Journal of Convex Analysis
Volume 9 (2002), No. 1, 185–224

Lagrangian Manifolds, Viscosity Solutions and Maslov Index*

D. McCaffrey
University of Sheffield, Dept. of Automatic Control and Systems Engineering, Mappin Street, Sheffield, S1 3JD, UK
david.mccaffrey@opc.shell.com

S. P. Banks
University of Sheffield, Dept. of Automatic Control and Systems Engineering, Mappin Street, Sheffield, S1 3JD, UK
s.banks@sheffield.ac.uk

Received December 13, 2000

Let M be a Lagrangian manifold, let the 1-form pdx be globally exact on M and let $S(x,p)$ be defined by $dS = pdx$ on M. Let $H(x,p)$ be convex in p for all x and vanish on M. Let $V(x) = \inf\{S(x,p) : p$ such that $(x,p) \in M\}$. Recent work in the literature has shown that (i) V is a viscosity solution of $H(x, \partial V/\partial x) = 0$ provided V is locally Lipschitz, and (ii) V is locally Lipschitz outside the set of caustic points for M. It is well known that this construction gives a viscosity solution for finite time variational problems - the Lipschitz continuity of V follows from that of the initial condition for the variational problem. However, this construction also applies to infinite time variational problems and stationary Hamilton-Jacobi-Bellman equations where the regularity of V is not obvious. We show that for $\dim M \leq 5$, the local Lipschitz property follows from some geometrical assumptions on M - in particular that the Maslov index vanishes on closed curves on M. We obtain a local Lipschitz constant for V which is some uniform power of a local bound on M, the power being determined by $\dim M$. This analysis uses Arnold's classification of Lagrangian singularities.

1. Introduction

This paper considers a geometrical approach to constructing stationary viscosity solutions to Cauchy problems involving Hamilton-Jacobi-Bellman (HJB) equations. This approach has recently been put forward by M.V. Day in [11]. The geometry involved is that of the Lagrangian manifolds in phase space on which the characteristic curves of the Cauchy problem lie. A certain regularity assumption, namely local Lipschitz continuity, had to be made by Day in order to show that the function he constructs is a viscosity solution. The main point of this paper is to investigate how this assumption follows from the topological and geometrical properties of the relevant Lagrangian manifold in the stationary case. There are several strands in the literature which have to be introduced in order to place this problem in its proper context. We do this now.

The first of these strands is optimal control and, more generally, Bolza type variational problems. Sufficient conditions for a solution to such a problem usually reduce to solving an HJB partial differential equation, see for instance Chapter 4 of [14] or Chapter 1 of [15]. In particular, optimising over a finite time horizon leads to a Cauchy problem where the initial condition (or alternatively the final condition) is determined by the initial

*This work was partially supported by EPSRC grant number GR/J 75241.

ISSN 0944-6532 / \$ 2.50

cost term in the cost functional. The characteristic curves for the Cauchy problem are the trajectories of the Hamiltonian system corresponding to the HJB equation. In the case of optimal control, this Hamiltonian system is given by some form of Pontryagin's maximum principle. The well known difficulty in formulating such problems is that, even with smooth initial data, the solution to the HJB equation is generally non-smooth. The characteristic curves start to cross at a finite distance from the initial manifold and the solution then becomes multi-valued.

This leads to the second strand, namely viscosity solutions, for which the best introductory reference is probably still [10]. One of the main motivations for their introduction was to provide an acceptable definition of the sense in which the value function for an optimal control problem can be said to solve Bellman's equation - see for instance Theorem 5.1 of [15], Theorem 2.3 of [17] or Theorem 1.10 of [22].

As already stated, the aspect of a viscosity solution which concerns us here is its connection with the geometry of the underlying Lagrangian manifold in phase space. This connection has already been explored in several places, so in order to make clear the interesting features of Day's construction, it is worth considering briefly how it differs from these other contributions.

Firstly, Hopf's original formula (Theorem 5a of [24]) for a generalised solution to an HJB equation in the case of a convex Hamiltonian is essentially geometric - the value of the solution at a point x is related to the value of the initial condition at a point y, the relationship being that y is the initial point on the minimising trajectory through x for the variational problem giving rise to the HJB equation. This formula was shown by Evans in Theorem 6.1 of [13] to be a viscosity solution. In this finite time variational setting, Lipschitz continuity of the generalised solution follows immediately from the corresponding Lipschitz continuity of the initial condition. In Examples 4.1.1 and 4.1.2 of [11], Day shows that his construction gives the same function as Hopf's formula. The Lipschitz property he requires in order to deduce that his construction is a viscosity solution follows from the above variational interpretation. However as the function constructed is already known to be a viscosity solution, this example doesn't expose the real application of his construction, which is to infinite time variational problems.

The other recent exploration of the geometry of viscosity solutions has been in papers such as [8] which have considered how to express classical second order conjugate point type necessary and sufficient conditions in the viscosity setting. These results essentially identify the points at which nearby trajectories start to cross and are no longer locally optimal. Again the analysis studies the evolution of an initial non-smooth manifold along the trajectories of a Hamiltonian system.

The interesting aspect of Day's construction is that it holds in the absence of any variational interpretation. In some sense it is independent of the evolution of initial manifolds along characteristics and it furnishes information on global optimality as opposed to local optimality. He makes this point himself in his paper, but we hope here to give an indication of the sense in which this is true.

This brings us to the third strand of the introduction. The particular case where Day's construction is of interest is in looking at stationary solutions to HJB Cauchy problems. In this case one cannot deduce the Lipschitz continuity of the solution from the evolution of a Lipschitz continuous initial condition, as the initial condition is the solution. No more is known about one than the other. Instead stationary solutions are usually arrived at as some sort of limit of solutions to finite time variational problems and regularity is difficult to prove. The key point of this paper is that the Lipschitz property actually follows from the topology of the relevant Lagrangian manifold in the stationary case.

To introduce this, we briefly review various approaches to finding stationary solutions. In linear quadratic optimal control, given appropriate stabilizability and detectability conditions, stationary solutions are found as the limit of the value functions for finite time optimal control problems as the final time tends to infinity - see for instance [29]. The corresponding Hamiltonian dynamics in phase space (i.e. the dynamics arising from the maximum principle) are hyperbolic and the limiting value function for the infinite time problem is the generating function for the stable Lagrangian plane at the equilibrium point. This idea is extended to non-linear infinite time optimal control in [4, 23] and to non-linear H_∞ control in [32, 33]. The key point is that the stable manifold theorem implies the existence of a stable Lagrangian manifold whose tangent plane at the origin is the stable plane for the phase space dynamics of the linearised control problem. In the region around the origin where the stable manifold has a well-defined projection onto state space, the value function for the infinite time problem is smooth and is the generating function for the stable manifold.

The extension to viscosity solutions of infinite time optimal control problems and H_∞ problems is done in, for instance, [15, 22] and [30] respectively. However the approach is not explicitly geometric. The stable manifold approach, at least as far as viscosity solutions are concerned, has until now been stuck at the point at which the projection onto state space becomes ill-defined. The connection is probably contained implicitly in the above cited works in that they all consider an exponentially discounted infinite cost function while the stable manifold theorem gives exponential bounds on the approach to the equilibrium point. So a transformation between the two viewpoints may be possible.

However, Day's construction does make explicit the link between the stable manifold and the stationary viscosity solution for infinite time optimal control - the full H_∞ problem is currently outside the scope of his construction as convexity of the Hamiltonian is required. As above, the relevant Lagrangian manifold is the stable manifold for the Hamiltonian system. This is simply connected - at least that portion of it which can be connected by Hamiltonian trajectories to a neighbourhood of the equilibrium point is. Closed curves on this manifold therefore have zero Maslov index. The purpose of this paper is to explain these terms, particularly Maslov index, and then show that, for manifolds of dimension ≤ 5, this implies local Lipschitz continuity of Day's construction and hence a stationary viscosity solution. The result for higher dimensions follows from an open conjecture on the form of so-called non-folded singularities on the Lagrangian manifold.

To end this introduction, we mention one of the other major approaches to defining unique generalised solutions to HJB equations, namely idempotent analysis. This is based on the observation that HJB equations become linear when considered with respect to the arithmetic operations of (max,+) rather than (+,$\times$), see [27, 20, 21, 19] for instance. This observation arose out of the study of logarithmic limits of short wave length asymptotic solutions to quantum tunnelling equations. It lead to the construction of quantum tunnelling canonical operators, usually based on heat transforms. We indicate in this paper the connection between Day's construction and the logarithmic limit of a quantum tunnelling canonical operator based on the Laplace transform. In particular we define an orientation on M which allows the correct choice of 'side' of the Laplace transform at any point.

The contents of the paper are as follows. The next section introduces the required background on Lagrangian manifolds, HJB equations and Day's construction. Section 3 then introduces the Maslov index. Section 4 explores the connection with idempotent solutions. Section 5 presents the main results of the paper - namely that vanishing Maslov index implies local Lipschitz continuity of Day's construction and, further, that the local Lipschitz constant can be expressed as some uniform power of a local bound on M where the power depends on $\dim M$. It follows that Day's construction is a viscosity solution. Lastly Section 6 covers some examples and areas for further work.

2. Lagrangian manifolds and HJB equations

This section reviews Day's proposed construction of viscosity solutions to HJB equations. It starts with a very brief survey of the required elements of symplectic geometry. There are many references in the literature for this material. Day himself gives a nice introduction in Sections 1 and 2 of [11]. Another control theoretic perspective is given in [32, 33]. More detailed expositions from the perspectives of classical mechanics or geometry can be found in, for instance, [3, 25, 28, 31].

Define phase space to be the real $2n$-dimensional vector space with coordinates (x, p) where $x \in \mathbb{R}^n$ and $p \in \mathbb{R}^n$. (Much of the following applies to general symplectic spaces but our ultimate applications all live in $\mathbb{R}^{2n}$.) On phase space there exists a canonical two-form $\omega = dp \wedge dx$. Let $\phi : M \to \mathbb{R}^{2n}$ be an n-dimensional submanifold of phase space on which the restriction of the canonical two-form vanishes, i.e.

$$\phi^*(dp \wedge dx) \equiv 0.$$

Then M is said to be a Lagrangian submanifold of phase space. This means that the one form pdx is locally exact on M.

Let I denote a subset of the set $\{1, \ldots, n\}$ and $\bar{I}$ denote its compliment. Let x^I denote the set of coordinates $\{x^i : i \in I\}$ and $p_{\bar{I}}$ denote the set $\{p_k : k \in \bar{I}\}$. Then it follows from the Lagrangian property that, at any point on M, there exists a collection of indices I such that $(x^I, p_{\bar{I}})$ form a local system of coordinates on M - see Section 2.1 of [28] or Proposition 4.6 of [25]. This means that M can be covered by an atlas of so-called canonical coordinate charts U_I in each of which its immersion into phase space is given by the relations

$$x^{\bar{I}} = x^{\bar{I}}\left(x^I, p_{\bar{I}}\right) \qquad p_I = p_I\left(x^I, p_{\bar{I}}\right). \tag{1}$$

Furthermore, it follows from the Lagrangian property that in each coordinate neighbourhood U_I, there exists a function $S_I(x^I, p_{\bar{I}})$ satisfying the equation

$$dS_I = \phi^*\left(p_I dx^I - x^{\bar{I}} dp_{\bar{I}}\right) \tag{2}$$

- see again Section 2.1 of [28] or Theorem 4.21 of [25]. (For brevity we cease, from now on, to distinguish between forms defined on M and forms defined on all of phase space.) The above equation means that M can be locally represented in the form

$$x^{\bar{I}} = -\frac{\partial S_I}{\partial p_{\bar{I}}} \qquad p_I = \frac{\partial S_I}{\partial x^I}. \tag{3}$$

S_I is called a generating function for M .

Now consider a C^2 real valued Hamiltonian function $H(x,p)$ defined on phase space. Associated with H is a Hamiltonian vector field X_H defined by the relation

$$i(X_H)\omega = dH$$

or in coordinate terms

$$\dot{x} = \frac{\partial H}{\partial p} \qquad \dot{p} = -\frac{\partial H}{\partial x}.$$

It follows again from the Lagrangian property that if H is constant on M then the vector field X_H is tangent to M, i.e. M is invariant with respect to the Hamiltonian flow corresponding to H.

This is the principle reason why Lagrangian manifolds are so important. They are formed by collections of Hamiltonian trajectories or characteristic curves for the HJB equation

$$H(x, \partial S/\partial x) = 0 \tag{4}$$

where $\partial S/\partial x$ denotes the vector of partial derivatives of S with respect to x. Suppose $H(x,p) = 0$ on M. Then, on any of the specific coordinate neighbourhoods U_I where $I = \{1,\ldots,n\}$, it follows from the relation (3) above that the local solution to (4) is given by the relevant S_I, up to an additive constant. Note that a HJB equation of the form

$$H(t, x, \partial S/\partial x) = -\partial S/\partial t \tag{5}$$

can be transformed into one of the form (4) by taking t and $-H$ to be canonical coordinates. This is equivalent to considering the Lagrangian manifold traced out in $\mathbb{R}^{2n+2}$ phase space by the evolution under the Hamiltonian flow of the Lagrangian manifold given by the initial condition.

The solution given by the above method of characteristics is only local. Day's construction attempts to extend it beyond the coordinate neighbourhood in which it is defined. He makes the following basic hypotheses in order to obtain his construction.

Hypotheses 2.1. (Day)

1. M is a Lagrangian submanifold of $\mathbb{R}^{2n}$.
2. H is a C^2 real valued function on $\mathbb{R}^{2n}$ with $H(x,p) = 0$ for all $(x,p) \in M$.
3. pdx is globally exact on M.
4. M is locally bounded.
5. M covers an open region Ω of state space $\mathbb{R}^n$ and has no boundary points over Ω.

The technical reasons for making these assumptions can be found in [11]. We only repeat the following comments for later use. Firstly, note that by the term submanifold, we mean that M is embedded rather than just immersed. Also, note that by boundary points we mean points in $\bar{M}\backslash M$, where $\bar{M}$ denotes the closure of M. Then, from Hypothesis (5) and the fact that M is embedded in $\mathbb{R}^{2n}$, it follows that if $x_n \to x \in \Omega$ with $(x_n, p_n) \in M$ for each n and $p_n \to p \in \mathbb{R}^n$, then $(x,p) \in M$ and $(x_n,p_n) \to (x,p)$ in the topology of M.

Secondly, Hypothesis (4) means that for each $x_0 \in \Omega$ there exists a $\delta > 0$ and $K < \infty$ such that $|p| \leq K$ for all $(x,p) \in M$ with $x \in B_\delta(x_0)$.

Lastly, the fact that M is Lagrangian means that the one-form pdx is closed when restricted to M. It thus defines a cohomomology class $[\phi^*(pdx)] \in H^1(M)$. Hypothesis (3) says that this class is trivial, or equivalently that there exists a function $S(x,p)$ defined globally on M which satisfies $dS = pdx$. For our analysis later on, it is useful to express this in terms of the local generating functions on canonical coordinate neighbourhoods U_I defined above. Let Φ_I denote the restriction of S to U_I. Then the generating function $S_I(x^I, p_{\bar{I}})$ is defined as

$$S_I = \Phi_I - x^{\bar{I}} p_{\bar{I}}.$$

It follows that this function S_I satisfies (2).

Conversely, we can express Hypothesis 2.1(3) in terms of the local generating functions S_I. Let U_I and U_J be any two canonical coordinate charts with a non-trivial intersection. Let $I_1 = I \cap J$, $I_2 = I \backslash J$, $I_3 = J \backslash I$ and $I_4 = \{1, \ldots, n\} \backslash (I \cup J)$. Then Hypothesis 2.1(3) is equivalent to the equation

$$S_I - S_J = p_{I_2} x^{I_2} - p_{I_3} x^{I_3} \tag{6}$$

holding in $U_I \cap U_J$. For then if the functions

$$\Phi_I = S_I + x^{\bar{I}} p_{\bar{I}} \tag{7}$$

are defined in each of the respective charts U_I, it follows from (6) that they agree on pairwise intersections. Thus they glue togther to give a smooth function S defined on the whole of M which coincides with Φ_I on each U_I and which satisfies on each chart

$$dS = d\Phi_I = pdx. \tag{8}$$

Equation (6) is the reformulation in Cech cohomology of the requirement that the class $[pdx]$ be trivial. It is known in quantum mechanics as Maslov's first quantization condition. We will return to this point below.

