The Challenge Problem from Paul J. Nahin's "When Least is Best": Solved

France Dacar, Jožef Stefan Institute

October 29, 2009

The Challenge Problem

Find an *analytical* derivation of the inequality

$$\int_0^{2\pi} \sqrt{a^2 \sin^2(t) + b^2 \cos^2(t)}\, dt \;\geqslant\; \sqrt{4\pi\big(\pi ab + (a-b)^2\big)}\,, \tag{1}$$

where a and b are any non-negative real numbers.

Where the Challenge Problem comes from

This is the challenge problem as stated by Paul J. Nahin at the end of the section 6.8 of "When Least is Best" (the derivation of the formula for the perimeter of an ellipse is omitted):

> Consider Figure 1, which shows an ellipse (divided into four quarters) with semimajor axes of lengths a and b. The area of this ellipse is given by πab. In Figure 2, the four quarters have been rearranged to form a new figure with

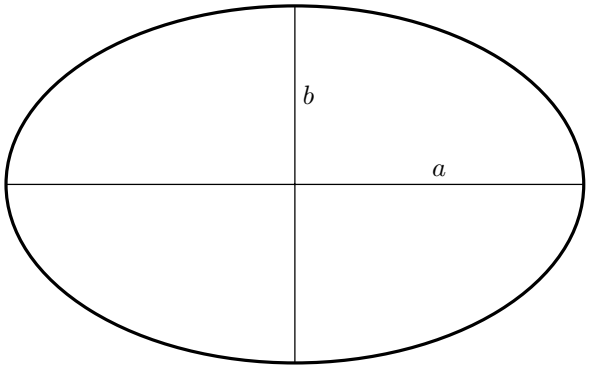

Figure 1: An ellipse.

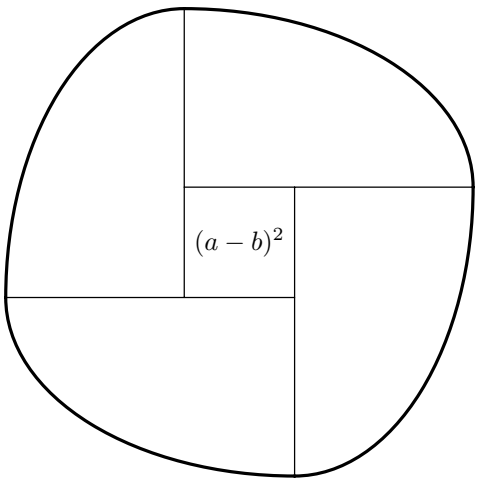

Figure 2: The ellipse of Figure 1 quartered and rearranged (same perimeter, increased area).

> area $\pi ab + (a-b)^2$. The crucial observation about these two figures is that they have the *same perimeter* (I'll call it P), given by

$$P \;=\; \int_0^{2\pi} \sqrt{a^2 \sin^2(t) + b^2 \cos^2(t)}\, dt\,.$$

Now, the isoperimetric theorem says that the area of a plane region with a perimeter $P = 2\pi R$ cannot exceed the area of a circle with radius R. That is,

$$A \leqslant \pi R^2 = \pi\left(\frac{P}{2\pi}\right)^2 = \frac{P^2}{4\pi}.$$

Thus, $P \geqslant \sqrt{4\pi A}$, and so, using the area of Figure 2 for A, we have

$$\int_0^{2\pi} \sqrt{a^2\sin^2(t) + b^2\cos^2(t)}\, dt \geqslant \sqrt{4\pi\big(\pi ab + (a-b)^2\big)}\,.$$

where the equality certainly holds when $a = b$.

Here's the challenge — there seems to be no 'easy' way to derive this inequality *directly*, by manipulating the integral on the left-hand side. That is, *I* can't see how to do it. If you try your hand at it and succeed, please write to me and tell me how you did it!

Some preparatory massaging

The integrand on the left hand side of inequality (1) is a periodic function with a period π, hence the integral from 0 to π is precisely one half of the integral from 0 to 2π, and so we may halve both sides and get an equivalent inequality

$$\int_0^{\pi} \sqrt{a^2\sin^2(t) + b^2\cos^2(t)}\, dt \geqslant \sqrt{\pi\big(\pi ab + (a-b)^2\big)}\,. \tag{2}$$

The right hand side of this inequality is symmetric in a and b. So is the left hand side: if we shift the interval of integration by $-\pi/2$ and introduce new integration variable $t + \pi/2$, then $\cos^2$ and $\sin^2$ change places. We can therefore assume that $a \leqslant b$. If $a = b = 0$, then both sides of the inequality are 0, so the inequality certainly holds in this trivial case, and we may assume, from now on, that $b > 0$.

