Notes on the stable manifold theorem
University of Houston, Math 6324, Fall 2012, Vaughn Climenhaga

The stable manifold theorem is one of the most important in the theory of non-linear ODEs and dynamical systems. Unfortunately, some of the standard introductory texts (Hirsch–Smale, Perko) either do not give a proof, or do not motivate the proof, while more advanced texts (Katok–Hasselblatt, Barreira–Pesin) are too high-powered to be appropriate in the setting of an introductory graduate course in ODEs, such as the one I am teaching now. So I'm taking this opportunity to turn my hastily scrawled hand-written notes on a (hopefully) properly-motivated proof into something that will be legible to myself and others on a more permanent basis.

The form of the theorem we will prove is this: let $U \subset \mathbb{R}^n$ be an open domain and $\varphi_t \colon U \to U$ the flow of a C^1 vector field $f \colon U \to \mathbb{R}^n$. Suppose that 0 is an equilibrium point for f and let $E^s \subset \mathbb{R}^n$ be the stable subspace for $Df(0)$ (the span of all generalised eigenvectors corresponding to eigenvalues with negative real part). Let $E^{cu} = E^c \oplus E^u \subset \mathbb{R}^n$ be the centre-unstable subspace for $Df(0)$ (corresponding to eigenvalues with zero or positive real part). Then there exists $r > 0$ and a C^1 function $\psi \colon B(0,r) \cap E^s \to E^{cu}$ such that the set $W^s := \operatorname{graph} \psi = \{x + \psi(x) \mid x \in B(0,r) \cap E^s\}$ has the following properties:

1. W^s is positively invariant;

2. given an initial condition $x \in W^s$, we have $\lim_{t\to\infty} \varphi_t(x) = 0$.

(Note that if f has an equilibrium point $\bar{x} \neq 0$, we can make the change of coordinates $y = x - \bar{x}$ and proceed as stated above. The choice $\bar{x} = 0$ simplifies the notation.)

PROOF: We present a proof using Perron's method – this is essentially the proof given in Perko's book, but with significantly more motivation and a more abstract viewpoint. The idea is to use the same general strategy that worked well for us in the proof of the Picard–Lindelöf theorem on local existence and uniqueness of solutions: prove existence and uniqueness of something by obtaining as the unique fixed point of a contraction on a complete metric space. In the Picard–Lindelöf theorem, the space was the space of approximate solutions, and the contraction was the Picard operator that transformed a candidate solution into something even closer to being a solution. Here we need to consider a slightly different space and operator.

1. We will consider the space of curves that approach the fixed point 0 with a certain exponential rate, as shown in Figure 1. In particular, for each $a \in E^s$ we will consider the set of such curves that begin at $a+y$ for some $y \in E^{cu}$. As before, we do not assume a priori that these curves are trajectories of the ODE.

2. We will define an integral operator on this space whose fixed points are solutions of the ODE. Then an application of the Banach fixed point theorem completes the proof.

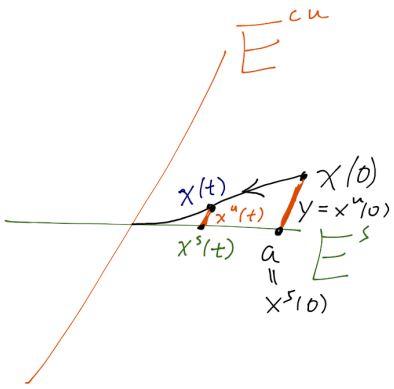

Figure 1: Approximate trajectories approaching 0 exponentially.

To this end, let $A = Df(0)$ be the linear part of f at the fixed point, and let $P = A|_{E^s}$ and $Q = A|_{E^{cu}}$ be the restrictions of A to the stable subspace and centre-unstable subspace, respectively. Let $-\alpha$ be the maximum value of $\operatorname{Re}\lambda$ for eigenvalues of P, and fix constants $0 < \gamma < \xi < \beta < \alpha$, as shown in Figure 2. Then all eigenvalues of P lie strictly to the left of $-\beta$ and all eigenvalues of Q lie strictly to the right of $-\gamma$. In particular, we can fix a norm on $\mathbb{R}^n$ with the property that

$$\begin{aligned} \|e^{Pt}\| &\leq e^{-\beta t}, \\ \|e^{-Qt}\| &\leq e^{\gamma t} \end{aligned} \tag{1}$$

for all $t \geq 0$. (Note that the second of these does not give decay to 0, but does give some control on growth.)