Given the existence of a smooth function $S(x,p)$ defined globally on M and satisfying $dS = pdx$, Day then proposes the following function

$$W(x) = \inf\{S(x,p) : p \text{ such that } (x,p) \in M\} \tag{9}$$

as a viscosity solution. Note that by Hypothesis 2.1(4), the infimum in (9) is achieved for every $x \in \Omega$ - see [11], Section 2.1.2. In Theorem 3 of [11] he proves the following.

Theorem 2.2. *(Day) If $H(x,p)$ is convex in p for each x and if $W(x)$ is (locally) Lipschitz continuous in Ω, then W is a viscosity solution of $H(x, \partial W(x)/\partial x) = 0$ in Ω.*

He also shows in Theorem 1 of the same paper that without the Lipschitz condition, W is a lower semi-continuous viscosity supersolution. Note, as already pointed out in the Introduction, in a finite time variational setting, the Lipschitz continuity of (9) would follow easily from that of the corresponding initial condition. However, our interest in the viscosity solution defined by (9) is that its definition is independent of any variational interpretation. In particular, it is applicable to the study of stationary solutions corresponding to stable or unstable Lagrangian manifolds of hyperbolic equilibrium points.

3. Maslov index

This paper will show that the Lipschitz continuity required in Theorem 2.2 follows if, amongst other conditions, all closed curves on the Lagrangian manifold M have Maslov index zero. We now define what is meant by the Maslov index.

The simplest definition involves the points on the Lagrangian manifold M at which the projection of M onto state space $\mathbb{R}^n$ is singular. Note, for later reference, that the images in state space of such singular points are called *caustic points*. Let Σ denote the set of singular points on M. It is shown in [1] that, in the generic case, Σ is an $(n-1)$-dimensional two-sided cycle in M - see also, for instance, Appendix 11 of [3] or Theorem 7.6 of [25]. The generic case can always be achieved by an arbitrarily small deformation in the class of Lagrangian manifolds. This means that a positive and a negative side of Σ can be consistently defined. The definition of the orientation goes as follows. Recall from above the notion (1) of canonical coodinate charts on M. In the neighbourhood of a simple singular point on Σ (i.e. one at which the rank of the projection onto state space drops by 1), a canonical system U_I can be chosen, where $I = \{1, \ldots, \hat{\imath}, \ldots, n\}$ for some $i \in \{1, \ldots n\}$, where $\hat{}$ denotes omission. This means that M is represented in a neighbourhood of the singular point in the form

$$x^i = x^i(x^I, p_i) \qquad p_I = p_I(x^I, p_i).$$

Singular points near the given one are then defined by the condition $\partial x^i / \partial p_i = 0$. For M in general position (i.e. up to a small deformation), this derivative changes sign on passing from one side of Σ to the other in the neighbourhood of the given simple singular point. The positive side of Σ is then taken to be the side where this derivative is positive. It is shown in [1] that this definition is independent of the chosen coordinate system.

Given this orientation of Σ, the Maslov index of a curve γ on M is then defined to be

$$\mathrm{ind}(\gamma) = \nu_+ - \nu_- \tag{10}$$

where ν_+ is the number of points where γ crosses from the negative to the positive side of Σ and ν_- is the number of points where γ crosses from the positive to the negative side of Σ - see for instance Appendix 11 of [3] or Definition 7.7 of [25]. This definition assumes that the endpoints of γ are non-singular and that γ only intersects Σ transversely in simple singular points. It is then extended to any curve on M by approximating such a curve with one of the form γ - it can be shown that the definition is independent of the approximating curve. As an example, the Maslov index of the circle $x^2 + p^2$ traversed anti-clockwise in 2-dimensional phase space is -2.

For those readers familiar with the Morse index from calculus of variations, it may help to note that the Maslov index is related to the Morse index as follows. An extremal for a variational problem with Hamiltonian H corresponds to a phase curve in $\mathbb{R}^{2n}$ phase space. As described above for equation (5), this can be lifted to a phase curve lying on an $(n+1)$-dimensional Lagrangian manifold in $\mathbb{R}^{2n+2}$ phase space by considering its evolution in time under the Hamiltonian flow. The Morse index of the extremal is equal to the Maslov index of the corresponding phase curve on the $(n+1)$-dimensional Lagrangian manifold - see [1], Theorem 5.2 or Appendix 11 of [3].

The definition in (10) is sufficient to allow us to use the concept of Maslov index in the proofs of this paper. However, it is worth noting that, in [1], Arnold actually proves that there is a 1-dimensional cohomology class g^* on M such that, for a closed curve γ on M, the evaluation of g^* on γ is equal to $\text{ind}(\gamma)$ defined above. Again, a summary can be found in Appendix 11 of [3]. This class g^* is called the Maslov class of M. It is constructed as the pull-back of a generator of cohomology on the universal bundle on a Grassmanian manifold. It is thus a characteristic class on M. This means that if M_1 and M_2 are Lagrangian manifolds, $F : M_1 \to M_2$ is a smooth map and g_1, g_2 denote the Maslov classes on M_1 and M_2 respectively, then $g_1 = F^*(g_2)$. Much of the work in Section 5 is aimed at proving that the Maslov index is also preserved under pull-back for certain types of non-closed curves.

The Maslov class can also be defined in terms of Cech cohomology - see Definition 7.4 of [25] and Section 2.3 of [28].

4. Idempotent Solutions

The principal application of the Maslov index has been to the construction of global asymptotic solutions to linear PDEs arising in mathematical physics. This has lead to the development of the idempotent analytic approach to solving HJB equations. A number of authors, e.g. [18], have considered the connections between viscosity and idempotent solutions to HJB equations. In this section we give a brief introduction to idempotent analysis with the aim of showing that Day's construction has already been used to obtain idempotent solutions to certain forms of stationary HJB equations. This holds out the possibility of showing a connection between viscosity and idempotent solutions in the stationary case.

The vanishing (mod 2) of the Maslov class is called the second Maslov quantization condition. The first quantization condition has already been introduced above in equation (6) - namely the vanishing of the class $[pdx]$. These two conditions appear as obstructions to the existence of global asymptotic solutions to linear pseudodifferential equations of the form

$$ih\frac{\partial \psi}{\partial t} = H\left(x, -ih\frac{\partial}{\partial x}\right)\psi, \quad \psi(x,0) = \exp\left(\frac{i}{h}S_0(x)\right)\phi_0(x) \tag{11}$$

as $h \to 0$. The construction of a local asymptotic solution is obtained by the famous Wentzel, Kramers and Brillouin (WKB) method. This involves looking for a solution of the form

$$\psi(x,t) = \exp\left(\frac{i}{h}S(x,t)\right)\phi(x,t).$$

The asymptotic phase S is the solution of the characteristic HJB equation

$$\frac{\partial S}{\partial t} + H\left(x, \frac{\partial S}{\partial x}\right) = 0, \quad S(x,0) = S_0(x)$$

corresponding to (11). The function $\phi(x,t)$ is the solution of a so-called transport equation. This representation only holds up to the first focal point of the Hamiltonian flow corresponding to H with respect to the initial manifold $M_0 = \{x, \partial S_0/\partial x\}$ in phase space.

It is extended to a global representation as follows. Consider another canonical coordinate chart U_I on the Lagrangian manifold M formed by the evolution of M_0 along the Hamiltonian flow given by H. Let S_I denote the generating function of M in U_I. Let

$\hat{p} = -ih\partial/\partial x$. Then there is a $1/h$-Fourier transform in x and p on phase space that takes the pseudolinear equation (11) to one for which

$$\psi_I(x^I, p_{\bar{I}}, t) = \exp\left(\frac{i}{h} S_I(x^I, p_{\bar{I}}, t)\right) \phi_I(x^I, p_{\bar{I}}, t)$$

is the local WKB solution. This is transformed back to a solution in coordinates $\{x^1, \ldots, x^n, t\}$ by an inverse $1/h$-Fourier transform. The construction of a global asymptotic solution on $\mathbb{R}^n \times \mathbb{R}$ then involves showing that the various representations are asymptotically equal on intersections $U_I \cap U_J$. The integrals involved in the inverse transforms are asymptotically expanded using the method of stationary phase, which involves the square root of the Jacobian of the change of coordinates between U_I and U_J. The first quantization condition guarantees that the phases S_I and S_J glue together into a global asymptotic phase. The second quantization condition guarantees that the branches of the complex square root function can be chosen such that the arguments of the square root cancel out globally on M. Details can be found, for instance, in Section 4.1 of [28]. The whole proceedure can be formulated without reference to Cauchy problems in terms of the so-called Maslov canonical operator. This maps functions ψ defined on M_0 globally to $\mathbb{R}^n$ and gives an asymptotic Green's function for equation (11).

The above method is typically applied to the Schrodinger equation

$$ih\frac{\partial \psi}{\partial t} = \left(-\frac{h^2}{2}\Delta + V(x)\right)\psi.$$

The connection with optimisation is via the search for the low level asymptotic eigenfunctions of the Schrodinger operator. This leads to the study of the large time asymptotics of the equation

$$h\frac{\partial u}{\partial t} = \left(\frac{h^2}{2}\Delta - V(x)\right)u, \tag{12}$$

i.e. to the study of asymptotic quantum tunnelling solutions - see [12], Section 1 or [26]. The logarithmic asymptotics of this equation also turn up in the study of large deviation problems in probability. The canonical operator giving the globally asymptotic (as $h \to 0$) solution of (12) is called the tunnelling canonical operator. It gives asymptotic solutions in which the principal term is of the form $\exp(-S(x,t)/h)$ where the entropy S is the generalised solution to the characteristic HJB equation

$$\frac{\partial S}{\partial t} + \frac{1}{2}\left(\frac{\partial S}{\partial x}\right)^2 - V(x) = 0, \quad S(x,0) = S_0(x). \tag{13}$$

This S is therefore also the generalised solution to an optimisation problem.

The tunnelling canonical operator differs from the Maslov canonical operator in that, in the neighbourhood of a focal point, the transformation between canonical charts on the Lagrangian manifold M corresponding to (13) is achieved by translating along the Hamiltonian flow given by $H_I = (1/2)\Sigma_{i \in I} p_i^2$ where $I \subset \{1, \ldots, n\}$. The corresponding transformation of the solution to (12) is therefore obtained by applying the solving operator for the heat equation. The resulting asymptotic integral expansion uses the Laplace method. This is simpler than the method of stationary phase. At any given point x, it identifies the branch of M on which the generating function S takes its minimum value over all the branches projecting onto x, rather than summing the contribution from all the branches as in the stationary phase method. The first Maslov quantization condition is still required to ensure the existence of a globally defined entropy S on M. The second quantization condition is not needed, however, as the Jacobian entering into Laplace's method is assumed to be positive and so there are no problems in taking its square root.

The logarithmic limit as $h \to 0$ of the asymptotic solution to (12) gives the so-called idempotent generalised solution to (13). This takes the form of a resolving operator R_t such that $S(x,t) = R_t S_0(x)$ is the generalised solution to (13) where R_t is defined on the set of functions bounded from below by the formula

$$R_t S_0(x) = \inf_{\xi}(S_0(\xi) + S(t,x,\xi)). \tag{14}$$

Here $S(t,x,\xi)$ is the value of the cost functional, for the variational problem corresponding to (13), along the minimising extremal starting at ξ at time $-t$ and ending at x at time 0. The operator R_t is linear with respect to the arithmetic operations $(\min, +)$, i.e.

$$\begin{aligned} R_t(\min(S_1, S_2)) &= \min(R_t S_1, R_t S_2), \\ R_t(\lambda + S(x)) &= \lambda + R_t S(x), \qquad \lambda \in \mathbb{R}. \end{aligned}$$

This is what is meant by an idempotent solution to (13). The fact that it is obtained by a linear resolving operator over an appropriate space allows the usual apparatus of analysis, i.e. weak solutions, distributions, Green's functions, convolutions, etc. and the corresponding numerical approximation schemes, to also be carried over to this space in order to find generalised solutions to HJB equations. A good review of this can be found in the introductory paper by Gunawardena in the volume [19]. A simple example is given in [27] of the 1-dimensional heat equation with a small parameter h which shows how the 'min' superposition law for the entropies arises from the log limit of the normal linear superposition law for solutions of the heat equation as $h \to 0$.

The connection between idempotent solutions and Day's construction (9) can now be stated. See Section 5 of [12] for more details. Suppose that in equation (13), the potential V is a smooth nonnegative function with a bounded matrix of second derivatives. Suppose also that $V(x) \to \infty$ as $|x| \to \infty$ and that $V(x)$ is zero at just a finite number of points $\xi_1, \dots \xi_l$ at each of which the matrix of second derivatives of V is non-singular. As already mentioned above, for the variational problem

$$J(c(.)) = \int_{-t}^{0} \left(\frac{1}{2} \dot{c}^2(\tau) + V(c(\tau)) \right) d\tau$$

corresponding to (13), let $S(t,x,\xi)$ denote the value of the cost functional along the minimising extremal starting at ξ at time $-t$ and ending at x at time 0. The cost functional takes its minimum value on a smooth curve. Now define

$$S_k(x) = \inf\{S(t,x,\xi_k) : t > 0\} = \lim_{t \to \infty} S(t,x,\xi_k)$$

and

$$S(x) = \min\{S_k(x) : k = 1, \dots, l\}.$$

It is pointed out in Proposition 1 of [12] that $S_k(x)$ and $S(x)$ are stationary idempotent solutions of (13). Furthermore, it is shown in Proposition 3 that $S_k(x)$ is the generating function for the essential parts of the unstable Lagrangian manifold through $(\xi_k, 0)$. Note for later reference that a *non-essential* point r on a Lagrangian manifold is one for which there is another point on the Lagrangian manifold with the same projection onto state space and at which the generating function has a lower value than at r. Thus it can be seen that the construction (9) considered by Day is already known to give stationary idempotent solutions and to relate them to the corresponding unstable manifold for the Hamiltonian flow when V satisfies the above assumptions.

In the same way as mentioned in the Introduction for viscosity solutions, the Lipschitz continuity of finite time idempotent solutions to the Cauchy problem (13) follows from the resolving operator representation (14) and the Lipschitz continuity of the initial condition S_0. However, this argument again fails to work for stationary idempotent solutions to (13). If it can be shown that Day's construction (9) is locally Lipschitz continuous then it will follow that, for the particular form of potential V considered here, stationary idempotent solutions to (13) coincide with stationary viscosity solutions.

We end this discussion of canonical tunnelling operators and idempotent solutions with the following remark. Recall from above that in the definition of the standard canonical tunnelling operator, the transformation between different canonical charts on the Lagrangian manifold defining the entropy is achieved with heat transforms. Day's construction (9) can be considered as the logarithmic limit of a canonical tunnelling operator. However in this case the transition between local representations for the entropy is achieved via the relation (6). This can be viewed as the logarithmic limit of a $1/h$-Laplace transform. In other words, the corresponding canonical tunnelling operator is constructed using Laplace transforms instead of heat transforms. The need to take account of the one-sidedness of the Laplace transform manifests itself in the discussion in Section 1.4 and Section 5 of [11] concerning time-reversals, 'forward-backward' transformations and the question of whether the corresponding variational problem involves an inf or a sup type cost functional. We deal with this point in the next section by slightly rephrasing Day's construction (9) to include the necessary orientation information to allow the correct choice of 'side' of the Laplace transform at any point.

5. Lipschitz Continuity

In this section we show that Day's construction (9) gives a viscosity solution on Lagrangian manifolds of dimension ≤ 5 if, amongst other conditions, the Lagrangian manifold has zero Maslov index on closed curves.