The integrand on the left hand side of (2) is symmetric with respect to the midpoint $\pi/2$ of the interval of integration, so we may once more halve boths sides of the inequality, to obtain

$$\int_0^{\pi/2} \sqrt{a^2\sin^2(t) + b^2\cos^2(t)}\, dt \geqslant \frac{1}{2}\sqrt{\pi\big(\pi ab + (a-b)^2\big)}\,. \tag{3}$$

Both sides of this inequality are homogenous of degree 1 in a and b: if we replace the parameters a, b with ua, ub, where the multiplier u is any nonnegative real number, each of the two sides is multiplied by u. We choose $u = 1/b$, write $x = a/b$, and obtain yet another inequality equivalent to the original inequality:

$$\int_0^{\pi/2} \sqrt{x^2\sin^2(t) + \cos^2(t)}\, dt \geqslant \frac{1}{2}\sqrt{\pi\big(\pi x + (1-x)^2\big)}\,, \qquad 0 \leqslant x \leqslant 1\,. \tag{4}$$

It is this final form of the challenge inequality which we shall prove. *Analytically.*

The first attempt, which fails

The idea of our analytical derivation is to replace the integrand in (4), which cannot be decently integrated, by a function lying below the integrand whose integral we *can* calculate and then show it to be greater or equal to the right hand side of (4).

But before we embark on the actual work, we introduce some shorthands and agree on some conventions. We shall write $f(t)$ for the integrand and I for the integral on the left hand

side of the inequality (4), and B for the lower bound on its right hand side; this inequality therefore reads simply $I \geqslant B$. Functions approximating $f(t)$ from below shall be denoted by $g_1(t)$, $g_2(t)$, $g_3(t)$, ..., and their integrals (from 0 to $\pi/2$) by J_1, J_2, J_3, ...

The integrand $f(t)$ and the integral I depend on the parameter x, as does the lower bound B, so we should by rights write them as, say, $f_x(t)$, I_x, and B_x (or $B(x)$). Likewise, any approximating function will depend on the parameter x, and possibly on some other parameter(s). Nevertheless, we shall mostly supress the urge to display parameters, because otherwise we would hardly be able to see our formulas for the thicket of parameters sprouting all over them in subscript and/or superscript positions. Instead we shall simply remember, about each particular animal under discussion, on which parameters it happens to depend.

And here begins our first attempt. It will fail, as announced in the section's title, but at the same time it will point out a promising direction of attack on the problem.

First let us take a look at the diagrams of functions $f(t)$ and $f(t)^2$ when $x = \frac{1}{3}$ (Figure 3).

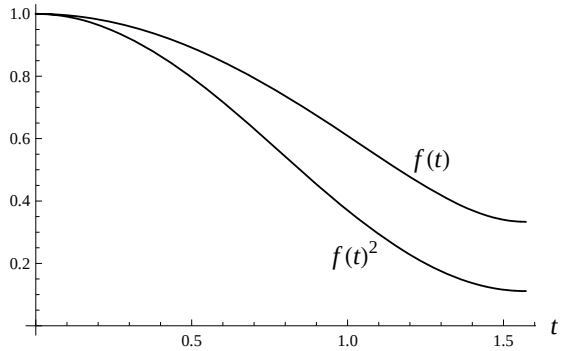

Figure 3: Functions $f(t)$ and $f(t)^2$, for $x = \frac{1}{3}$.

