MAA 4212, Spring 2009—Miscellaneous non-book problems

F1. **Picard iteration.** Recall that our proof of existence and uniqueness of solutions to the initial value problem

$$\frac{dy}{dt} = f(t,y), \quad y(t_0) = y_0 \tag{1}$$

used the Contracting Mapping Fixed-Point Theorem (CMT), which was itself proved by looking at the sequence of iterates of a point under a contraction defined on some complete metric space. In the FTODE, the metric space was a closed ball in $C(\overline{I})$ for some closed interval $\overline{I}$, and the contraction was the map H defined by

$$H(g)(t) = y_0 + \int_{t_0}^{t} f(s, g(s))\, ds.$$

Recall that the CMT gives us not just existence and uniqueness of a fixed point, but a way of constructing the fixed point: start with *any* point in the metric space, and follow the sequence of points obtained by repeatedly applying the contraction. Carrying out this procedure in the context of the FTODE is called *Picard iteration.* We start with a function g_0 (usually the constant function $t \mapsto y_0$) defined on some neighborhood of t_0, define $g_1 = H(g_0), g_2 = H(g_1)$, etc. The proof of the FTODE (via the CMT) shows that if we take δ small enough, (1) will have a unique solution on $(t_0 - \delta, t_0 + \delta)$, and the sequence $\{g_n\}$ will converge uniformly to this solution on that interval. Thus, Picard iteration gives us a (not necessarily efficient) way to produce the solution of (1). Of course, for some f we can solve (1) explicitly, in closed form, rather than express the solution as the limit of some sequence; that's what you did in the first few weeks of MAP 2302. For such ODEs, there is no point to doing Picard Iteration other than for fun, to see what happens, or as an exercise in learning. (Even for ODEs that we can't solve by MAP 2302 methods, there are usually much more efficient ways of computing solutions than to use Picard Iteration.)

While the *proof* of the FTODE requires us to choose δ sufficiently small, in practice when we do Picard Iteration we don't worry about how small δ needed to be for that proof, just how small it needs to be for the sequence we actually produce to converge.

Try Picard Iteration for the following IVPs. In each case, take g_0 to be the constant function with value y_0.

(a) $\frac{dy}{dt} = y, \quad y(0) = 1$. Find a formula for $g_n(t)$. If you do this right, what you find should be a sequence that converges on the whole real line, uniformly on any bounded interval. Where have you seen this sequence $\{g_n\}$ before?

The IVP in (a) is very, very special. Ordinarily the phenomenon that you saw (or should have seen) in (a) does not occur. For example:

(b) $\frac{dy}{dt} = y^2, \quad y(0) = 1$. Find a "semi-explicit" formula for $g_n(t)$, of this form:

$$g_n(t) = p_n(t) + q_n(t), \tag{2}$$

where p_n is a polynomial of degree n that you give an explicit formula for, and q_n is a polynomial whose terms have degree ranging from $n+1$ to 2^n-1. (You're not expected to find a formula for q_n, but for fun you might try to find a formula for the coefficient of t^{n+1}.) If you do this right, you should be able to show explicitly (not by recourse to the FTODE or CMT) that the more manageable sequence $\{p_n\}$ converges pointise on $(-1,1)$ (uniformly on compact subsets) to the solution you could have found in MAP 2302, $t \mapsto \frac{1}{1-t}$. (The proof of the FTODE shows indirectly that the messier series $\{g_n\}$ converges uniformly to this function on $(-\delta,\delta)$ for δ some sufficiently small. Hence the sequence $\{q_n\}$ converges uniformly to 0 on this interval, but that's not obvious from (2).)

(c) $\frac{dy}{dt} = e^y,\ \ y(0) = 0$. The solution of this IVP is easily found by MAP 2302 methods: $y(t) = -\log(1-t)$. See just how bad Picard Iteration is for this example by computing as many g_n as you can in closed form, and seeing how soon you get stuck.

(d) $\frac{dy}{dt} = t^2+y^2,\ \ y(0) = 0$. This one is just to give you an example in which $f(t,y)$ has some t-dependence (in (a)-(c) f depended only on y), and also an example in which you cannot find a closed-form solution by any methods you learned in MAP 2302. Compute g_1 and g_2 explicitly, and show by induction that each $g_n(t)$ is a polynomial whose coefficient of t^m is 0 for all m not congruent to 3 mod 4. (Thus we can write $g_n(t) = t^3 h_n(t^4)$ for some polynomial h_n.)

F2. **Estimating values of certain series.** Given a convergent series whose value we don't know how to compute *exactly*, it's of interest to know how to get a good estimate of the sum. Broadly speaking, we'd like to get as close as we can to the true value of the sum (and know a bound on how far off we might be), while doing as little computation as possible. In this problem, you will compare two estimation schemes for series to which the Integral Test applies.