By Theorem 2.2, it is sufficient to show that W is locally Lipschitz in Ω. Furthermore, by other results of Day, we only need to show that this holds at certain points of Ω. To see this, recall from Section 2 that a point $x \in \Omega$ is called a caustic point if there exists a point $(x, p) \in M$ at which the projection $\pi|_M$ onto state space is singular. Recall also the definition of a non-essential point on M made at the end of the previous section. Since Hypothesis 2.1(4) guarantees that S achieves its minimum in (9), an *essential caustic* point $x \in \Omega$ can be defined to be a caustic point for which π is singular at every $(x, p) \in M$ at which S achieves its minimum over x. Denote the set of essential caustic points by C_*. This is, of course, a subset of the set of caustics in Ω. It is shown in Theorem 2 of [11] that

Theorem 5.1. *(Day) W defined by (9) is continuous in $\Omega\backslash C_*$ and locally Lipschitz in the interior of $\Omega\backslash C_*$.*

So we can concentrate our attention on the set C_* of essential caustic points in Ω. We start by formally stating the hypotheses under which we will work. These will be assumed to hold in addition to those already assumed by Day and listed above in Hypotheses 2.1. The same notation applies as there. In the following, for any $(x,p) \in M$, let γ_p denote the integral curve corresponding to H which passes through (x,p), i.e. the integral curve for the canonical equations

$$\dot{x} = \partial H/\partial p, \ \dot{p} = -\partial H/\partial x \tag{15}$$

where H is the Hamiltonian function referred to in Hypotheses 2.1(2). From the fact that M is Lagrangian and H is constant on M, it follows that γ_p lies on M. The term phase flow corresponding to H refers to the transformation of phase space given by the solution to (15). Also, in the following let Σ denote the set of singular points for $\pi|_M$ on M.

Hypotheses 5.2.

1. The Maslov index of any closed curve on M is zero.
2. There exists a neighbourhood $U \subset M$ on which the projection $\pi|_M$ onto state space is non-singular and such that $\pi(U) \subset \Omega$.
3. For any point $(x,p) \in M$ lying over $x \in \Omega$, the integral curve γ_p passes through U and intersects Σ at most once between (x,p) and U.
4. For any $x \in \Omega$ downstream of U, let $(x,p^*) \in M$ be a point at which $S(x,.)$ achieves its minimum over all p such that $(x,p) \in M$. Then the integral curve γ_{p*} does not intersect Σ between (x,p^*) and U. Similarly, for $x \in \Omega$ upstream of U, let $(x,p^*) \in M$ be a point at which $-S(x,.)$ achieves its minimum over M. Then γ_{p*} does not intersect Σ between (x,p^*) and U. The notions of upstream and downstream will be defined shortly.
5. For any $x \in \Omega$ and any (x,p), $(x,q) \in M$ which lie over x, let s and t respectively denote the parameterisations of γ_p and γ_q given by the phase flow corresponding to H. Then traversing both γ_p and γ_q in the direction of increasing s and t, respectively, corresponds to travelling either away from U towards both (x,p) and (x,q) or away from both (x,p) and (x,q) towards U.
6. $H(x,p)$ is convex in p for each $x \in \Omega$.

Some comments about the above are in order. As with Day's hypotheses, they are attempts to extract the geometrical essence of the problem independent of any variational interpretation. We examine each of the hypotheses in turn in the context of two main classes of Lagrangian manifolds.

The first class is those arising from finite time variational problems where, if M_0 denotes the Lagrangian manifold corresponding to the initial or final cost term, then M is the $(n+1)$-dimensional Lagrangian manifold traced out in $\mathbb{R}^{2n+2}$ phase space by the evolution of M_0 along the Hamiltonian flow. This M is clearly simply connected. The fact that $H = 0$ on M follows, as mentioned above after equation (5), from the technique used to extend phase space from $2n$ to $(2n+2)$ dimensions.

As already noted, the Lipschitz property is easily established in the finite time case via other arguments. However, this is not the case for the second class of manifolds, namely those arising as stable or unstable Lagrangian manifolds of hyperbolic equilibrium points of Hamiltonian systems. These are the manifolds in which we are primarily interested and they correspond to stationary solutions to HJB equations arising from infinite time variational problems. Stable and unstable manifolds are not in general simply connected. However, as shown in Section 5 of [11], it is easy to construct large portions of them which are, as follows. Let $x = 0$, $p = 0$ be a hyperbolic equilibrium point for the dynamics given by (15) and let M^+ denote the associated stable manifold. Let Ω be an open region in state space containing 0 with the following properties:

(a) Ω is covered by M^+ - i.e. for every $x \in \Omega$ there is some $(x, p) \in M^+$ and
(b) Ω is forward invariant for (15) - i.e. for every $(x, p) \in M^+$ with $x \in \Omega$, if γ_p denotes the integral curve for (15) with $\gamma_p(0) = (x, p)$, then $\pi(\gamma_p(t)) \in \Omega$ for all $t \geq 0$.

Then the required Lagrangian manifold M is the submanifold of M^+ consisting of those $(x, p) \in M^+$ with $x \in \Omega$. This M is clearly simply connected since it can be pulled back onto a simply connected neighbourhood of $(0, 0)$ on M^+. As noted in [32], M is Lagrangian since the canonical two-form ω is invariant under the phase flow and all vector fields on M vanish as the phase flow converges to $(0, 0)$. Similarly, $H = 0$ on M since H is constant on phase curves and these all converge to $(0, 0)$ at which $H = 0$.

For the the unstable manifold M^- associated with $(0, 0)$, the same construction gives a simply connected Lagrangian submanifold M provided the region Ω is backward invariant for (15).

We can now look at the reasonableness of the above hypotheses in the context of these two classes of Lagrangian manifolds. Note first that Hypothesis 5.2(1) is equivalent to the vanishing of the Maslov class $g^* \in H^1(M, \mathbb{R})$. This will clearly be satisfied if M is simply connected, which is true for both classes of examples.

Hypothesis 5.2(2) is also natural for both classes. In the finite time case, U would be given by a small time neighbourhood of M_0 on M. In the infinite time case, U would be a small neighbourhood of the equilibrium point on the stable or unstable manifold. The fact that U has a non-singular projection onto state space follows from the assumptions which guarantee that the equilibrium point is hyperbolic. These typically take the form of assumptions on the linearised dynamics at the equilibrium point which guarantee the existence of a local smooth solution to the infinite time variational problem.

Hypothesis 5.2(3) is again natural. In the finite time case, it follows from the fact that every point on M will flow into or out of the initial manifold M_0 under the Hamiltonian flow. In the infinite time case, it follows from the definition of a stable or unstable manifold. The requirement that γ_p intersects Σ at most once restricts attention to regions of the manifold on which the Hamiltonian trajectories from U have passed through at most one singularity. We will return to this in the final section of the paper.

There is some content in Hypothesis 5.2(4). It would follow from the differentiability of the value function along an optimal state trajectory. This is well known for the finite time case provided the optimal trajectory is unique - see for instance [7, 5]. Points at which two optimal trajectories meet correspond to the formation of shocks in the evolution of M viewed as the graph in phase space of the vector function $p = p(t, x)$ - see Section

4.2 of [11] and [6]. At such points, the value function is no longer differentiable and the minimising point (x, p^*) in formula (9) jumps from one branch of M to another. The results below show that the jump cannot occur at a singular point in the projection of M onto state space. In the infinite time case, it would seem reasonable to expect this hypothesis to be satisfied on some sizeable submanifold, constructed as above, of a stable or unstable manifold. We leave for the future the question of deciding just how restrictive this hypothesis is.

Hypothesis 5.2(5) rules out the possibility of reaching (x, p) from U by following the phase flow in forward time along γ_p while reaching (x, q) by following it in reverse time along γ_q. This means that all branches of M lying over the same point $x \in \Omega$ can be given the same H-dependent 'orientation' with respect to U. This is the orientation referred to at the end of the previous section in connection with the choice of the correct side in the $1/h$-Laplace transform used to construct a canonical tunnelling operator. This hypothesis allows us to make the following definition.

Definition 5.3. If the direction of travel along γ_p from U to (x, p) corresponds to following the phase flow for H in forward time then we will say that x is *downstream* from U, otherwise x is *upstream* from U.

Lastly, we have included the convexity of H (already required by Theorem 2.2) as an explicit hypothesis in 5.2(6) because it is central to the following arguments. We comment in the final section about how the approach of this paper may be extended to cover the case where H is neither convex nor concave.

There are two more hypotheses which we will require, one of so-called 'transverse connectivity' and the other a compactness condition, but their statements require certain technical notions to be developed later in this section. They will be stated after those developments as Hypotheses 5.12 and 5.22. The first is a general position assumption and so is not restrictive in the sense that the space of manifolds which satisfies it is dense in the space of all Lagrangian manifolds. The second excludes certain cases which we consider to be pathological. As with Hypothesis 5.2(4), we leave for the future the question of deciding just how restrictive this exclusion is.

We now rephrase Day's construction (9) to take account of the orientation given by Hypothesis 5.2(5). The motivation for doing so is given by Example 4.1.1 of [11] in which Day considers the variational problem

$$W(t, x) = \inf_{x(t)=x} \left\{ \Phi(x_0) + \int_0^t L(\dot{x}(s))ds \right\}$$

with convex integrand L. He applies his construction (9) on the $(n+1)$-dimensional Lagrangian manifold M formed by the evolution of the initial manifold $(x_0, \partial\Phi/\partial x_0)$ under the Hamiltonian flow. The convex Hamiltonian on M is given by $H^+(t, x, \sigma, p) = \sigma + H(p)$ where H is the Legendre transformation of L and (t, σ) are the extra canonical coordinates required to extend $\mathbb{R}^{2n}$ phase space to $\mathbb{R}^{2n+2}$, as described after equation (5) above. On M, $H^+ = 0$, i.e. $\sigma = -H$. As is obvious from the variational interpretation, the function defined by (9) gives a Lipschitz viscosity solution to $H^+ = 0$ on the portion downstream from the initial manifold, confirming Theorem 2.2. However, the function so defined is discontinuous on the upstream portion of M.

To obtain a viscosity solution on the upstream portion of M using a variational argument, the correct variational problem to consider would be

$$\begin{aligned} W(-t,x) &= \inf_{x(-t)=x}\left\{\Phi(x_0)+\int_0^{-t} L(\dot{x}(s))ds\right\} \\ &= -\sup_{x(-t)=x}\left\{\int_{-t}^{0} L(\dot{x}(s))ds-\Phi(x_0)\right\} \end{aligned}$$

where $t>0$. So the cost functional is of supremum type, the value function to solve for is $\hat{W}=-W$ and it satisfies the final condition $\hat{W}(0,x_0)=-W(0,x_0)=-\Phi(x_0)$, which matches the initial condition for the downstream portion. Also, $p=\partial W/\partial x=-\partial\hat{W}/\partial x$ and $\sigma=-H=\partial W/\partial t=-\partial\hat{W}/\partial t$. Now to construct $\hat{W}$ from M the analogue of formula (9) would be

$$\hat{W}(x)=\sup\{S(x,p): p \text{ such that } (x,p)\in M\}. \tag{16}$$

It is clear from either a variational argument or from inspection of the pictures in Example 4.1.1 of [11] that $\hat{W}$ is Lipschitz on the upstream portion of M. Since $-H^+(t,x,\sigma,p)=0$ is concave with respect to σ and p, the analogue of Theorem 2.2 says that $\hat{W}$ is a viscosity solution of $-H^+(t,x,-\partial\hat{W}/\partial t,-\partial\hat{W}/\partial x)=0$ on the upstream portion of M. By Remark 1.4 of [9], this is equivalent to $W=-\hat{W}$ being a viscosity solution of $H^+(t,x,-\partial W/\partial t,-\partial W/\partial x)=0$ on the upstream portion of M.

So if $H^+(t,x,\sigma,p)=0$ on all of M, then the appropriate equation to solve on the upstream portion of M is $H^+(t,x,-\partial W/\partial t,-\partial W/\partial x)=0$ and the correct definition for a viscosity solution on this portion of M is $W=-\sup\{S(x,p)\}$, rather than equation (9). On the downstream portion, the appropriate equation to solve is $H^+(t,x,\partial W/\partial t,\partial W/\partial x)=0$ and the correct definition for a viscosity solution is given by (9). Given that Hypothesis 5.2(5) allows a consistent definition on M of the notions of upstream and downstream, we are thus motivated to redefine Day's construction (9) as follows.

Definition 5.4. If $x\in\Omega$ is downstream of U then define

$$W(x)=\inf\{S(x,p): p \text{ such that } (x,p)\in M\}, \tag{17}$$

while if x is upstream of U then define

$$\begin{aligned} W(x) &= -\sup\{S(x,p): p \text{ such that } (x,p)\in M\} \\ &= \inf\{-S(x,p): p \text{ such that } (x,p)\in M\}. \end{aligned} \tag{18}$$

This is the sense in which, viewing the above construction as the logarithmic limit of a canonical tunnelling operator based on a $1/h$-Laplace transform, then the orientation given by Hypothesis 5.2(5) enables a coherent choice of side in the transform.

It is clear, from the discussion preceding the above definition, that Theorems 2.2 and 5.1 apply in the same way to W defined by (17) and (18) as to W defined by (9). It is also clear that if, in Hypothesis 5.2(6), H is assumed to be concave instead of convex and if, in Hypothesis 5.2(4), the reference to 'minimum' is replaced by 'maximum', then all the results of this paper still apply provided (i) inf is replaced by sup and S by $-S$ in formulae (17) and (18) and (ii) the appropriate equations to solve are taken to be $H(x,\partial W/\partial x)=0$ upstream and $H(x,-\partial W/\partial x)=0$ downstream. This is equivalent to reversing time and swapping the definitions of which portions of M are upstream and downstream. It is well known in the literature on control and viscosity solutions that reversing time corresponds to solving the equation $-H = 0$ instead of $H = 0$, and thus concave Hamiltonians are converted into convex ones. This is used by Day in Section 5 of [11] in order to convert the supremum cost functional and concave Hamiltonian for an L_2-gain problem into an infimum cost functional and convex Hamiltonian to which the construction (9) can be applied. It is interesting to note that this observation is natural from the perspective of symplectic geometry. This is because, as noted above, $(t, -H)$ are the pair of symplectic coordinates used to embed $\mathbb{R}^{2n}$ phase space into $\mathbb{R}^{(2n+2)}$ phase space. So if time is reversed, i.e. its sign is changed, then the sign of H must be changed as well in order to preserve the orientation of the canonical 2-form $\omega = dy \wedge dx - dH \wedge dt$ on $\mathbb{R}^{(2n+2)}$.

The next step is a key use of the convexity of H to show that the Maslov index is monotonic along any of the Hamiltonian integral curves lying on M.

Lemma 5.5. *For $x \in \Omega$ and any $(x, q) \in M$, let γ_q denote the integral curve for H which lies on M and which connects the point (x, q) to U. Consider the direction of travel along γ_q which corresponds to following the phase flow for H in forward time. Then the Maslov index is non-decreasing along γ_q, with respect to this direction of travel.*

Proof. Recall that $H(x, p)$ is assumed to be C^2. Further, since it is convex in p for all x, it follows that at any x and p, the Hessian of H with respect to p is positive semi-definite, i.e. for any $h, \hat{x}, \hat{p} \in \mathbb{R}^n$,

$$h^T \frac{\partial^2 H}{\partial p^2} \mid_{\hat{x},\hat{p}} h \geq 0.$$

Let Σ denote the set of singular points on M and let $(\hat{x}, \hat{p})$ be a point of intersection of γ_q with Σ. Recall from the definition of the Maslov index (10) that, up to an arbitrarily small deformation in the class of Lagrangian manifolds, we can assume that γ_q intersects Σ transversely in a simple singular point. Thus, in the neighbourhood of $(\hat{x}, \hat{p})$, points on Σ are defined by the condition $\partial x^i / \partial p_i = 0$ for some $i \in \{1, ..., n\}$. Let e_i denote the standard basis vector with a 1 in the ith position and zeros elsewhere and let t denote the parameterisation of γ_q given by the solution of $\dot{x} = \partial H / \partial p$, $\dot{p} = -\partial H / \partial x$. Then at $(\hat{x}, \hat{p})$

$$\frac{d}{dt} \frac{\partial x^i}{\partial p_i} = \frac{\partial}{\partial p_i} \frac{dx^i}{dt} = \frac{\partial^2 H}{\partial p_i^2} = e_i^T \frac{\partial^2 H}{\partial p^2} e_i \geq 0,$$

i.e. at $(\hat{x}, \hat{p})$ the dervative $\partial x^i / \partial p_i$ cannot decrease. Since $(\hat{x}, \hat{p})$ is a point of transversal intersection with Σ, this derivative has to change sign. Therefore it must go from negative to positive. It follows that the Maslov index of γ_q with the given orientation must increase by $+1$ at $(\hat{x}, \hat{p})$. □

Corollary 5.6. *Suppose $x \in \Omega$ is downstream of U and $(x, q) \in M$. Then the Maslov index of γ_q is 0 or $+1$ when traversed from U to (x, q). In particular, if $(x, p^*) \in M$ is the point at which $S(x, .)$ achieves its minimum over all p such that $(x, p) \in M$, then the Maslov index of γ_{p*} equals 0. Conversely, if $x \in \Omega$ is upstream of U, then the Maslov index of γ_q is 0 or -1 when traversed from U to (x, q) and, if $(x, p^*) \in M$ is the point at which $-S(x, .)$ achieves its minimum over all p such that $(x, p) \in M$, then the Maslov index of γ_{p*} equals 0.*

Proof. The Maslov index of any curve lying in U is zero and γ_q intersects Σ at most once between U and (x,q). Furthermore, γ_{p*} does not intersect Σ between U and (x,p^*). $\square$

The main result of this section is to show that, what we will call, folded singularities on M cannot be minimising points for S downstream of U (or $-S$ upstream of U). The main tool used to prove this will be the Maslov index. These arguments, as with all the above references to Maslov index, make use of the notion of transversality or general position, i.e. they hold up to an arbitrarily small perturbation of M within the class of Lagrangian manifolds. In particular, we will make extensive use of specific assumptions about the transversality of a connecting curve between points (x,p) and (x,q) on M. For clarity, we state these assumptions explicitly.