The diagram of $f(t)^2$ is a vertically scaled and shifted cosine wave $\cos(2t)$:

$$f(t)^2 \;=\; x^2 \cdot \tfrac{1}{2}\big(1 - \cos(2t)\big) + \tfrac{1}{2}\big(1 + \cos(2t)\big) \;=\; \tfrac{1}{2}(1 + x^2) + \tfrac{1}{2}(1 - x^2)\cos(2t)\,.$$

We get $f(t)$ by taking the square root of $f(t)^2$, which deforms the cosine wave by scaling it vertically depending on the ordinate. For a tiny segment of the diagram of $f(t)^2$ around $f(t)^2 = y$ the scaling factor is $\big(\sqrt{y}\big)' = 1/\big(2\sqrt{y}\big)$, which decreases when y increases, and that means that taking the square root squashes higher parts of the cosine wave $f(t)^2$ in comparison with its lower parts. Now, if we vertically scale and shift the undeformed cosine wave $\cos(2t)$ so as to obtain a function $g_1(t)$ with the same values at $t = 0$ and $t = \pi/2$ as $f(t)$,

$$g_1(t) \;=\; \tfrac{1}{2}(1 + x) + \tfrac{1}{2}(1 - x)\cos(2t)\,, \tag{5}$$

then the diagram of $g_1(t)$ should lie below the diagram of $f(t)$ — and so it does, as a glance at Figure 4 convinces us. Since a glance at a convincing figure does not constitute a proof, we compute the difference $f(t)^2 - g_1(t)^2$ to see if it is always non-negative:

$$f(t)^2 - g_1(t)^2 \;=\; \tfrac{1}{4}(1 - x)^2\big(1 - \cos^2(2t)\big) \;\geqslant\; 0\,.$$

It is, and the diagram of $g_1(t)$ does lie below the diagram of $f(t)$. Actually, it lies somewhat too low, so that the gap between the two diagrams is ominously large. Still we push on, since we want to see how good, or bad, is our approximation $g_1(t)$ of $f(t)$. The integral of $g_1(t)$ is

$$J_1 \;=\; \tfrac{1}{4}\pi(1 + x)\,.$$

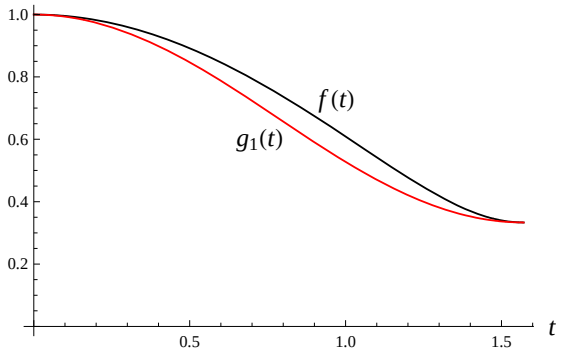

Figure 4: $g_1(t) \leqslant f(t)$ for $0 \leqslant t \leqslant \pi/2$.

To compare J_1 with the lower bound B of I, we compute the difference

$$J_1^2 - B^2 \;=\; -\tfrac{1}{16}\pi(4-\pi)(1-x)^2\,,$$

and find that it is strictly negative for $0 \leqslant x < 1$ and is zero at $x = 1$. How bad is this? If we compare $J_1 - B$ with $I - B$ (Figure 5), we see that $J_1 - B$ is negative and almost the opposite

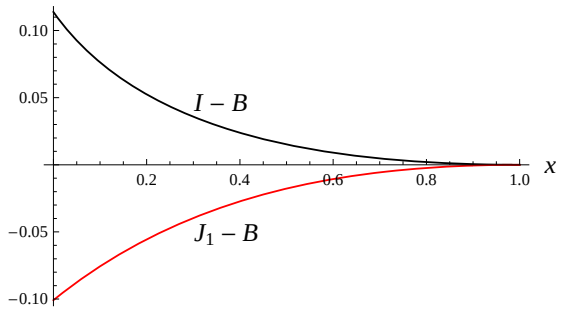

Figure 5: $J_1 - B$ compared to $I - B$.

of $I - B$, while we wanted it to be positive and less than $I - B$.

So our first attempt is a total, wide miss.

Success!

What now? We can try to improve the lower bound $g_1(t)$ of $f(t)$, by increasing it to $g_2(t)$ in such a way that we will still manage to compute the integral J_2 of $g_2(t)$, with the resulting expression simple enough, so that we will be able to decide whether it is greater than B or not.