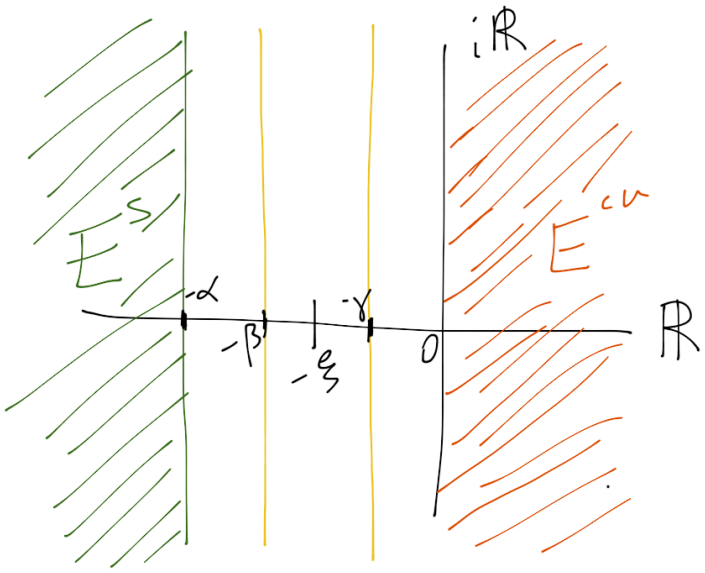

Figure 2: Bounds on eigenvalues corresponding to E^s and E^{cu}.

We work in the space of all C^1 curves approaching 0 with exponential rate at least ξ:

$$X := \left\{ x \colon [0,\infty) \to \mathbb{R}^n \,\middle|\, x \text{ is } C^1, \|x\|_\xi := \sup_{t\geq 0} |x(t)|e^{\xi t} < \infty \right\}. \tag{2}$$

This is a complete metric space with distance given by

$$d(x,y) = \|x-y\| = \sup_{t\geq 0} |x(t)-y(t)|e^{\xi t}.$$

(The reader familiar with Banach spaces will note that X is in fact a Banach space.)

Fixing $a \in E^s$, we consider the subspace $X_a = \{x \in X \mid x^s(0) = a\}$. Here and throughout we will write $x^s \in E^s$ and $x^{cu} \in E^{cu}$ for the parts of x lying in the stable subspace and centre-unstable subspace, respectively, so that $x = x^s + x_{cu}$. The goal is to find, for each sufficiently small $a \in E^s$, some $x \in X_a$ that is a trajectory of the ODE $\dot{x} = f(x)$. Then we can put $\psi(a) = x^{cu}(0)$ and conclude that ψ has the properties claimed in the theorem. (In fact we will not prove that ψ is C^1, which takes some more work, but positive invariance and exponential stability will follow from what we show.)

Having defined our metric space X_a, we need to define an operator on it whose fixed points are trajectories of the system. As in the Picard–Lindelöf theorem we would like to define an integral operator. We cannot simply use the one defined there ($(\mathcal{P}x)(t) = x(0) + \int_0^t f(x(s))\,ds$) for at least two reasons:

1. we need $\mathcal{P}$ to preserve the property $x(t) \to 0$, which this does not;
2. we will need to have the possibility that $(\mathcal{P}x)(0) \neq x(0)$, otherwise the operator will preserve the subsets of X_a on which $x^{cu}(0)$ is constant, and in particular each of these subsets could have a solution of the ODE, but the whole point is that X_a should only contain one value of $x^{cu}(0)$ corresponding to a solution.

Recall the idea from variation of constants: treat the ODE as a perturbation of a linear system, and write down an expression that is constant for the linear system. Then differentiating this expression gives the effect of the perturbation at each time, and integrating yields an expression for the solution of the full non-linear system. In other words, we write $f(x) = Ax + F(x)$, where $F(x) = o(|x|)$, so that the ODE becomes

$$\dot{x} = f(x) = Ax + F(x). \tag{3}$$

If $F(x) \equiv 0$ then the solution is $x(t) = e^{At}x(0)$, and in particular $e^{-At}x(t)$ is constant. So we differentiate this expression and use (3) to get

$$\begin{aligned}\frac{d}{dt}(e^{-At}x(t)) &= -Ae^{-At}x(t) + e^{-At}(Ax(t) + F(x(t))) \\ &= e^{-At}F(x(t)).\end{aligned}$$