Definition 5.7. A regular curve segment in state space is an embedding $b:[0,1]\to\mathbb{R}^n$ such that the closed curve $\{b(s) : s\in[0,1]\}$ has no self-intersections. In particular, $b(0)\neq b(1)$.

Definition 5.8. Let M be a Lagrangian manifold in $\mathbb{R}^{2n}$ phase space and let (x,p) and (x,q) be points on M lying over the same point x in $\mathbb{R}^n$ state space. Then (x,p) and (x,q) are said to be connected over a regular curve segment in state space if there exists an embedded curve $\beta:[0,1]\to M$ such that

1. $\beta(0)=(x,p)$ and $\beta(1)=(x,q)$,
2. β has no self-intersections and
3. the projection $\pi(\{\beta(t) : t\in[0,1]\})$ of β onto state space coincides with the image of a regular curve segment b in state space, i.e. $\pi(\{\beta(t) : t\in[0,1]\})=\{b(s) : s\in[0,1]\}$.

Definition 5.9. Suppose (x,p) and $(x,q)\in M$ are connected over a regular curve segment b in state space. So there exists some $s\in[0,1]$ such that $b(s)=x$. We can choose an open interval I of $\mathbb{R}$ containing $[0,1]$ and extend b to an embedding $b:I\to\mathbb{R}^n$ such that the image $b(I)$ is contained in Ω and has no self-intersections. Then (x,p) and (x,q) are transversely connected over a regular curve segment in state space if for all $t\in[0,1]$, the maps b and $\pi|_M$ are transverse at $\pi(\beta(t))$. This means that if $\hat{x}=\pi(\beta(t))=b(s)$ for some $s\in[0,1]$ and if U_1 denotes a small neighbourhood of $\beta(t)$ on M then $T_{\hat{x}}(\pi|_{U_1}(M))+T_{\hat{x}}(b(I))=\mathbb{R}^n$.

Definition 5.10. M is said to have the property that all its branches are transversely connected over regular curve segments in state space if the property holds for any pair (x,p) and $(x,q)\in M$.

Example 5.11. Suppose M is the Lagrangian manifold traced out in extended $\mathbb{R}^4$ phase space by an initial manifold $M_0=\{x,\partial S_0/\partial x\}$ in $\mathbb{R}^2$ phase space under the Hamiltonian flow corresponding to the finite time Cauchy problem

$$H(x,\partial S/\partial x)=-\partial S/\partial t,\quad S(x,0)=S_0(x).$$

Then all branches on M are transversely connected over regular curve segments in state space.

Proof. If $(x,p),(x,q)\in M$ then, since the 2nd coordinate x_2 in state space is time t, it follows that (x,p) and (x,q) are simultaneous. So they can be connected by a curve of simultaneous points on M. Since the Hamiltonian flow has everywhere a unit component in the time direction, (x, p) and (x, q) are transversely connected. □

In the general case, which includes higher dimensions and infinite time problems, transverse connectivity can be achieved in the same way as other transversality conditions, i.e. by an arbitrarily small perturbation within the class of Lagrangian manifolds. Thus, the following hypothesis is not restrictive in that it is satisfied by a space of manifolds which is dense in the space of all Lagrangian manifolds.

Hypothesis 5.12. All branches of M are transversely connected over regular curve segments in state space.

We saw in Section 3 how the Maslov index behaves on closed curves under pull-back. We now examine how it behaves on transverse connecting curves under pull-back. By definition, these are not closed. The pull-back we consider is into a two dimensional phase space. We aim, eventually, to evaluate S along the pulled back connecting curve in this 2-dimensional phase space.

Lemma 5.13. *Let (x, p) and $(x, q) \in M$ be transversely connected over a regular curve segment in state space. With the notation of the preceding definitions we can consider $I \subset \mathbb{R}$ to be embedded in the s-coordinate axis (i.e. the state coordinate) in $\mathbb{R}^2$ phase space with symplectic coordinates (s, r). Let $b^*(\beta)$ denote the pull-back under b of the curve β from $\mathbb{R}^{2n}$ phase space into $\mathbb{R}^2$ phase space. Then $b^*(\beta)$ is a Lagrangian submanifold of $\mathbb{R}^2$. Furthermore, if β_1 is any sub-segment of β and ind denotes the Maslov index, then ind$(b^*(\beta_1)) =$ ind(β_1), where the indices are calculated with respect to some parameterisation of β_1 and the corresponding induced parameterisation on $b^*(\beta_1)$.*

Proof. Given the smooth map $b : I \longrightarrow \mathbb{R}^n$ and the curve β on M lying over $\{b(s) : s \in [0, 1]\} \subset \mathbb{R}^n$, the pull-back $b^*(\beta)$ in $\mathbb{R}^2$ phase space is defined as the submanifold

$$\left\{ (\hat{s}, \hat{r}) : \hat{r} = \hat{p}\frac{dx}{ds} \mid_{\hat{x}}, \hat{x} = b(\hat{s}) \text{ and } (\hat{x}, \hat{p}) \in \beta \subset \mathbb{R}^{2n} \right\}$$

of $\mathbb{R}^2$. Note that this is well defined since for any given $(\hat{x}, \hat{p}) \in \beta$, there is a unique $\hat{s}$ such that $\hat{x} = b(\hat{s})$ and the corresponding $\hat{r}$ such that $(\hat{s}, \hat{r}) \in b^*(\beta)$ is then fixed by $\hat{r} = \hat{p}(dx/ds) \mid_{\hat{x}}$. Now it follows, by transversality, that $b^*(\beta)$ is Lagrangian (see for example Proposition 3.3.11 of [31]).

Next, consider a point $(\hat{x}, \hat{p})$ on β which does not lie on the singular locus Σ_M for the projection of M onto $\mathbb{R}^n$ state space. Then in a neighbourhood of $(\hat{x}, \hat{p})$ on M, M can be parameterised by x, i.e. it can be expressed in the form $\{(x, p(x)) : x \in$ open set in $\mathbb{R}^n\}$ for some smooth vector function p of x. Now, as noted above, there is a unique $(\hat{s}, \hat{r}) = b^*(\hat{x}, \hat{p})$. Then in a neighbourhood of $(\hat{s}, \hat{r})$ on $b^*(\beta)$, we see that $b^*(\beta)$ can be expressed in the form $\{(s, p(b(s))(db/ds)|_{b(s)}) : s \in$ open set in $\mathbb{R}\}$. In other words, $b^*(\beta)$ can be parameterised by s and so $(\hat{s}, \hat{r})$ is not a point of intersection of $b^*(\beta)$ with $\Sigma_{b^*(\beta)}$, the singular locus for the projection of $b^*(\beta)$ onto $\mathbb{R}$ state space.

Conversely, suppose $(\hat{x}, \hat{p})$ on β is a point of intersection with Σ_M. Then $\pi|_M$ is singular at $(\hat{x}, \hat{p})$. But, by transversality, if U_1 denotes a small neighbourhood of $(\hat{x}, \hat{p})$ on M, then $T_{\hat{x}}(\pi|_{U_1}(M)) + T_{\hat{x}}(b(I)) = \mathbb{R}^n$. Since dim$T_{\hat{x}}(b(I)) = 1$, the rank of $\pi|_M$ must drop by only

1 at $(\hat{x}, \hat{p})$ and, furthermore, the direction in which $\pi|_M$ becomes singular must lie along the tangent to b at $\hat{x} = b(\hat{s})$. Thus, if $\pi_{\mathbb{R}^2}$ denotes the projection of (s, r)-phase space onto (s)-state space and if $(\hat{s}, \hat{r}) = b^*(\hat{x}, \hat{p})$, then $\pi_{\mathbb{R}^2}|_{b^*(\beta)}$ is singular at $(\hat{s}, \hat{r})$. So $(\hat{s}, \hat{r})$ is a point of intersection of $b^*(\beta)$ with $\Sigma_{b^*(\beta)}$. A similar transversality argument shows that the orientation of both Σ_M and $\Sigma_{b^*(\beta)}$ as cycles in their respective Lagrangian manifolds are the same.

Now consider the parameterisation of β given by t, i.e. the smooth map $\beta : t \longmapsto (x(t), p(t))$. Since b is injective, this induces a smooth parameterisation of $b^*(\beta)$, namely

$$t \longmapsto (s(t), r(t)) = \left(b^{-1}(x(t)), p(t) \left(\frac{db}{ds} \right) |_{b(s)=x(t)} \right). \tag{19}$$

Furthermore, since rank$\beta = 1$ for all t, it follows that the rank of the above map onto $b^*(\beta)$ is also equal to 1 for all t. To see this, note that if $dx/dt \neq 0$, then $ds/dt \neq 0$ since b^{-1} is an injective map from $\{b(s) : s \in [0, 1]\} \subset \mathbb{R}^n$ to $[0, 1]$.

Conversely, if $dx/dt = 0$, then $dp/dt \neq 0$, since rank$\beta = 1$. We also have $db/ds \neq 0$ since rank$b = 1$ for all s. Now, since β lies on M, a vector tangent to β is also tangent to M. Then, from the fact that $dx/dt = 0$, it follows that M has a tangent vector $v = (dx/dt, dp/dt)$ at $(x(t), p(t))$ which has null projection onto state space. In other words, $\pi|_M$ is singular in a neighbourhood of $(x(t), p(t))$ on M. By the transverse connectivity condition, the rank of $\pi|_M$ can only drop by one and, as above, the direction in which it becomes singular lies along $(db/ds)|_{x(t)}$. So we can find vectors $w_i = (w_{xi}, w_{pi})$, $i = 1, ...n-1$ spanning the remaining $(n-1)$-dimensions of the tangent space to M at $(x(t), p(t))$, such that each $w_{xi} \neq 0$ and is orthogonal to $(db/ds)|_{x(t)}$ in state space. Now, if $\omega = dp \wedge dx$ denotes the canonical two form in $\mathbb{R}^{2n}$, then for each $i = 1, ..., n-1$,

$$\omega(v, w_i) = (dp/dt).w_{xi} - (dx/dt).w_{pi} = 0$$

since v and w_i are tangent vectors to a Lagrangian manifold. But then, since $dx/dt = 0$, it follows that dp/dt, considered as a vector in $\mathbb{R}^n$, is orthogonal to each w_{xi}, $i = 1, ..., n-1$. It follows that dp/dt is colinear with $(db/ds)|_{x(t)}$, considered as vectors in $\mathbb{R}^n$. Thus,

$$\frac{dr}{dt} = \frac{dp}{dt}. \left(\frac{db}{ds} \right) |_{x(t)} \neq 0$$

as required.

So now, with respect to these parameterisations of both β and $b^*(\beta)$ by t, it follows from paragraphs two and three of this proof that any subsegment β_1 of β intersects the singular cycle Σ_M the same number of times with the same orientation as the pulled-back subsegment $b^*(\beta_1)$ intersects $\Sigma_{b^*(\beta)}$. The Maslov indices of the subsegment and its pull-back are therefore equal. □

In order to evaluate S along the pulled back connecting curve $b^*(\beta)$, we need to establish conditions under which there exists a symplectic diffeomorphism from a 2-dimensional neighbourhood V_1 of $b^*(\beta)$ in $\mathbb{R}^2$ onto a 2-dimensional neighbourhood V_2 of β in $\mathbb{R}^{2n}$. This is done in the next three lemmas.

Lemma 5.14. *Let (x,p) and $(x,q) \in M$ be transversely connected over a regular curve segment in state space. With the notation of the preceding lemma, let β_t for $t \in (0,1]$ denote an initial open segment of β, i.e. $\beta_t = \{\beta(\tau) : \tau \in (0,t]\}$. Suppose that for all $t \in (0,1]$, $ind(\beta_t)$ is equal to either 0 or -1. Then $b^*(\beta)$ has no self-intersections in $\mathbb{R}^2$. In particular, for the points $(x,p) = \beta(0)$ and $(x,q) = \beta(1)$, $b^*(x,p) \neq b^*(x,q)$.*

Proof. Suppose there exists a point $(\hat{s},\hat{r})$ of self intersection on $b^*(\beta)$. Then there exist distinct points $(\hat{x},\hat{p}_1)$ and $(\hat{x},\hat{p}_2)$ on β with $\hat{x} = b(\hat{s})$ and

$$\hat{r} = \hat{p}_1 \frac{dx}{ds} \mid_{\hat{x}} = \hat{p}_2 \frac{dx}{ds} \mid_{\hat{x}} .$$

The points $(\hat{x},\hat{p}_1)$ and $(\hat{x},\hat{p}_2)$ on β are parameterised by t, say $(\hat{x},\hat{p}_1) = \beta(t_1)$ and $(\hat{x},\hat{p}_2) = \beta(t_2)$ for $t_1 < t_2 \in [0,1]$. So, if we let $b^*(\beta)(t)$ denote the parameterisation of $b^*(\beta)$ with respect to t given in (19), then $(\hat{s},\hat{r}) = b^*\beta(t_1) = b^*\beta(t_2)$. Now, it was shown in the proof of the previous lemma that this parameterisation of $b^*(\beta)$ has rank 1 for all t. So $b^*(\beta)(t)$ cannot 'double back' on itself in $\mathbb{R}^2$, i.e. using the notation of equation (19), there is no point on $b^*(\beta)$ at which $ds/dt = dr/dt = 0$. It follows that the image $\{b^*\beta(t) : t \in [t_1,t_2]\}$ is a closed curve in $\mathbb{R}^2$ formed by the intersection of $b^*(\beta)$ with itself at the point $(\hat{s},\hat{r})$.

We can assume, up to an arbitrarily small perturbation, that $b^*(\beta)$ intersects itself transversely at points such as $(\hat{s},\hat{r})$. So these points are isolated on $b^*(\beta)$. There are thus finitely many such points since $b^*(\beta)$ is compact.

Consider now the set of all pre-image points of $(\hat{s},\hat{r})$, i.e. the set of points on β defined by

$$\varphi(\hat{s},\hat{r}) = \{(\hat{x},\hat{p}) \in \beta : \hat{x} = b(\hat{s}),\ \hat{r} = \hat{p}(dx/ds) \mid_{\hat{x}}\}.$$

Let the elements of $\varphi(\hat{s},\hat{r})$ be ordered by the parameterisation of β with respect to t. It follows, from the fact that the parameterisation $b^*(\beta)(t)$ has rank 1 for all t, that the elements of $\varphi(\hat{s},\hat{r})$ are isolated on β. So, since β is compact, there are only finitely many of them.