Let us try this. We pick α in the interval $0 < \alpha < \pi/2$, grab the diagram of $g_1(t)$ at the point $\big(\alpha, g_1(\alpha)\big)$ and slide it along the vertical $t = \alpha$ upwards to the point $\big(\alpha, f(\alpha)\big)$, deforming the diagram of $g_1(t)$ into the diagram of a function $g_2(t)$ (Figure 6). During the deformation the endpoints $(0,1)$ and $(1,x)$ of the diagram of $g_1(t)$ stay fixed, and the two arcs of the diagram, from $(0,1)$ to $\big(\alpha, g_1(\alpha)\big)$ and from $\big(\alpha, g_1(\alpha)\big)$ to $(1,x)$, are getting uniformly scaled in the vertical direction. Describing $g_2(t)$ by a formula, we have

$$g_2(t) \;=\; \begin{cases} \dfrac{f(\alpha) - \cos(2\alpha) + \big(1 - f(\alpha)\big)\cos(2t)}{1 - \cos(2\alpha)} & \text{if } 0 \leqslant t \leqslant \alpha\,, \\[2ex] \dfrac{f(\alpha) + x\cos(2\alpha) + \big(f(\alpha) - x\big)\cos(2t)}{1 + \cos(2\alpha)} & \text{if } \alpha \leqslant t \leqslant \pi/2\,. \end{cases} \tag{6}$$

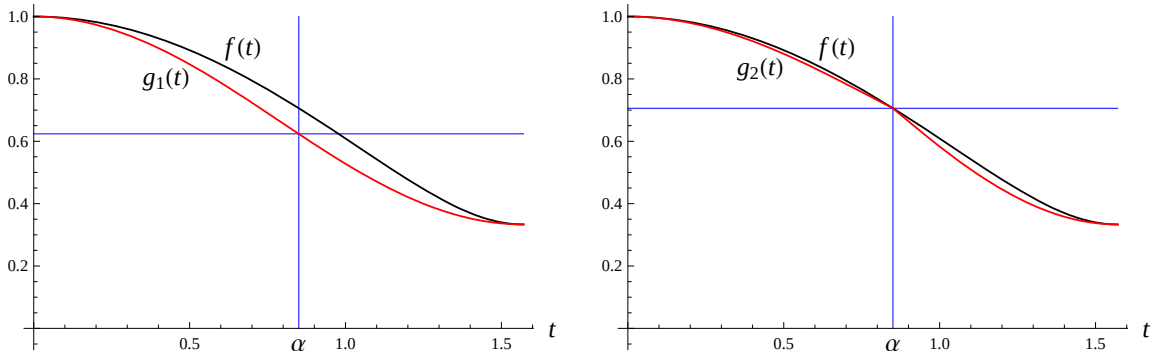

Figure 6: Increasing the lower bound $g_1(t)$ of $f(t)$ to a better lower bound $g_2(t)$.

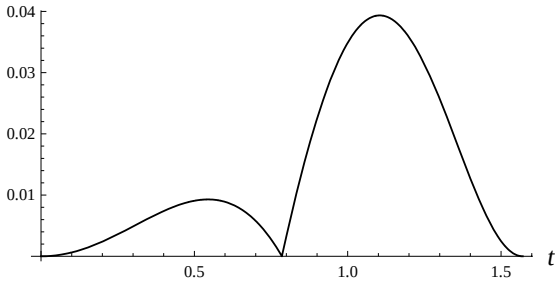

Figure 7: The difference $f(t) - g_2(t)$, for $\alpha = \pi/4$ and $x = \frac{1}{3}$.

The important question here is, of course, is $g_2(t)$ still below $f(t)$ for all t? The plot of the difference $f(t) - g_2(t)$, for $\alpha = \pi/4$ and $x = \frac{1}{3}$, is encouraging (Figure 7). Plotting the difference for various values of x and α we invariably get a diagram entirely above the t-axis. To actually prove that always $g_2(t) \leqslant f(t)$, we could proceed just as we did with $g_1(t) \leqslant f(t)$, namely calculate the difference $f(t)^2 - g_2(t)^2$, factor it (separately for each of the two intervals $0 \leqslant t \leqslant \alpha$ and $\alpha \leqslant t \leqslant \pi/2$), and then find, by examining the factors, that this difference is in fact always non-negative; this *would* work, though rather messily. However, there seems to be some more general principle at work here; if we can discover what it is, we may be able to take a shortcut and conclude straightaway that $g_2(t) \leqslant f(t)$. But of course: the square root is a concave function! We have already exploited the concavity of the square root in our 'intutive reasoning' about why the diagram of $g_1(t)$ must lie below the diagram of $f(t)$.