Integrating, we see that (3) is equivalent to

$$x(t) = e^{At}x(0) + \int_0^t e^{A(t-s)}F(x(s))\,ds, \tag{4}$$

or more generally, by integrating from T to t for some $t \geq 0$,

$$x(t) = e^{A(t-T)}x(T) + \int_T^t e^{A(t-s)}F(x(s))\,ds. \tag{5}$$

Now by decomposing everything into the part lying in E^s and the part lying in E^{cu}, we see that $x(t)$ is a solution of (3) if and only if we have

$$\begin{aligned} x^s(t) &= e^{Pt}x^s(0) + \int_0^t e^{P(t-s)}F^s(x(s))\,ds, \\ x^{cu}(t) &= e^{Q(t-T)}x^{cu}(T) + \int_T^t e^{Q(t-s)}F^{cu}(x(s))\,ds \end{aligned} \tag{6}$$

for some (and hence every) $T \geq 0$. The reason for using non-zero values of T in the second equation is that we have good control on $e^{Q\tau}$ when $\tau < 0$ but not when $\tau > 0$, and so by sending $T \to \infty$ we may observe that

$$|e^{Q(t-T)}x^{cu}(T)| \leq \|e^{Q(t-T)}\||x(T)| \leq e^{-\gamma(t-T)}\|x\|e^{-\xi T} = e^{-\gamma t}\|x\|e^{(-\xi+\gamma)T} \to 0,$$

recalling the relationship between ξ, γ illustrated in Figure 2. In particular, we may replace (6) with

$$x^{cu}(t) = -\int_t^\infty e^{Q(t-s)}F^{cu}(x(s))\,ds. \tag{7}$$

The conclusion is that if we write

$$(\mathcal{P}x)(t) = e^{Pt}x^s(0) + \int_0^t e^{P(t-s)}F^s(x(s))\,ds - \int_t^\infty e^{Q(t-s)}F^{cu}(x(s))\,ds, \tag{8}$$

then $x = \mathcal{P}x$ if and only if x solves (3). One may also understand the motivation behind sending $T \to \infty$ as follows: for a fixed T, we expect the first term from (6) to grow exponentially as $t \to \infty$, which makes it difficult to verify the property $(\mathcal{P}x)(t) \to 0$. By eliminating this term (and the corresponding growth in the integral), we obtain an expression that is more tractable in the limit $t \to \infty$.

Now we show that $\mathcal{P}$ maps X_a to itself, and that it is a contraction, which will complete the proof by an application of the Banach fixed point theorem. First we observe that because the non-linear part F of the vector field f is C^1 with $DF(0) = 0$, it is Lipschitz on small neighbourhoods of 0, and the Lipschitz constant can be made arbitrarily small by making the neighbourhood small enough. More precisely, for every $\varepsilon > 0$ there exists $r > 0$ such that if $|x|, |y| \leq r$ then $|F(x) - F(y)| \leq \varepsilon|x - y|$.

Now we make a computation that shows both claims in the previous paragraph. Fix $x, y \in X$ and let r, ε be as above. Recall that by the definition

of $d(x,y)$ we have $|x(t)-y(t)| \le d(x,y)e^{-\xi t}$ for all $t \ge 0$, and write $\Delta^s = |x^s(0)-y^s(0)|$. Then

$$\begin{aligned}
&|(\mathcal{P}x)(t) - (\mathcal{P}y)(t)| \\
&\quad \le \|e^{Pt}\||x^s(0)-y^s(0)| + \int_0^t \|e^{P(t-s)}\||F^s(x(s)) - F^s(y(s))|\,ds \\
&\qquad\qquad + \int_t^\infty \|e^{Q(t-s)}\||F^{cu}(x(s)) - F^{cu}(y(s))|\,ds \\
&\quad \le e^{-\beta t}\Delta^s + \int_0^t e^{-\beta(t-s)}\varepsilon|x(s)-y(s)|\,ds + \int_t^\infty e^{-\gamma(t-s)}\varepsilon|x(s)-y(s)|\,ds \\
&\quad \le e^{-\beta t}\Delta^s + \varepsilon d(x,y)\left(\int_0^t e^{-\beta(t-s)}e^{-\xi s}\,ds + \int_t^\infty e^{-\gamma(t-s)}e^{-\xi s}\,ds\right) \\
&\quad \le e^{-\beta t}\Delta^s + \varepsilon d(x,y)\left(e^{-\beta t}\left[\frac{e^{(\beta-\xi)s}}{\beta-\xi}\right]_{s=0}^{t} + e^{-\gamma t}\left[\frac{e^{(\gamma-\xi)s}}{\gamma-\xi}\right]_{s=t}^{\infty}\right).
\end{aligned}$$