Let $t_1(\hat{s},\hat{r})$ be the value of t on the first occasion that $b^*\beta(t)$ passes through $(\hat{s},\hat{r})$. Then the point $(\hat{x},\hat{p}_1) = \beta(t_1(\hat{s},\hat{r}))$ is the minimal element of $\varphi(\hat{s},\hat{r})$. Similarly, let $t_2(\hat{s},\hat{r})$ be the value of t on the second occasion that $b^*\beta(t)$ passes through $(\hat{s},\hat{r})$. Then $(\hat{x},\hat{p}_2) = \beta(t_2(\hat{s},\hat{r}))$ is the second element of $\varphi(\hat{s},\hat{r})$.

Now define a function f from the set of self intersection points on $b^*(\beta)$ into the interval $[0,1]$ by $f(\hat{s},\hat{r}) = t_2(\hat{s},\hat{r})$. Since this function is bounded below and its domain is finite, it must achieve a minimum value at some point of self intersection on $b^*(\beta)$. Denote this point by $(\hat{v},\hat{u})$. It follows that, in terms of the parameterisation of $b^*(\beta)$ with respect to t, $t_2(\hat{v},\hat{u})$ is the first time at which $b^*(\beta)$ intersects itself and it does so at the point $(\hat{v},\hat{u})$.

Thus, the initial segment of $b^*(\beta)$ from $t = 0$ up to $t = t_1(\hat{v},\hat{u})$ does not intersect itself. By the previous lemma, the Maslov index of this segment is equal to the Maslov index of the corresponding initial segment $\beta_{t_1(\hat{v},\hat{u})}$ of β. By hypothesis this is either 0 or -1.

Furthermore, the segment of $b^*(\beta)$ from $t = t_1(\hat{v},\hat{u})$ up to $t = t_2(\hat{v},\hat{u})$ is a simple closed curve in $\mathbb{R}^2$ with no self intersections. The Maslov index of this segment must therefore be ± 2.

Now the Maslov index of the initial segment of $b^*(\beta)$ from $t = 0$ up to $t = t_2(\hat{v}, \hat{u})$ must equal the sum of the indices of the segments from $t = 0$ up to $t = t_1(\hat{v}, \hat{u})$ and $t = t_1(\hat{v}, \hat{u})$ up to $t = t_2(\hat{v}, \hat{u})$. But it also must equal the index of the corresponding initial segment $\beta_{t_2(\hat{v},\hat{u})}$ of β which, again by hypothesis, is either 0 or -1. No such combination is possible. This contradiction shows that there are no points of self intersection on $b^*(\beta)$. $\square$

Lemma 5.15. *Let (x, p) and $(x, q) \in M$ be transversely connected over a regular curve segment in state space and, with the notation of the preceding lemmas, suppose that for all $t \in (0, 1]$, ind(β_t) = 0 or -1. Then the natural map, which we denote by φ, from $b^*(\beta) \to \beta$ is a smooth embedding of $b^*(\beta)$ as a submanifold of $\mathbb{R}^{2n}$.*

Proof. Consider a point $(s, r) \in b^*(\beta)$. Then by definition of b^*, there exists a point $(x, p) \in \beta$ such that $x = b(s)$ and $r = p(dx/ds) \mid_{b(s)}$. Then φ is defined as the map $(s, r) \to (x, p)$. By the previous lemma, φ is well defined. It is clearly smooth and it is onto β by definition of b. To see that it is injective, suppose (s, r) and (v, u) both map onto (x, p). Since b has no self intersections and $x = b(s) = b(v)$, it follows that $s = v$. Then, since $r = p(dx/ds) \mid_{b(s)}$ and $u = p(dx/ds) \mid_{b(v)}$, it follows that $r = u$. So φ is a diffeomorphism onto β. Its differential is therefore a monomorphism of bundles $d\varphi : T(b^*(\beta)) \to T\mathbb{R}^{2n}$. φ is thus a smooth embedding $b^*(\beta) \to \beta \subset \mathbb{R}^{2n}$. $\square$

Lemma 5.16. *Let (x, p) and $(x, q) \in M$ be transversely connected over a regular curve segment in state space and suppose that for all $t \in (0, 1]$, ind(β_t) = 0 or -1. Then, with the notation of the preceding lemmas, there exists a 2-dimensional neighbourhood V_1 of $b^*(\beta)$ in $\mathbb{R}^2$ and a 2-dimensional neighbourhood V_2 of β in $\mathbb{R}^{2n}$ such that φ can be extended to a diffeomorphism $\hat{\varphi} : V_1 \to V_2$. Furthermore, if $\omega_1 = dr \wedge ds$ and $\omega_2 = dp \wedge dx$ are the canonical two-forms on $\mathbb{R}^2$ and $\mathbb{R}^{2n}$ phase space, then $\hat{\varphi}^*(\omega_2|_{V_2}) = \omega_1|_{V_1}$, i.e. the pull-back under φ of the symplectic structure on $\mathbb{R}^{2n}$ restricted to V_2 is equal to the symplectic structure on $\mathbb{R}^2$ restricted to V_1.*

Proof. Since $\varphi : b^*(\beta) \to \beta \subset \mathbb{R}^{2n}$ is an embedding, its differential $d\varphi$ is a monomorphism of bundles $T(b^*(\beta)) \to T\mathbb{R}^{2n}$. This can be decomposed into a composition,

$$T(b^*(\beta)) \to \varphi^*(T\mathbb{R}^{2n}) \to T\mathbb{R}^{2n}.$$

The factor bundle $\varphi^*(T\mathbb{R}^{2n})/T(b^*(\beta))$ is called the normal bundle to the embedding φ and is denoted $\nu(b^*(\beta))$. The space of this bundle has dimension $2n$. If we consider the restriction $\varphi^*(T\mathbb{R}^{2n})$ of the tangent bundle $T\mathbb{R}^{2n}$ to the submanifold $b^*(\beta)$, then it decomposes into an orthogonal direct sum of two bundles

$$\varphi^*(T\mathbb{R}^{2n}) = T(b^*(\beta)) \oplus T(b^*(\beta))^\perp$$

where $\perp$ denotes the orthogonal complement. It follows from the definition above that there is an isomorphism

$$\chi : \nu(b^*(\beta)) \simeq T(b^*(\beta))^\perp \subset T\mathbb{R}^{2n}.$$

Now let $\xi \in \nu(b^*(\beta))$ be an arbitrary vector at the point $(s, r) \in b^*(\beta)$. Then $\chi(\xi)$ is a vector in $\mathbb{R}^{2n}$ based at the point $(x, p) \in \beta$, where $x = b(s)$ and $r = p(dx/ds) \mid_{b(s)}$. Furthermore, $\chi(\xi)$ is not tangent to β at (x, p). Let $\mathbf{v} \in \mathbb{R}^{2n}$ be the end point defined by this vector. Then this defines a map

$$\psi : \xi \to \mathbf{v} : \nu(b^*(\beta)) \to \mathbb{R}^{2n}.$$

This is analogous to the definition of the standard exponential map associated with a Riemannian metric in differential geometry. It is trivial to check that the differential of this map is an isomorphism for all null vectors $\xi \in \nu(b^*(\beta))$. Thus there exists a neighbourhood $W \subset \nu(b^*(\beta))$ of the null section $b^*(\beta)$ which is mapped diffeomorphically by ψ onto a neighbourhood $X \subset \mathbb{R}^{2n}$ of the submanifold β. It is clear that the restriction of ψ to the null section $b^*(\beta)$ coincides with φ.

We can now construct $\hat{\varphi}$ as the restriction of ψ to a certain line sub-bundle of $\nu(b^*(\beta))$. To do this, recall first that $b^*(\beta)$ is a curve in $\mathbb{R}^2$ phase space with standard symplectic coordinates (s, r) where the interval I defining the domain of the curve $b : I \to \mathbb{R}^n$ is embedded in the s-coordinate axis. The canonical 2-form on this space is $\omega_1 = dr \wedge ds$. At any point $(s, r) \in b^*(\beta)$, let $\mathbf{e}_1$ denote the unit tangent to $b^*(\beta)$ in the direction of increasing t, where $b^*(\beta)$ is parameterised by t. Then there exists a unique unit vector $\mathbf{e}_2 \in \mathbb{R}^2$ such that $\omega_1(\mathbf{e}_2, \mathbf{e}_1) = +1$.

Similarly, the restriction of the canonical 2-form $\omega_2 = \sum_{i=1}^n dp_i \wedge dx_i$ on $\mathbb{R}^{2n}$ to the neighbourhood X of β defines a symplectic structure on $\nu(b^*(\beta))$. At the point $(s, r) \in b^*(\beta)$, there is a unique unit vector $\hat{\mathbf{e}}_2$ in the tangent space to the fibre of $\nu(b^*(\beta))$ at (s, r) such that $\omega_2(\hat{\mathbf{e}}_2, \mathbf{e}_1) = +1$, i.e. the plane spanned by $\mathbf{e}_1$ and $\hat{\mathbf{e}}_2$ is a non-null 2-subplane of the tangent space to the fibre of $\nu(b^*(\beta))$ at $(s, r) \in b^*(\beta)$.

The restriction of ψ to the symplectic line sub-bundle of $\nu(b^*(\beta))$ defined by $\hat{\mathbf{e}}_2$ gives a diffeomorphism onto a 2-dimensional neighbourhood V_2 of β in $\mathbb{R}^{2n}$.

On the other hand, the argument used to construct ψ can be applied to the identification of $\mathbf{e}_2$ with $\hat{\mathbf{e}}_2$ to produce a diffeomorphism from the same line sub-bundle of $\nu(b^*(\beta))$ onto a neighbourhood V_1 of $b^*(\beta)$ in $\mathbb{R}^2$.

The appropriate composition of one of these diffeomorphisms with the inverse of the other is the required map $\hat{\varphi}$.

The fact that $\hat{\varphi}$ preserves the symplectic structure at points (s, r) on $b^*(\beta)$ follows from

$$\begin{aligned}
\hat{\varphi}^* \left(\sum_{i=1}^n dp_i \wedge dx_i \right) &= \sum_{i=1}^n dp_i \wedge \left(\frac{dx_i}{ds} \right) |_{b(s)} ds \\
&= d \left(\sum_{i=1}^n p_i \left(\frac{dx_i}{ds} \right) |_{b(s)} \right) \wedge ds \\
&= dr \wedge ds.
\end{aligned}$$

That $\hat{\varphi}$ preserves the symplectic structure at other points of V_1 follows from the above canonical identification of symplectic basis vectors in $\mathbb{R}^2$ and $T(\nu(b^*(\beta)))$ and from the identification, already used in the construction of ψ, of a vector in one of these spaces with its respective end point in V_1 or V_2. □

So we have now established the existence of a symplectic diffeomorphism from a 2-dimensional neighbourhood V_1 of $b^*(\beta)$ in $\mathbb{R}^2$ onto a 2-dimensional neighbourhood V_2 of β in $\mathbb{R}^{2n}$. The required condition is that any initial open segment on β has Maslov index 0 or -1. Clearly, the same results would have been obtained under the alternative condition that any initial open segment has index 0 or +1. We will see later on that the first of these conditions holds downstream and the second upstream on M. We can now evaluate S along the pull back $b^*(\beta)$ of the transverse connecting curve between (x,p) and (x,q).

Proposition 5.17. *Let (x,p) and $(x,q) \in M$ be transversely connected over a regular curve segment in state space by a curve β on M with $\beta(0) = (x,p)$ and $\beta(1) = (x,q)$. Suppose, with the notation of the preceding lemmas, that for all $t \in (0,1]$, ind$(\beta_t) = 0$ or -1. Let S denote the smooth function defined globally on M by $dS = pdx$. Then $S(x,p)$ is not a minimising value for $S(x,\hat{p})$ over all $\hat{p}$ such that $(x,\hat{p}) \in M$. Similarly, if for all $t \in (0,1]$, ind$(\beta_t) = 0$ or +1, then $-S(x,p)$ is not a minimising value for $-S(x,\hat{p})$ over all $\hat{p}$ such that $(x,\hat{p}) \in M$.*

Proof. Suppose $\mathrm{ind}(\beta_t) = 0$ or -1. The same argument works for the other case. Let $(x(t),p(t)) = \beta(t)$ for $t \in [0,1]$. So $(x(0),p(0)) = (x,p)$ and $(x(1),p(1)) = (x,q)$. Then, for $t \in (0,1]$,

$$S(x(t),p(t)) = \int_{\beta_t} pdx + S(x,p)$$

where $\beta_t = \{\beta(\tau) : \tau \in (0,t]\} \subset M$ is the initial open segment of β from (x,p) to $(x(t),p(t))$. Now, by the previous lemmas, $\beta \subset V_2 = \hat{\varphi}(V_1)$ where V_1 is a neighbourhood of $b^*(\beta)$ in $\mathbb{R}^2$. So, since $\beta_t = \varphi(b^*(\beta_t))$,

$$S(x(t),p(t)) - S(x,p) = \int_{b^*(\beta_t)} \hat{\varphi}^*(pdx) = \int_{b^*(\beta_t)} rds.$$

Again, by the previous lemmas, the curve $b^*(\beta)$ is a Lagrangian submanifold in $\mathbb{R}^2$ with no self intersections. In addition, for any $t \in (0,1]$, the initial open segment $b^*(\beta_t)$ is Lagrangian with no self intersections and Maslov index either 0 or -1. Also, since the start and end points of β have the same x-coordinates in $\mathbb{R}^{2n}$ phase space, the start and end points of $b^*(\beta)$ have the same s-coordinate in (s,r) phase space. Furthermore, s cannot be constant on $b^*(\beta)$ since then x is constant on β which contradicts transverse connectivity over state space.

Let $(s,r(0))$ be the initial point of $b^*(\beta)$ and let $(s(t),r(t))$ be the point corresponding to $(x(t),p(t))$ on β. At the end point of $b^*(\beta)$ we have $s(1) = s$. So there exists at least one $t \in (0,1]$ such that $s(t) = s$. Let t_1 be the first such $t \in (0,1]$. So t_1 is also the parameter value of the first time that $x(t) = x$ on β for $t \in (0,1]$.

Now, by the transversality condition, singularities in the projection of $b^*(\beta)$ onto the s-axis can be assumed to be isolated on $b^*(\beta)$. So either at $(s,r(0))$ itself, or at a point arbitrarily close to it on $b^*(\beta)$ if $(s,r(0))$ is singular, the tangent vector to $b^*(\beta)$ with respect to the parameterisation by t has a non-zero s-component. We now show that if the sign of this component is negative then $r(t_1) < r(0)$. Otherwise, if it is positive then $r(t_1) > r(0)$.

To see this, suppose it is negative. Now at t_1 the point $(s(t_1),r(t_1))$ on $b^*(\beta)$ has $s(t_1) = s$ again. So the s-component of the tangent vector to $b^*(\beta)$ must have changed sign from negative to positive at some point $(s(\tau_1),r(\tau_1))$ on $b^*(\beta_{t_1})$ for $\tau_1 \in (0,t_1)$. Let this be the first such point on $b^*(\beta_{t_1})$. Then this point is the first singular point for the projection of $b^*(\beta_{t_1})$ onto the s-axis.

Now, by transversality of β, we can assume that this is a simple singular point at which $b^*(\beta)$ intersects $\Sigma_{b^*(\beta)}$ transversely. So $\mathrm{ind}(b^*(\beta_t))$ must change by ± 1 at $t = \tau_1$. Since

$(s(\tau_1), r(\tau_1))$ is the first singular point on the initial open segment $b^*(\beta_{t_1})$, it follows that $\mathrm{ind}(b^*(\beta_t)) = 0$ for $t < \tau_1$. Then, from the hypothesis, $\mathrm{ind}(b^*(\beta_t)) = -1$ for $t \in (\tau_1, \tau_1 + \delta)$ for some $\delta > 0$. So, from the definition of Maslov index (10), the r-component of the tangent to $b^*(\beta_{t_1})$ must be negative at τ_1.

If $(s(\tau_1), r(\tau_1))$ is the only such singularity on $b^*(\beta_{t_1})$, i.e. the only point at which the s-component of the tangent vector changes sign, then since the initial closed segment $\{(s, r(0))\} \cup b^*(\beta_{t_1})$ has no self intersections, it follows that $r(t_1) < r(0)$.