Here is how it goes. Let Y be a nonempty interval of real numbers (which may be bounded, or may be unbounded in one or both directions), and let φ be a real-valued convex function defined on Y. Let $a < b$ be real numbers, and let $h(t)$ be a function defined for $a \leqslant t \leqslant b$ and taking values in Y; moreover, suppose that $h(a) \neq h(b)$ and that $h(t)$ lies between $h(a)$ and $h(b)$ for $a \leqslant t \leqslant b$.

Concavity of $\varphi(t)$ means that its diagram bulges above any chord connecting two of its points; spelled out formally, $\varphi(t)$ satisfies the condition

$$\varphi\big((1-\lambda)y + \lambda z\big) \;\geqslant\; (1-\lambda)\varphi(y) + \lambda\varphi(z) \tag{7}$$

for all y and z in Y and all real numbers λ in the unit interval $0 \leqslant \lambda \leqslant 1$. For any t (in the interval on which h is defined) we can write

$$h(t) \;=\; (1-\lambda_t)h(a) + \lambda_t h(b)\,, \tag{8}$$

where the coefficients λ_t and $1 - \lambda_t$ are

$$\lambda_t \;=\; \frac{h(t) - h(a)}{h(b) - h(a)}\,, \qquad 1 - \lambda_t \;=\; \frac{h(b) - h(t)}{h(b) - h(a)}\,,$$

and $0 \leqslant \lambda_t \leqslant 1$. Using (7) on (8), we get the inequality

$$\varphi\big(h(t)\big) \;\geqslant\; (1-\lambda_t)\varphi\big(h(a)\big) + \lambda_t \varphi\big(h(b)\big) \;=:\; h_\varphi(t)\,; \tag{9}$$

which is an equality when $t = a$ or $t = b$,

$$h_\varphi(a) \;=\; \varphi\big(h(a)\big)\,, \qquad h_\varphi(b) \;=\; \varphi\big(h(b)\big)\,. \tag{10}$$

The function $h_\varphi(t)$ has the form

$$h_\varphi(t) \;=\; u + v h(t)\,, \tag{11}$$

where the coefficients u and v are constants (i.e. they do not depend on t), given by

$$u \;=\; \frac{h(b)\varphi\big(h(a)\big) - h(a)\varphi\big(h(b)\big)}{h(b) - h(a)}\,, \qquad v \;=\; \frac{\varphi\big(h(b)\big) - \varphi\big(h(a)\big)}{h(b) - h(a)}\,; \tag{12}$$

that is, the diagram of $h_\varphi(t)$ is obtained by vertically scaling and shifting the diagram of $h(t)$. The diagram of $h_\varphi(t)$ lies below the diagram of $\varphi\big(h(t)\big)$, and has the same endpoints with it. Note that if $h(t) = c + dk(t)$, where c and d are constants, and $k(t)$ has the property we have required of $h(t)$ ($k(a) \neq k(b)$ and $k(t)$ is between $k(a)$ and $k(b)$ for $a \leqslant t \leqslant b$), then

$$h_\varphi(t) \;=\; (u + cv) + (dv)k(t) \;=\; u_1 + v_1 k(t)\,, \tag{13}$$

where u_1 and v_1 are the constants

$$u_1 \;=\; \frac{k(b)\varphi\big(h(a)\big) - k(a)\varphi\big(h(b)\big)}{k(b) - k(a)}\,, \qquad v_1 \;=\; \frac{\varphi\big(h(b)\big) - \varphi\big(h(a)\big)}{k(b) - k(a)}\,. \tag{14}$$

Formulas (13) and (14) give a valid 'subterpolated' function $h_\varphi(t)$ even in the degenerate case when $d = 0$ and $h(t)$ is the constant c, since then $h_\varphi(t)$ is that same constant.

Taking $h(t) = f(t)^2$, $\varphi(y) = \sqrt{y}$, $k(t) = \cos(2t)$, and $0 \leqslant a < b \leqslant \pi/2$, we see that

$$f(t) \;\geqslant\; \frac{\cos(2b)f(a) - \cos(2a)f(b)}{\cos(2b) - \cos(2a)} \;+\; \frac{f(b) - f(a)}{\cos(2b) - \cos(2a)}\cos(2t) \qquad \text{for } a \leqslant t \leqslant b\,. \tag{15}$$

The desired inequality $f(t) \geqslant g_2(t)$, for $0 \leqslant t \leqslant \pi/2$, is now an immediate consequence.