Using the fact that $\gamma - \xi < 0$ and $\beta - \xi > 0$, we obtain

$$\begin{aligned}
|(\mathcal{P}x)(t) - (\mathcal{P}y)(t)| &\le e^{-\beta t}\Delta^s + \varepsilon d(x,y)\left(\frac{e^{-\xi t} - e^{-\beta t}}{\beta-\xi} - \frac{e^{-\xi t}}{\gamma-\xi}\right) \\
&\le e^{-\xi t}\Delta^s + \varepsilon d(x,y)e^{-\xi t}\left(\frac{1}{\beta-\xi} + \frac{1}{\xi-\gamma}\right) \\
&= \left(\Delta^s + \frac{\varepsilon(\beta-\gamma)}{(\beta-\xi)(\xi-\gamma)}d(x,y)\right)e^{-\xi t}.
\end{aligned}$$

Writing $L = \frac{\beta-\gamma}{(\beta-\xi)(\xi-\gamma)}$ and recalling the definition of $\|\cdot\|_\xi$ in (2), this gives

$$\|\mathcal{P}x - \mathcal{P}y\|_\xi \le \Delta^s(x,y) + \varepsilon L d(x,y) \tag{9}$$

Given $x \in X$, by putting $y = 0$ we see that

$$\|\mathcal{P}x\|_\xi \le |x^s(0)| + \varepsilon L\|x\|_\xi < \infty,$$

and so $\mathcal{P}$ maps X to itself. Moreover, if $x \in X_a$ then it is apparent from (8) that $\mathcal{P}x \in X_a$ as well. Finally, if $x, y \in X_a$ for some $a \in E^s$, then $\Delta^s(x,y) = 0$ and (9) gives

$$d(\mathcal{P}x, \mathcal{P}y) \le \varepsilon L d(x,y).$$

By choosing r small enough we can guarantee that $\varepsilon L < 1$, and hence $\mathcal{P}$ is a contraction. Thus it has a unique fixed point $\bar{x} \in X_a$. This is a trajectory of the ODE (3) which approaches 0 exponentially (with rate at least ξ) and has $\bar{x}^s(0) = a$, so we put $\psi(a) = \bar{x}^{cu}(0)$ and conclude that ψ has the properties claimed in the statement of the theorem. $\square$

A few remarks are in order. The proof here follows the proof in Perko's book (p. 107–111), but includes more abstract language and more complete motivation. We have shown that ψ exists and its graph W^s is positively invariant with trajectories converging exponentially to 0. One can also show that ψ is C^1 and $D\psi(0) = 0$, so that W^s is tangent to E^s. This proof is not in Perko's book – he refers to the text of Coddington and Levinson (p. 332–333). If I feel ambitious I may try to include this part of the proof at a later point.

The proof given above also shows that $\bar{x} \in X_a$ is the *only* trajectory with $x^s(0) = a$ that converges exponentially to 0 with rate at least ξ. In particular, any such trajectory lies in the stable manifold W^s. However, it is possible that there are trajectories lying outside of W^s that converge to 0 more slowly – for example, there may be trajectories converging to 0 with subexponential speed along the centre direction E^c.

Ultimately, the only property of the subspace E^{cu} that we used in the proof was the fact that all eigenvalues of $Q = Df(0)|_{E^{cu}}$ have real part strictly greater than $-\gamma$. In fact, the entire proof above is valid if we replace E^{cu} with a subspace $E^2 \subset \mathbb{R}^n$ such that

1. $\mathbb{R}^n = E^s \oplus E^s$; and

2. there exists $-\xi < 0$ such that every eigenvalue of $P = Df(0)|_{E^s}$ has real part strictly less than $-\xi$, and every eigenvalue of $Q = Df(0)|_{E^s}$ has real part strictly greater than $-\xi$.

In this setting, we still get a stable manifold, which is often called a *strong stable manifold* and denoted W^{ss}, because it corresponds to the directions in which the contraction is the strongest. There may be trajectories lying outside of W^{ss} that approach 0 exponentially, but the rate at which they do so is less than ξ.