Otherwise, let $\tau_2 \in (0, t_1)$ be the parameter value of the next such singularity on $b^*(\beta_{t_1})$. At this point, the s-component of the tangent must change sign from positive to negative. Since the Maslov index has to change by ± 1, the only possibility allowed by the hypothesis is that it increases by $+1$ from -1 to 0. So the r-component of the tangent again must be negative at τ_2.

Now, $s(\tau_2) < s$ since t_1 is the first time on $(0, 1]$ at which $s(t) = s$. Also, at τ_2 the s-component of the tangent to $b^*(\beta_{t_1})$ goes negative again. So there must exist at least one more singular point in the projection of $b^*(\beta_{t_1})$ onto the s-axis. A repetition of the above argument shows that the r-component of the tangent must again be negative at this singular point. Again, if this is the last such singular point before t_1, then since $\{(s, r(0))\} \cup b^*(\beta_{t_1})$ has no self intersections, it follows that $r(t_1) < r(0)$.

Now $b^*(\beta_{t_1})$ is compact and, as noted above, singular points on $b^*(\beta_{t_1})$ can be assumed to be isolated. So there are only finitely many singular points on $b^*(\beta_{t_1})$. So the above argument can be repeated a finite number of times to show that $r(t_1) < r(0)$.

The same argument also shows that $r(t_1) > r(0)$ if the s-component of the tangent vector to $b^*(\beta)$ is positive at $(s, r(0))$, or at a point arbitrarily close to it on $b^*(\beta)$ if $(s, r(0))$ is singular.

So now consider the value of $S(x(t_1), p(t_1))$. By the definition of t_1, we have $x(t_1) = x$ and t_1 is the smallest $t \in (0, 1]$ such that $x(t) = x$ on β. Now, from above,

$$S(x, p(t_1)) - S(x, p) = S(x(t_1), p(t_1)) - S(x, p) = \int_{b^*(\beta_{t_1})} r ds.$$

Let l denote the straight line segment in (s, r)-phase space from $(s(t_1), r(t_1)) = (s, r(t_1))$ to $(s, r(0))$. Since s is constant on l, we have

$$\int_l r ds = 0.$$

So if C denotes the closed curve from $(s, r(0))$ to $(s, r(t_1))$ and back to $(s, r(0))$ formed by $b^*(\beta_{t_1}) \cup l$ then C has no self intersections and

$$S(x, p(t_1)) - S(x, p) = \oint_C r ds = \int\int_A dr \wedge ds$$

where A is the region of (s, r)-phase space bounded by C.

Now, consider the s-component of the tangent vector to $b^*(\beta)$ with respect to the parameterisation by t. If this is negative at $(s, r(0))$, or at a point arbitrarily close to it on $b^*(\beta)$

in the case where $(s, r(0))$ is singular, then from above $r(t_1) < r(0)$. Thus C is traversed anti-clockwise. Conversely, if it is positive at $(s, r(0))$, or at a point arbitrarily close to it on $b^*(\beta)$, then $r(t_1) > r(0)$ and again C is traversed anti-clockwise. In either case, the oriented area of A is negative and so

$$S(x, p(t_1)) < S(x, p),$$

i.e. there exists a point $(x, p(t_1)) \in M$ at which the value of S is less than at (x, p). So (x, p) is not a minimising point for $S(x, \hat{p})$ over all $\hat{p}$ such that $(x, \hat{p}) \in M$. $\square$

We now formalise the definition of a common type of singularity in the projection of M onto state space. Our aim is to show that a singular point of this type can be transversely connected to another point on M by a curve which satisfies the conditions of the previous proposition.

Definition 5.18. A folded singularity $(x, p) \in M$ is a singular point in the projection $\pi : M \to \mathbb{R}^n$ such that

1. there exists a smooth curve $\alpha : [-1, 1] \to M$ such that $\alpha(0) = (x, p)$ and $\pi(\alpha(t)) = \pi(\alpha(-t))$ for all $t \in (0, 1]$.
2. the locus $\{\pi(\alpha(t)) : t \in (0, 1]\}$ defines a straight line in state space
3. let $\mathbf{e} = \pi(\alpha(1)) - \pi(\alpha(0)) = \pi(\alpha(1)) - x$, i.e. the vector which lies on the above defined straight line in state space and 'points' away from x into $\pi(M)$. Then for any sequence $p_m \in \mathbb{R}^n$ such that $(x - \frac{1}{m}\mathbf{e}, p_m) \in M$ for all m, $p_m \nrightarrow p$ as $m \to \infty$.

Lemma 5.19. *Let $(x, p) \in M$ be a folded singularity. Then there exists a distinct point $(x, q) \in M$ for some $q \in \mathbb{R}^n$. It thus follows that (x, p) and (x, q) can be transversely connected over a regular curve segment in state space by a curve β on M.*

Proof. From the definition of a folded singularity, let $\mathbf{e} = \pi(\alpha(1)) - x$ and consider the point $x - s\mathbf{e}$ for any $s \in (0, 1]$. Since the region Ω of state space covered by M is open and since $x \in \Omega$, we can assume that $x - s\mathbf{e} \in \Omega$ for all $s < \varepsilon$ for some sufficiently small $\varepsilon > 0$. Then since M covers Ω, there exists for each $s \in (0, \varepsilon]$ a value $q(s) \in \mathbb{R}^n$ such that $(x - s\mathbf{e}, q(s)) \in M$. Furthermore, if x is downstream of U, then we can chose $q(s)$ such that $(x - s\mathbf{e}, q(s))$ is a minimising point for $S(x - s\mathbf{e}, \hat{q})$ over all $\hat{q}$ such that $(x - s\mathbf{e}, \hat{q}) \in M$. Similarly, if x is upstream, then we can chose $(x - s\mathbf{e}, q(s))$ to be a minimising point for $-S(x - s\mathbf{e}, \hat{q})$. Then by Lemma 1 of [11], the sequence $q(\frac{1}{m})$, $m > \frac{1}{\varepsilon}$, $m \in \mathbb{N}$ is bounded and for any limit point q, the point (x, q) lies on M. Chose one such limit point for a convergent subsequence $(x - \frac{1}{m}\mathbf{e}, q(\frac{1}{m}))$. Then by the definition of a folded singularity, $p \neq q$. So, by our transversality or general position assumption that all branches of M can be transversely connected over regular curve segments in state space, it follows that (x, p) and (x, q) can be so connected. $\square$

Lemma 5.20. *With the notation of the previous lemma, in a small neighbourhood of a folded singularity (x, p) on M, β can be assumed to coincide with one branch of the curve α appearing in the definition of the folded singularity.*

Proof. Consider all points on β which lie over x, including (x, p) and (x, q). Denote these points by (x, p_λ) for $\lambda \in \Lambda$. The aim is to apply at each (x, p_λ) an arbitrarily small deformation to β which

1. is in a direction orthogonal to β and along M and which preserves the condition of transversality of β
2. is such that the deformed curve, which we denote by $\hat{\beta}$, still passes through each (x, p_λ)
3. is such that all the branches of $\hat{\beta}$ for $\lambda \in \Lambda$ project onto the same deformation $\hat{b}$ of the regular curve segment b in state space
4. is such that the vector $\mathbf{e}$ appearing in the definition of the folded singularity is tangent to $\hat{b}$ at x.

Recall that, by definition, $\mathbf{e}$ is tangent at x to the curve $\{\pi(\alpha(t)) : t \in (0, 1]\}$ in state space and so this last requirement will ensure that $\hat{\beta}$ coincides with one branch of α at (x, p).

Let U_1 be a neighbourhood of (x, p) on M. Then, by the definition of a folded singularity, $\mathbf{e} \notin T_x(\pi|_{U_1}(M))$. Also, by transversality, the rank of π drops by at most 1 at (x, p). It follows that

$$T_x(\pi|_{U_1}(M)) + \mathrm{span}(\mathbf{e}) = \mathbb{R}^n.$$

Let U_λ be a neighbourhood of (x, p_λ) on M and let $\mathbf{f}$ denote a tangent vector at x to the curve b in state space. Now, for each λ, the projected curve $\pi|_{U_\lambda}(\beta)$ coincides with the locus of the curve b in state space. So, for each λ, a vector which is tangent to $\pi|_{U_\lambda}(\beta)$ at x must lie in span($\mathbf{f}$). Note that, for those λ for which $\pi|_{U_\lambda}$ is non-singular at (x, p_λ), any vector in span($\mathbf{f}$) is tangent to $\pi|_{U_\lambda}(\beta)$ at x. However, for those λ for which $\pi|_{U_\lambda}$ is singular at (x, p_λ), only one direction in span($\mathbf{f}$) will give a well defined tangent to $\pi|_{U_\lambda}(\beta)$ at x.

Now, by transversality, at each λ

$$T_x(\pi|_{U_\lambda}(M)) + \mathrm{span}(\mathbf{f}) = \mathbb{R}^n$$

i.e. if $\pi|_{U_\lambda}$ is singular at (x, p_λ) then the rank of π drops by 1 in the direction spanned by $\mathbf{f}$. If $\pi|_{U_\lambda}$ is non-singular, then the above equation holds trivially. Thus, in each neighbourhood U_λ, we can find a deformation of β which, when projected onto state space, takes $\mathbf{f}$ into $\mathbf{e}$. These deformations satisfy the above four requirements. □

Having established the existence of the point (x, q), we now deal with the possibility that $\pi|_M$ is singular at (x, q). Note, there may be more than one point (x, q) corresponding to (x, p) which can be constructed, as in the above lemmas, as the limit of a sequence of minimising points for S or $-S$ on M. The following applies to any such point.

Lemma 5.21. *With the notation of the previous two lemmas, the point $(x, q) \in M$ is either such that $\pi|_M$ is non-singular at (x, q) and ind$\gamma_{(x,q)} = 0$, or is such that, given some arbitrarily small $\varepsilon > 0$, β can be extended by an amount ε beyond (x, q) to a point $(\hat{x}, \hat{q}) = \beta(1 + \varepsilon) \in M$ such that ind$\gamma_{(\hat{x},\hat{q})} = 0$. Here $\gamma_{(x,q)}$ (respectively $\gamma_{(\hat{x},\hat{q})}$) denotes the integral curve for H which lies on M and which connects the point (x, q) (respectively $(\hat{x}, \hat{q})$) to U.*

Proof. If (x, p) lies downstream of U, then, by construction, (x, q) is the limit of a sequence of minimising points $(x(m), q(m)) = (x - \frac{1}{m}\mathbf{e}, q(\frac{1}{m}))$ for S at each of which ind$\gamma_{(x(m),q(m))} = 0$, where the index is calculated with respect to traversing γ from U to

$(x(m), q(m))$. If (x, p) lies upstream, then the same argument works with S replaced by $-S$. Denote the path defined by this sequence of points on M by C. The vector $\mathbf{e}$ is tangent to the projection of C onto state space and so, by the previous lemma, the smooth continuation of C on M beyond the point (x, q) coincides with β. By the transversality of β, if C intersects the singular cycle Σ at (x, q), then it must do so in an isolated simple singular point. In this case, we can take the point $(\hat{x}, \hat{q})$ to be $(x(m_0), q(m_0))$ for $m_0 > 1/\varepsilon$.

Conversely, if C does not intersect Σ at (x, q), i.e. $\pi|_M$ is non-singular at (x, q), then C does not meet Σ in a small neighbourhood of (x, q) on M either. This follows again by the transversality of β and its extension over $b(I)$. There is thus a segment C' of the curve C from (x, q) to some $(x(m_0), q(m_0))$ for sufficiently large m_0 such that the ind$C' = 0$. Now, by construction, $\text{ind}(-\gamma_{(x(m_0),q(m_0))}) = 0$, where the minus sign indicates that the curve is traversed in reverse, i.e. from $(x(m_0), q(m_0))$ to U. Also, if C'' denotes the connecting curve in U between the starting points of $\gamma_{(x,q)}$ and $\gamma_{(x(m_0),q(m_0))}$, then ind$C'' = 0$. Now, by hypothesis, the closed curve

$$\gamma_{(x,q)} \cup C' \cup (-\gamma_{(x(m_0),q(m_0))}) \cup C''$$

on M has Maslov index zero. Hence, ind$\gamma_{(x,q)} = 0$ as required. □

We have shown in the previous lemma that, by the construction of (x, q) as the limit of a sequence of minimising points on M, it follows that (x, q) (or a point close to it) lies on a branch of M on which ind$\gamma_{(x,q)} = 0$, and further that this holds for each such (x, q). Now for each (x, q), the corresponding connecting curve β coincides with one or other branch of the curve α appearing in the definition of (x, p). We will see in the proof of the next lemma that one of these branches can be identified with zero and the other with non-zero Malsov index. Our eventual aim is to show that if the folded singularity (x, p) is downstream, then $S(x, q) < S(x, p)$ for at least one (x, q), and conversely $-S(x, q) < -S(x, p)$ upstream. In order to do this we need to know that for this (x, q), the corresponding β coincides with the branch of α identified with non-zero index. This requires the following hypothesis to be satisfied by M.

Hypothesis 5.22. Let (x, p) be any folded singularity on M downstream of U. Then there exists at least one point $(x, q) \in M$ constructed, as in the previous three lemmas, as the limit of a sequence of minimising points for S on M, with the following property. Let β be the transverse connecting curve between (x, p) and (x, q) on M and let b be the corresponding regular curve segment in state space. Then b can be extended at both ends in such a way that, outside of a compact interval on b containing x, b is covered by a corresponding extension of β on M consisting of minimising points for S. Conversely, if (x, p) is upstream then the same condition holds with S replaced by $-S$.

This is the last hypothesis that we require. The consequences of it not holding will be seen in the proof of the next lemma - essentially every extension of b is covered by an infinite sequence of folded singularities. We consider this case to be somewhat pathological and thus have excluded it from the present analysis by imposing the above hypothesis. As with Hypothesis 5.2(4), we leave for the future the question of deciding just how restrictive it is.

The consequence of the above hypothesis being satisfied is that the corresponding connecting curve β satisfies a certain local property which, in rough terms, can be stated as follows. If β leaves (x,p) to the left, when viewed in $\mathbb{R}^2$ phase space under the pull-back b^*, then it must approach (x,q) from the left, and vice versa. A more precise statement would be that the s-component of the tangent to $b^*(\beta)$ has opposite signs at $b^*((x,p))$ and $b^*((x,q))$. However, this is not quite true as the s-component is zero at $b^*((x,p))$. Also, $\pi|_M$ can be singular at (x,q), in which case, as in the previous lemma, we have to consider a point $(\hat{x},\hat{q})$ close to (x,q) on β at which $\mathrm{ind}\gamma_{(\hat{x},\hat{q})} = 0$. We formalise this property in the following definitions and then deal with these technicalities in the following lemma. The name we give to this property reflects the fact that it follows from M being regular over b outside of a compact interval on b containing x.

Definition 5.23. Suppose $(x,p) \in M$ is a folded singularity and let $(x,q) \in M$ be a point constructed, as above, as the limit of a sequence of minimising points for S or $-S$ on M. Let $\beta : [0,1] \to M$ and $b : I \to \mathbb{R}^n$ denote the transverse connecting curve and corresponding regular curve segment in state space. So $\beta(0) = (x,p)$, $\beta(1) = (x,q)$ and, for a given choice of parameterisation, there exists $s \in [0,1]$ such that $b(s) = \pi(\beta(0)) = \pi(\beta(1))$. Then β is said to be 'regular at the boundary' if one or other of the following conditions holds true. On the one hand, suppose that $\pi(\beta(\varepsilon_1)) = b(s-\varepsilon_2)$ for some arbitrarily small ε_1, $\varepsilon_2 > 0$. Then we can find ε_3, ε_4, $\varepsilon_5 > 0$ such that either (i) $\pi(\beta(1-\varepsilon_3)) = b(s-\varepsilon_4)$ if $\pi|_M$ is non-singular at (x,q) or otherwise (ii) $\pi(\hat{x},\hat{q}) = b(s+\varepsilon_5)$ where $(\hat{x},\hat{q})$ is the point refered to in the previous lemma. On the other hand, suppose that $\pi(\beta(\varepsilon_1)) = b(s+\varepsilon_2)$. Then either (i) $\pi(\beta(1-\varepsilon_3)) = b(s+\varepsilon_4)$ if $\pi|_M$ is non-singular at (x,q) or otherwise (ii) $\pi(\hat{x},\hat{q}) = b(s-\varepsilon_5)$.