At this point we are tempted (but only for a moment or two), to compute the integral J_2 of $g_2(t)$ and then find $\alpha = \alpha_{\text{opt}}$ (which depends on the parameter x) that maximizes the integral. If we try to carry out this plan, we soon find ourselves wading knee-deep in messy formulas while trying to solve an equally messy transcendental equation. The dependences of α_{opt} and the maximal value of J_2 on the parameter x are far from simple, and our chances of proving that the maximal value of J_2 is above B, by directly manipulating some nice formulas obtained from optimization, are nil. (We shall soon be able to conclude that the maximal value of J_2 is above B, by showing that the value of J_2 for a particular value of α is above B.) This is one of those occasions when "keep it simple" is the best 'optimization policy'.

So this is what we do — we choose $\alpha = \pi/4$, and compute the integral J_2 of the corresponding 'subterpolated' function $g_2(t)$ of $f(t)$:

$$J_2 \;=\; \frac{1}{2}\left(1 + x + \sqrt{2}\left(\frac{\pi}{2} - 1\right)\sqrt{1 + x^2}\right). \tag{16}$$

With slightly trembling fingers we type into MATHEMATICA the request to plot the difference $J_2 - B$ as a function of the parameter x, and... there it is (Figure 8), and *it is non-negative*!

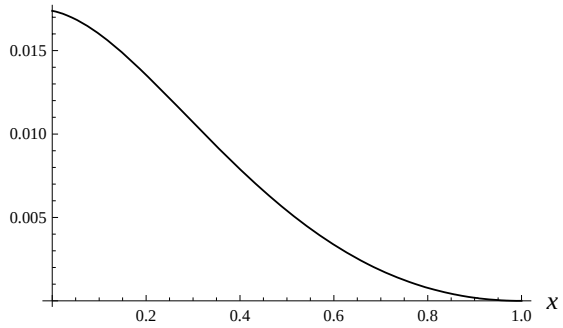

Figure 8: The difference $J_2 - B$ (for $\alpha = \pi/4$) as a function of x.

That is, it appears to be non-negative, and all it remains to do is to prove that it really is non-negative. We start with the inequality $2J_2 \geqslant 2B$,

$$1 + x + \sqrt{2}\left(\frac{\pi}{2} - 1\right)\sqrt{1+x^2} \;\geqslant\; \sqrt{\pi\big(\pi x + (1-x)^2\big)}\,, \tag{17}$$

which we are going to square and rearrange, then square and rearrange again to get the right hand side down to zero, and finally factor the left hand side. This final left hand side will be — let's see — a biquadratic polynomial in x; looking at the diagram of $J_2 - B$ in Figure 8 we have a strong hunch that the factor $(1-x)^2$ will appear in the factorization, so the other factor will be quadratic, hence easy to handle. Now we proceed with our plan. We square (17), multiply by 2, rearrange:

$$2\sqrt{2}(\pi-2)(1+x)\sqrt{1+x^2} \;\geqslant\; (-6{+}6\pi{-}\pi^2) + (-4{-}4\pi{+}2\pi^2)x + (-6{+}6\pi{-}\pi^2)x^2\,. \tag{18}$$

The numerical values, to five decimal places, of the coefficients of the quadratic polynomial on the right hand side are 2.97995, 3.17284, and 2.97995. Once more we square (18), then rearrange and factor, and obtain the inequality

$$(1-x)^2(a_0 + a_1 x + a_2 x^2) \;\geqslant\; 0\,, \tag{19}$$

where the coefficients of the second quadratic factor are

$$\begin{aligned} a_0 &= -4 + 40\pi - 40\pi^2 + 12\pi^3 - \pi^4 \;\doteq\; 1.54576\,, \\ a_1 &= 8 + 16\pi - 8\pi^3 + 2\pi^4 \;\doteq\; 5.03345\,, \\ a_2 &= a_0\,, \end{aligned} \tag{20}$$

from which it is clear that inequality (19) holds for $0 \leqslant x \leqslant 1$; since this inequality implies the inequality $J_2 \geqslant B$ (actually, all inequalities in our derivation are equivalent to each other for any $x \geqslant 0$), we are done.