need to obtain information about the probability that $\varphi(n+1) > \varphi(n)$ for n randomly chosen in $\{1, 2, \ldots, L\}$.

Idea. The average difference is cool, but it doesn't exactly get us to where we want for the original problem. I suppose we could try and run a linear regression of some form on the differences, but this isn't exactly what we're after. What we truly want is a regression of some form that guarantees all points are below the regression but also minimizes the average squared distance of each point from the regression (or maximizes the likelihood, but is that even quite well defined in this case?). This is a pretty cool thing to think about even outside the context of the problem, so let's explore it.

Actually, if we take the convex hull of all the points, this bounding line should probably be the extension of one of the top lines if the regression we're doing is linear.

Let's consider a very simple one dimensional linear regression case (you'd probably have to just do any more complicated cases numerically idk). In other words, we have a list of data points $(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)$, and we wish to bound them as tightly above by a function

$$f(x) = \beta x + \alpha.$$

Then, formally, we wish to minimize

$$\sum_{i=1}^{k} (f(x_k) - y_k)^2$$

by choosing optimal β, α, subject to the following inequalities:

$$\begin{aligned} & f(x_i) - y_i \geq 0 \\ \iff\ & \beta x_i + \alpha - y_i \geq 0, \end{aligned}$$

for all i from 1 to k.

In playing around a bit on Desmos and such, we are inclined to make the following claim

Claim. WLOG, assume all points (x_i, y_i) have positive coordinates (we can do so because if the points are not positive, we can just shift them). Then the bounding regression line is