Definition 5.24. Suppose for every folded singularity $(x,p) \in M$ that the connecting curve to at least one of the points $(x,q) \in M$ is regular at the boundary. Then M is said to be regular at the boundary.

Lemma 5.25. *A Lagrangian manifold satisfying the above hypotheses is regular at the boundary.*

Proof. Let $(x,p) \in M$ be a folded singularity and suppose x lies downstream of U. The same argument works upstream. Let α denote the curve appearing in the definition of the folded singularity. Let $(x,q) \in M$ be any point constructed, as in the above lemmas, as the limit of a sequence of minimising points for S on M. Let $\beta : [0,1] \to M$ denote the transverse connecting curve on M from (x,p) to (x,q) and let $b : I \to \mathbb{R}^n$ denote the corresponding regular curve segment in state space. Note we have $\pi(\{\alpha(t) : t \in [-1,1]\}) \subset \pi(\{\beta(\tau) : \tau \in [0,1]\}) = \{b(s) : s \in [0,1]\}$. Then, by the previous lemma, either $\pi|_M$ is non-singular at (x,q) and $\mathrm{ind}\gamma_{(x,q)} = 0$ or β can be extended to a point $(\hat{x},\hat{q}) = \beta(1+\hat{\varepsilon}) \in M$ at which $\mathrm{ind}\gamma_{(\hat{x},\hat{q})} = 0$, for some arbitrarily small $\hat{\varepsilon} > 0$. Since, in the first case, we can find an open ball on M centred on (x,q) on which $\pi|_M$ is non-singular, we can assume in this case also that β can be extended to a point $(\hat{x},\hat{q}) = \beta(1+\hat{\varepsilon})$ at which $\mathrm{ind}\gamma_{(\hat{x},\hat{q})} = 0$, with $\hat{\varepsilon}$ small enough to ensure β does not intersect Σ between $\beta(1)$ and $\beta(1+\hat{\varepsilon})$.

Consider now one particular (x,q) and the corresponding curves β and b. One branch of α coincides with β and, by transversality of β, α intersects Σ transversely in a simple singular point at (x,p). So the Maslov index of α must change by ± 1 on passage through (x,p). Chose an arbitrarily small $\varepsilon > 0$ and consider $\mathrm{ind}\gamma_{\alpha(t)}$ for $t \in (-\varepsilon,\varepsilon)$, where the index is calculated with respect to traversing γ from U to $\alpha(t)$. Note, consistent with the notation in the previous lemma, $\gamma_{\alpha(t)}$ denotes the integral curve for H which lies on M and connects the point $\alpha(t)$ to U. By Corollary 5.6, the only possibilities for $\mathrm{ind}\gamma_{\alpha(t)}$ are 0 or +1. Let α_t denote the segment of α from $\alpha(-\varepsilon)$ to $\alpha(t)$. Now, $\mathrm{ind}\alpha_t$ only changes at $t = 0$ and it does so by ± 1. So, since the closed curves $\gamma_{\alpha(-\varepsilon)} \cup \alpha_t \cup -\gamma_{\alpha(t)} \cup C_t$ on M all have Maslov index zero for some connecting curves C_t in U, it follows that $\mathrm{ind}\gamma_{\alpha(t)}$ is constant on each branch of α. So the parameterisation t of α can be chosen such that $\mathrm{ind}\gamma_{\alpha(t)} = 0$ for $-\varepsilon \leq t < 0$ and $\mathrm{ind}\gamma_{\alpha(t)} = +1$ for $0 < t \leq \varepsilon$.

Suppose first that β coincides with the branch $\{\alpha(t) : t \in [-\varepsilon, 0)\}$ on which $\mathrm{ind}\gamma_{\alpha(t)} = 0$, i.e. parameterise β by $\tau \in [0,1]$ so that for $\tau \in (0,\varepsilon]$ there exists some $t \in [-\varepsilon, 0)$ such that $\beta(\tau) = \alpha(t)$. Then $\mathrm{ind}\gamma_{\beta(\varepsilon)} = 0$. Let $\tilde{\beta}$ denote the segment of β from $\beta(\varepsilon)$ through $\beta(1) = (x,q)$ and onto $\beta(1+\hat{\varepsilon}) = (\hat{x},\hat{q})$. Now the closed curve $\gamma_{\beta(\varepsilon)} \cup \tilde{\beta} \cup -\gamma_{(\hat{x},\hat{q})} \cup C$ on M has Maslov index zero, for some connecting curve C in U. Since the first, third and fourth terms in this union have index zero, it follows that $\mathrm{ind}\tilde{\beta} = 0$. A repetition of the argument used in the proof of Proposition 5.17 then shows that the s-component of the tangent to $b^*(\beta)$ has the same sign at $b^*(\beta(\varepsilon))$ and $b^*(\beta(1+\hat{\varepsilon}))$. Note, in applying this argument, that if $\tilde{\beta}_\tau$ denotes the segment of β from $\beta(\varepsilon)$ to $\beta(\tau)$, then the requirement that $\mathrm{ind}\tilde{\beta}_\tau = 0$ or +1 follows from Corollary 5.6 and consideration of the closed curve $\gamma_{\beta(\varepsilon)} \cup \tilde{\beta}_\tau \cup -\gamma_{\beta(\tau)} \cup C_\tau$ on M for some connecting curve C_τ in U.

Since $\pi(\beta(0)) = \pi(\beta(1)) = b(s_0)$ for some $s_0 \in [0,1]$, it follows that the s-component of the tangent to $b^*(\beta)$ must change sign an even number of times and at least twice between $\beta(\varepsilon)$ and $\beta(1+\hat{\varepsilon})$. Furthermore, we can find points $\beta(t_1)$ and $\beta(t_2)$ on β for some $t_1, t_2 \in (0,1]$ at which two of these sign changes occur such that, if $\pi(\beta(t_i)) = b(s_i)$ for some $s_i \in [0,1]$, $i = 1,2$, then $s_1 \leq s_0 \leq s_2$. Now each of these points at which the sign changes corresponds to a transverse intersection of β with Σ in a simple singular point. Further, such a singular point is clearly a folded singularity on M. We can also assume that, in fact, $s_1 < s_0 < s_2$. For otherwise, it must be the case that $\pi|_M$ is singular at (x,q) and then the construction, in the proof of the previous lemma, of the point $(\hat{x},\hat{q}) = \beta(1+\hat{\varepsilon})$ requires the s-component of the tangent to $b^*(\beta)$ to have opposite signs at $b^*(\beta(\varepsilon))$ and $b^*(\beta(1+\hat{\varepsilon}))$, contrary to the deduction of the previous paragraph.

Now, let $\pi(\beta(\varepsilon)) = b(s_3)$ for some s_3 and suppose that $s_3 < s_0$, i.e. at the folded singularity (x,p), the projection of M onto state space lies to the left of s_0 on b. A similar argument will hold if $s_3 > s_0$. Then β must pass through a folded singularity $\beta(t_2)$ for some $t_2 \in (0,1]$ whose projection $b(s_2)$ lies to the right of s_0 on b, i.e. $s_0 < s_2$. Also, at $\beta(t_2)$ the projection of M onto state space lies to the left of s_2 on b. Further, the argument used in the proof of Proposition 5.17 shows that β approaches this singularity via points $\beta(\tau)$ at which $\mathrm{ind}\gamma_{\beta(\tau)} = +1$ for $\tau \in (t_2 - \bar{\varepsilon}, t_2)$, for some $\bar{\varepsilon} > 0$. By Corollary 5.6, these cannot therefore be minimising points for S on M. Thus, if β_{t_2} denotes the segment of β up to, but not including, $\beta(t_2)$ then this particular extension of b to the right of s_0 is covered by a curve β_{t_2} on M which does not consist of minimising points for S. There are potential minimising points lying over this extension of b, namely $\beta(\tau)$ for $\tau \in (t_2, t_2+\bar{\varepsilon})$ at which $\mathrm{ind}\gamma_{\beta(\tau)} = 0$, but these points lie beyond the folded singularity $\beta(t_2)$ on β.

This argument can be repeated to show that further extensions of b to the right, formed by projecting extensions of this particular β, must pass through a sequence $\pi(\beta(t_\lambda))$ of projections of folded singularities. Further, on the approach to each $\pi(\beta(t_\lambda))$, b is covered by an extension of β on M which does not consist of minimising points for S until after it has passed through $\beta(t_\lambda)$. Then, since the projections of folded singularities on b are isolated, it follows that for this particular (x,q) we cannot find an extension of b satisfying Hypothesis 5.22. The same conclusion holds for any other (x,q) for which the corresponding β coincides with the branch of α on which $\mathrm{ind}\gamma_{\alpha(t)} = 0$.

In order to satisfy this hypothesis, there must therefore exist at least one (x,q) for which the corresponding β coincides with the branch $\{\alpha(t) : t \in (0,\varepsilon]\}$ on which $\mathrm{ind}\gamma_{\alpha(t)} = +1$. Then $\mathrm{ind}\gamma_{\beta(\varepsilon)} = +1$. Again, let $\tilde{\beta}$ denote the segment of β from $\beta(\varepsilon)$ to $\beta(1+\hat{\varepsilon}) = (\hat{x},\hat{q})$ and consider the closed curve $\gamma_{\beta(\varepsilon)} \cup \tilde{\beta} \cup -\gamma_{(\hat{x},\hat{q})} \cup C$ on M. Since the third and fourth terms in this union have index zero and the first has index $+1$, it follows that $\mathrm{ind}\tilde{\beta} = -1$. The argument of Proposition 5.17 then shows that the s-component of the tangent to $b^*(\beta)$ has opposite signs at $b^*(\beta(\varepsilon))$ and $b^*(\beta(1+\hat{\varepsilon}))$. Thus this particular β is regular at the boundary. Since (x,p) was arbitrary, it follows that M is regular at the boundary. □

Lemma 5.26. *Let $(x,p) \in M$ be a folded singularity and $(x,q) \in M$ be a point which can be connected to (x,p) by a transverse connecting curve β on M which is regular at the boundary. Then we can assume without loss of generality that (x,q) is in fact a non-singular point for $\pi|_M$.*

Proof. Suppose $\pi|_M$ is singular at (x,q) and that x lies downstream of U. The same argument works upstream. Then, with the notation of Definition 5.23 and a particular choice of parameterisation s of b, suppose that $\pi(x,p) = \pi(x,q) = b(\hat{s})$ and

$$\pi(\beta(\varepsilon_1)) = b(\hat{s} - \varepsilon_2) \tag{20}$$

for some $\hat{s} \in [0,1]$. It follows from regularity at the boundary that $\pi(\hat{x},\hat{q}) = b(\hat{s}+\varepsilon_5)$. Now $(\hat{x},\hat{q}) = \beta(1+\varepsilon)$ for some $\varepsilon > 0$, i.e. $(\hat{x},\hat{q})$ lies on the extension of β beyond $(x,q) = \beta(1)$. By transversality, β intersects Σ in a simple singular point at (x,q). The sign of the s-component of the tangent vector to $b^*(\beta)$ in $\mathbb{R}^2$ phase space must therefore change sign at (x,q). It follows that

$$\pi(\beta(1-\varepsilon_3)) = b(\hat{s} + \varepsilon_4). \tag{21}$$

Thus, since b has no self intersections and $\pi\{\beta(t) : t \in [0,1]\} = \{b(s) : s \in [0,1]\}$, it follows from (20) and (21) that there must exist some $\hat{t} \in (0,1)$ such that $\pi(\beta(\hat{t})) = b(\hat{s}) = x$ and also such that there exist arbitrarily small $\varepsilon_i > 0$ satisfying

$$\pi(\beta(\hat{t} - \varepsilon_6)) = b(\hat{s} - \varepsilon_7) \tag{22}$$

and

$$\pi(\beta(\hat{t} + \varepsilon_8)) = b(\hat{s} + \varepsilon_9). \tag{23}$$

Now, by transversality of β, the rank of $\pi|_M$ at $\beta(\hat{t})$ can drop by at most 1 in the direction tangent to b. Thus, from (22) and (23), $\pi|_M$ is non-singular at $\beta(\hat{t})$. We can thus take (x,q) to be the point $\beta(\hat{t})$ without loss of generality. □

We can now prove the main result of this paper.

Theorem 5.27. *Suppose (x,p) is a folded singularity on M. If x is downstream of U then (x,p) cannot be a minimising point for $S(x,\hat{p})$ over all $(x,\hat{p}) \in M$. Conversely, if $x \in \Omega$ is upstream, then (x,p) cannot be a minimising point for $-S$.*

Proof. Suppose x lies downstream of U. The same argument works upstream. Let α denote the curve appearing in the definition of the folded singularity. Then, by the previous lemma, (x,p) can be transversely connected via a curve β on M to a point $(x,q) \in M$ at which $\pi|_M$ is non-singular. Furthermore, β is regular at the boundary and coincides with one branch of α. Our aim is to show that any initial open segment of β has index 0 or -1.

Now, as shown in the proof of Lemma 5.25, ind$\gamma_{\alpha(t)}$ is constant on each branch of α and the parameterisation t of α can be chosen such that ind$\gamma_{\alpha(t)} = 0$ for $-\varepsilon \le t < 0$ and ind$\gamma_{\alpha(t)} = +1$ for $0 < t \le \varepsilon$. Furthermore, since β is regular at the boundary, the same proof shows that it must coincide with the branch $\{\alpha(t) : t \in (0,\varepsilon]\}$.

Let β_τ denote the open segment of β from $\beta(0) = (x,p)$ to $\beta(\tau)$. Then, since the singular point (x,p) can be assumed by transversality to be isolated on β, we can assume that β_τ does not intersect Σ for $\tau \in (0,\varepsilon]$. Thus ind$\beta_\tau = 0$ for $\tau \in (0,\varepsilon]$.

Now for $\tau \in (\varepsilon, 1]$, let $\tilde{\beta}_\tau$ denote the open segment of β from $\beta(\varepsilon)$ to $\beta(\tau)$. Consider the closed curve $\gamma_{\beta(\varepsilon)} \cup \tilde{\beta}_\tau \cup -\gamma_{\beta(\tau)} \cup C_\tau$ on M for some connecting curve C_τ in U. By the choice of the branch of α, there exists some $t \in (0,\varepsilon]$ such that $\beta(\varepsilon) = \alpha(t)$. So the Maslov index of the first term in this union is $+1$. The fourth term has index 0 and, by Corollary 5.6, the third term has index 0 or -1. Thus, since the total closed curve must have index zero, it follows that ind$\tilde{\beta}_\tau = 0$ or -1 for $\tau \in (\varepsilon, 1]$. Hence ind$\beta_\tau = 0$ or -1 for $\tau \in (0,1]$, i.e. any open initial segment of β has index 0 or -1. It then follows from Proposition 5.17 that $S(x,q) < S(x,p)$. So (x,p) is not a minimising point for S. □

Corollary 5.28. *The set of essential caustics C_* only contains projections of non-folded singularities for $\pi|_M$, i.e. given any $\bar{x} \in C_*$ there exists some $\bar{p} \in \mathbb{R}^n$ such that $(\bar{x},\bar{p}) \in M$ is a non-folded singularity for $\pi|_M$ and S or $-S$ achieves its minimum at $(\bar{x},\bar{p})$, depending on whether $\bar{x}$ is upstream or downstream. Then, furthermore, for any x in a small neighbourhood of $\bar{x}$, there is a point $(x,p) \in M$ which lies in a small neighbourhood of $(\bar{x},\bar{p})$ on M, i.e. on the same branch of M as $(\bar{x},\bar{p})$.*

Proof. From the definition of a folded singularity, it follows that if $(\bar{x},\bar{p}) \in M$ is a non-folded singularity, then given any sequence $x_n \to \bar{x}$ in $\mathbb{R}^n$, we can find a corresponding sequence p_n in $\mathbb{R}^n$ such that $(x_n, p_n) \in M$ for all n and $(x_n, p_n) \to (\bar{x},\bar{p})$ as $n \to \infty$. □

Recall from Section 2 that, at any point $(x,p) \in M$, we can chose a collection of indices $I \subseteq \{1,\dots,n\}$ such that in a neighbourhood of (x,p) the generating function for M is of the form $S_I(x^I, p_{\bar{I}})$, where $\bar{I}$ denotes the complement of I. Furthermore, if $\Phi_I = S_I + x^{\bar{I}} p_{\bar{I}}$, then the function S appearing in the definitions (17) and (18) of W is given by $S = \Phi_I$ in the neighbourhood of M on which S_I is defined.