The proof of the Integral Test (but not simply the *statement* of the test in p. 161/9) gives an important error-bound. This error bound, stated below, comes from the following: if f is a monotone-decreasing function on $[N,\infty)$, where $N \in \mathbf{N}$, and $\int_N^\infty f(x)\,dx$ converges, then $\sum_{n=N+1}^\infty$ converges and

$$\int_{N+1}^\infty f(x)\,dx \le \sum_{n=N+1}^{\infty} f(n) \le \int_N^\infty f(x)\,dx. \tag{3}$$

(a) Prove the double-inequality (3). (You may have already done this when you did p. 161/9; in that case, do it again.)

(b) Let f be monotone-decreasing function on $[1,\infty)$, let $a_n = f(n)$ for $n \in \mathbf{N}$, and for $N \in \mathbf{N}$ let $s_N(\vec{a}) = \sum_{n=1}^N a_n$. Let $E_N(\vec{a}) = (\sum_{n=1}^\infty a_n) - s_N(\vec{a})$. Up to sign (a matter of convention), $E_N(\vec{a})$ is the *error in estimating* $\sum_{n=1}^\infty a_n$ *by* $s_N(\vec{a})$. Since this error can also be written as $\sum_{n=N+1}^\infty a_n$, (3) implies that

$$\int_{N+1}^\infty f(x)\,dx \le E_N(\vec{a}) \le \int_N^\infty f(x)\,dx. \tag{4}$$

Because the left-most integral is ≥ 0, the right-hand inequality gives us an upper bound on the magnitude of the error, namely

$$|E_N(\vec{a})| = \left| \left(\sum_{n=1}^{\infty} a_n \right) - s_N(\vec{a}) \right| \leq \int_N^{\infty} f(x)\ dx. \tag{5}$$

But we have thrown away valuable information: the left-hand inequality in (4) gives us a *lower* bound on the error—i.e. it tells us that if we estimate $\sum_{n=1}^{\infty} a_n$ by $s_N(\vec{a})$, then our estimate will be too small by *at least* a certain amount. We can use this fact to get a sharper estimate without doing significantly more computation.

For f and $\{a_n\}$ as above, define

$$\tilde{s}_N(\vec{a}) = s_N(\vec{a}) + \frac{1}{2}\left(\int_{N+1}^{\infty} f(x)\ dx + \int_N^{\infty} f(x)\ dx \right).$$

In other words, we add to $s_N(\vec{a})$ the average of the upper and lower bounds on the error $E_N(\vec{a})$ given by (4). The error in estimating $\sum_{n=1}^{\infty} a_n$ by $\tilde{s}_N(\vec{a})$ (modulo sign-convention, again) is then $\tilde{E}_N(\vec{a}) := (\sum_{n=1}^{\infty} a_n) - \tilde{s}_N(\vec{a})$.

Show that

$$\left| \tilde{E}_N(\vec{a}) \right| \leq \frac{1}{2} a_N. \tag{6}$$

This error-bound can be phrased qualitatively as "when the series is estimated this way, the error is at most one-half the last term in the partial sum".

(c) To get a sense of how much sharper the estimate $\tilde{s}_N(\vec{a})$ is than is $s_N(\vec{a})$, compare the two error-bounds for $\sum_{n=1}^{\infty} \frac{1}{n^2}$ as follows. For $N \in \mathbf{N}$ let $\mathrm{erbd}(N), \widetilde{\mathrm{erbd}}(N)$ denote the error-bounds given by the right-hand sides of (4), (6) respectively (with $f(x) = \frac{1}{x^2}$). For $\epsilon > 0$, let $N(\epsilon), \tilde{N}(\epsilon)$ denote the smallest values of N for which $\mathrm{erbd}(N), \widetilde{\mathrm{erbd}}(N)$, respectively, is $\leq \epsilon$.

(i) Compute $\mathrm{erbd}(N), \widetilde{\mathrm{erbd}}(N)$ and show that $\lim_{N\to\infty} \frac{\widetilde{\mathrm{erbd}}(N)}{\mathrm{erbd}(N)} = 0$. Interpret this in terms of the accuracy-gain for "worst-case scenarios" of the two estimation schemes when you truncate the sum after the N^{th} term. ("Worst-case scenario" means that the *actual* error equals the upper bound on the error given by the estimation scheme.)

(ii) Compute $N(\epsilon), \tilde{N}(\epsilon)$ and show that $\lim_{\epsilon\to 0} \frac{\tilde{N}(\epsilon)}{N(\epsilon)} = 0$. Interpret this in terms of how much work you have to do, in the two estimation schemes, to get the same degree of guaranteed accuracy.

(d) For $\sum_{n=1}^{\infty} \frac{1}{n^2}$ compute $N(.1), \tilde{N}(.1), N(.005), \tilde{N}(.005)$. If you've done problem D5e, then, using your calculator (bet you never thought I'd say that!), find $s_3, \tilde{s}_3$, and $\sum_{n=1}^{\infty} \frac{1}{n^2}$ (the latter by your formula from D5e), and use these to compute the *actual* errors $E_3, \tilde{E}_3$ and the ratio $\tilde{E}_3/E_3$. (The last computation should exhibit that the accuracy-gain, in this example, is even more impressive than you might have expected from part (c)(i) above. If it doesn't exhibit this, then redo D5e, because your answer to it is wrong!)