Now, if $\pi|_M$ is non-singular at $(x_1,p_1) \in M$ then $I = \{1,\dots,n\}$ and so S_I is a function of x alone. Then from the definition (3) of the generating function S_I, there exists some $\delta > 0$ such that the set of (x,p) with $x \in B_\delta(x_1)$ and $p = \partial S_I/\partial x$ defines a neighbourhood of (x_1,p_1) on M and $S = S_I$ in this neighbourhood.

Let x_0 be an interior point of $\Omega \backslash C_*$ and suppose x_0 is downstream of U. The same argument works upstream with S replaced by $-S$. Then from the Hypothesis 2.1(4) of local boundedness, there is a $K_{x_0} < \infty$ and a $\delta_{x_0} > 0$ such that $|p| \le K_{x_0}$ for all $(x,p) \in M$ with $x \in B_{\delta_{x_0}}(x_0)$. Then it is shown in the proof of Theorem 2 of [11] that this same K_{x_0} is the Lipschitz constant which applies in the local Lipschitz property for W at x_0. The proof of this fact starts by taking δ_{x_0} small enough to ensure $B_{\delta_{x_0}}(x_0) \subseteq \Omega \backslash C_*$. Then any $x_1 \in B_{\delta_{x_0}}(x_0)$ is not in C_*. So $\pi|_M$ is non-singular at some minimising point $(x_1, p_1) \in M$ for $S(x_1, .)$ and so, if S_I denotes the generating function for M at this point, then $W(x_1) = S(x_1) = S_I(x_1)$. Then, for any $x_2 \in B_\delta(x_1) \subseteq B_{\delta_{x_0}}(x_0)$, $|\partial S_I(x_2)/\partial x| < K_{x_0}$ and so

$$\begin{aligned} W(x_2) &\leq S_I(x_2) \\ &\leq S_I(x_1) + K_{x_0} |x_2 - x_1| \\ &= W(x_1) + K_{x_0} |x_2 - x_1| . \end{aligned} \tag{24}$$

An argument using the convexity of $B_{\delta_{x_0}}(x_0)$ then shows that the above holds for all x_1, $x_2 \in B_{\delta_{x_0}}(x_0)$. Interchanging x_1 and x_2 leads to

$$|W(x_1) - W(x_2)| \leq K_{x_0} |x_1 - x_2|$$

for all x_1, $x_2 \in B_{\delta_{x_0}}(x_0)$, which is the local Lipschitz property.

Now let $x_1 \in C_*$ and again suppose x_1 is downstream of U. Then there exists some non-folded singularity $(x_1, p_1) \in M$ at which $S(x_1, .)$ achieves its minimum. For some strict subset $I \subset \{1, \ldots, n\}$, let S_I denote the generating function for M in a neighbourhood of (x_1, p_1) and let $\Phi_I = S_I + x^{\bar{I}} p_{\bar{I}}$ denote the restriction of S to this neighbourhood. Then, from the previous corollary, there is a $\delta = \delta(x_1) > 0$ such that for all $x \in B_\delta(x_1)$, there exists p such that (x, p) lies on the branch of M defined by S_I, i.e. such that

$$x^{\bar{I}} = -\frac{\partial S_I}{\partial p_{\bar{I}}} \qquad p_I = \frac{\partial S_I}{\partial x^I}$$

holds true at (x, p). Now, consider the function $W_{x_1}(x)$ defined in $B_\delta(x_1)$ by replacing S by Φ_I in (17). Then clearly $W(x_1) = S(x_1, p_1) = W_{x_1}(x_1)$ and $W(x) \leq W_{x_1}(x)$ for $x \in B_\delta(x_1)$. Similarly, if x_1 is upstream then W_{x_1} is defined by replacing S by Φ_I in (18). Then we make the following conjecture regarding the local functions W_{x_1}.

Conjecture 5.29. For each $x_1 \in C_*$, the function W_{x_1} is continuous in $B_\delta(x_1)$. Furthermore, for a given Lagrangian manifold M, there is a single integer $m \geq 1$ with the following property: for any $x_1 \in C_*$ and for any $K \in (1, \infty)$ such that $|p| \leq K$ for all $(x, p) \in M$ with $x \in B_\delta(x_1)$, the inequality

$$W_{x_1}(x_2) \leq W_{x_1}(x_1) + K^m |x_2 - x_1|$$

holds true for all $x_2 \in B_\delta(x_1)$.

We show first that it follows from this conjecture that W is, in fact, locally Lipschitz continuous in the whole of Ω. We then finish the section by showing that such a value m does exist for Lagrangian manifolds of dimension ≤ 5.

Theorem 5.30. *W is locally Lipschitz continuous at every point in Ω.*

Proof. This is a generalisation of Theorem 2 of [11] and its proof. Note, this theorem is stated above as Theorem 5.1. We show first that W is continuous in Ω. Let $x_1 \in \Omega$ and suppose x_1 is downstream of U. If $x_1 \notin C_*$ then the proof of Theorem 2 of [11] shows that W is continuous at x_1. Conversely, if $x_1 \in C_*$ then, as above, there exists a continuous function W_{x_1} of x defined in the open ball $B_\delta(x_1)$ such that $W(x_1) = S(x_1, p_1) = W_{x_1}(x_1)$ and $W(x) \le W_{x_1}(x)$ for $x \in B_\delta(x_1)$, where $(x_1, p_1) \in M$ is some non-folded singularity at which $S(x_1, .)$ achieves its minimum. Hence

$$\limsup_{x \to x_1} W(x) \le W_{x_1}(x_1) = W(x_1)$$

showing that W is upper-semicontinuous at x_1. Since it shown in Theorem 1 of [11] that W is lower semicontinuous in general, it follows that W is continuous at x_1. If x_1 is upstream of U, then the above argument can be repeated with S replaced by $-S$. Note, the same comment applies to the proof of lower semicontinuity in Theorem 1 of [11], viz. replacing S by $-S$ in the proof of that theorem shows that W is lower semicontinuous at any x upstream of U.

We now prove the local Lipschitz property. Suppose $x_0 \in \Omega$ and, as above, chose some $K_{x_0} < \infty$ and $\delta_{x_0} > 0$ such that $|p| \le K_{x_0}$ for all $(x, p) \in M$ with $x \in B_{\delta_{x_0}}(x_0)$. Let $B = B_{\delta_{x_0}}(x_0)$. Note that B is convex and consider any $x_1 \in B$. If $x_1 \notin C_*$ then as above in (24), there exists $B_\delta(x_1) \subseteq B$ such that for any $x_2 \in B_\delta(x_1)$,

$$W(x_2) \le W(x_1) + K_{x_0} |x_2 - x_1| .$$

Conversely, if $x_1 \in C_*$ then from the conjecture we can find a ball $B_\delta(x_1) \subseteq B$ and a continuous function W_{x_1} defined on $B_\delta(x_1)$ such that for all $x_2 \in B_\delta(x_1)$,

$$\begin{aligned} W(x_2) &\le W_{x_1}(x_2) \\ &\le W_{x_1}(x_1) + (K_{x_0})^m |x_2 - x_1| \\ &= W(x_1) + (K_{x_0})^m |x_2 - x_1| . \end{aligned}$$

Since $m \ge 1$ and K_{x_0} can be chosen, if necessary, to be ≥ 1, we thus have that for any $x_1 \in B$ there exists $B_\delta(x_1) \subseteq B$ such that

$$W(x_2) \le W(x_1) + (K_{x_0})^m |x_2 - x_1| \tag{25}$$

for any $x_2 \in B_\delta(x_1)$. The argument used in the proof of Theorem 2 of [11] can now be applied to show that (25) holds for x_2 along any ray from x_1 up to ∂B. Since B is convex, this means (25) holds for all $x_1, x_2 \in B$. Interchanging x_1 and x_2 allows us to conclude that

$$|W(x_1) - W(x_2)| \le (K_{x_0})^m |x_1 - x_2|$$

for all $x_1, x_2 \in B$. Thus W is locally Lipschitz at x_0. □

We have seen above in (24) that at points in the interior of $\Omega \backslash C_*$, the local Lipschitz constant for W is determined by the local bound on M. It is now clear from the preceding proof that at points in the closure of C_*, the value m in the conjecture defines the degree to which the local Lipschitz constant can grow in relation to the local bound on M.

Corollary 5.31. *W is a viscosity solution of $H(x, \partial W(x)/\partial x) = 0$ for all $x \in \Omega$ downstream of U and a viscosity solution of $H(x, -\partial W(x)/\partial x) = 0$ for all $x \in \Omega$ upstream of U.*

Proof. Follows from the previous result and Theorem 2.2. □

We now show that the above conjecture is true for Lagrangian manifolds of dimension ≤ 5. For manifolds of these dimensions in general position, the caustics or, more precisely, the singularities of the projection onto state space have been classified by Arnold - see Section 11 of [2] or Appendix 12 of [3]. This classification takes the form of a generating function $S_I(x^I, p_{\bar{I}})$ for M in the neighbourhood of the corresponding singularity. Each singularity is labelled by a corresponding simple Lie group. For a Lagrangian manifold M in general position of dimension $n \leq 5$, all singularities in the projection of M onto state space can be reduced by a Lagrangian equivalence to one of the following normal forms. For $n = 1$ there is a single type of singularity labelled A_2. For $n = 2$ there is, in addition, a type labelled A_3. For $n = 3$ there are, in addition, types labelled A_4 and D_4. For $n = 4$ there are, in addition, types labelled A_5 and D_5. For $n = 5$ there are, in addition, types labelled A_6, D_6 and E_6.

For all the singularities of type A_k, the generating function is of the form $S_{n-1}(x_2, ..., x_n, p_1)$ and the manifold is given by

$$x_1 = -\frac{\partial S_{n-1}}{\partial p_1} \qquad p_i = \frac{\partial S_{n-1}}{\partial x_i} \tag{26}$$

for $i = 2, ..., n$. Note, we are now using the notation x_i and p_i to denote the coordinate values of a point (x, p) in $\mathbb{R}^n$, rather than to denote distinct points in $\mathbb{R}^n$. The function S appearing in the definitions (17) and (18) of W is then given by $S = S_{n-1} + x_1 p_1$ in the neighbourhood of the singularity.

For the singularities of types D_k and E_k, the generating function is of the form $S_{n-2}(x_3, ..., x_n, p_1, p_2)$ and the manifold is given by

$$x_j = -\frac{\partial S_{n-2}}{\partial p_j} \qquad p_i = \frac{\partial S_{n-2}}{\partial x_i} \tag{27}$$

for $j = 1, 2$ and $i = 3, ..., n$. The function S is then given by $S = S_{n-2} + x_1 p_1 + x_2 p_2$ in the neighbourhood of the singularity.

Note that the sign convention used here is the opposite of that used by Arnold. Note also that the generating function need not depend on every x_i argument. For instance, a singularity of type A_3 has a generating function which depends only on x_2 and p_1. If $n > 2$, then in a neighbourhood of a singularity of type A_3, $p_i = 0$ on M for $i > 2$.

It can easily be verified that, of the above list, singularities of types A_2, A_4, A_6, D_4^+, D_6^+ and E_6 are all folded. Note D_4 and D_6 are given respectively by generating functions

$$S_{n-2} = \pm p_1^2 p_2 \pm p_2^3 + x_3 p_2^2$$

and

$$S_{n-2} = \pm p_1^2 p_2 \pm p_2^5 + x_5 p_2^4 + x_4 p_2^3 + x_3 p_2^2.$$

D_4^+ and D_6^+ denote the versions of these two singularities in which the first two terms in the respective generating function have the same sign.

The remaining singularities A_3, A_5, D_4^-, D_5 and D_6^- are all non-folded. Note, in this list, D_4^- and D_6^- denote the versions of the above two generating functions in which the first two terms have opposite signs. For $n \leq 5$, the set of essential caustics C_* can thus only contain projections of these types of singularities. In each case, the function S, defined as above by $S_{n-1} + x_1 p_1$ or $S_{n-2} + x_1 p_1 + x_2 p_2$ in a neighbourhood of the singularity, is a polynomial in x and p. The largest total degree of these polynomials is six.

Let the singularity be denoted $(\hat{x}, \hat{p})$ and let S_I denote the corresponding normal form generating function S_{n-1} or S_{n-2}. Then, since $(\hat{x}, \hat{p})$ is non-folded, there is a $\delta > 0$ such that for all $x \in B_\delta(\hat{x})$, there exists at least one p such that (x, p) lies on one of the branches of M defined in a neighbourhood of $(\hat{x}, \hat{p})$ by S_I. The values of $S(x, p)$ over these branches give rise to a multi-valued function of x in $B_\delta(\hat{x})$. If $\hat{x}$ is downstream of U, then the minimum of these values at any given $x \in B_\delta(\hat{x})$ defines the function $W_{\hat{x}}(x)$ described in the above conjecture. If $\hat{x}$ is upstream, then the minimum of the values of $-S(x, p)$ over these branches defines $W_{\hat{x}}(x)$ at any $x \in B_\delta(\hat{x})$.

The defining relations (26) or (27) for M allow us to express the polynomial S of x and p as a multi-valued algebraic expression in x alone, i.e. depending on powers and roots of x. On any single branch of M in a neighbourhood of $(\hat{x}, \hat{p})$, this gives rise to a well defined algebraic function of x. Since the generating functions for each of the above listed types of non-folded singularities give rise to only finitely many branches of M in a neighbourhood of $(\hat{x}, \hat{p})$, it follows that $W_{\hat{x}}$ is the minimum of a finite number of algebraic functions of x. $W_{\hat{x}}$ is thus clearly continuous in $B_\delta(\hat{x})$.

For any given type of non-folded singularity, the value m required by the above conjecture is obtained by using the relations (26) or (27) to express S as a polynomial in x_i and p_j of maximum degree 1 in any of the x_i terms. The corresponding value of m is then the highest degree of the p_j terms. We can in fact obtain a different m for each dimension n of M from one up to five as follows.

For $n = 1$ there are no non-folded singularities and so $m = 1$.

For $n = 2$, there is only one type of non-folded singularity, viz. A_3. This has generating function

$$S_{n-1} = \pm p_1^4 + x_2 p_1^2$$

and from (26) it follows that, in a neighbourhood of such a singularity,

$$x_1 = \mp 4p_1^3 - 2x_2 p_1.$$

Thus

$$S_{n-1} = -\frac{1}{4} x_1 p_1 + \frac{1}{2} x_2 p_1^2$$

and

$$S = \frac{3}{4} x_1 p_1 + \frac{1}{2} x_2 p_1^2.$$

Thus, for manifolds of dimension 2, we can take $m = 2$ in the above conjecture.

For $n = 3$, there is only one type of non-folded singularity, viz. D_4^-. This has generating function

$$S_{n-2} = \pm p_1^2 p_2 \mp p_2^3 + x_3 p_2^2.$$

Using (27) we get that

$$S = \frac{2}{3} x_1 p_1 + \frac{2}{3} x_2 p_2 + \frac{1}{3} x_3 p_2^2$$

and so, for manifolds of dimension 3, we can take $m = 2$ again.

For $n = 4$, there are two types of non-folded singularity, viz. A_5 and D_5. Repeating the above analysis yields a value of $m = 4$ for A_5 type singularities and $m = 3$ for D_5 type singularities. Thus, for manifolds of dimension 4, we can take $m = 4$.

Lastly, for $n = 5$, there is only one type of non-folded singularity, viz. D_6^-, for which a repeat of the above analysis yields a value of $m = 4$. Thus, for manifolds of dimension 5, we can again take $m = 4$. The results of the last five paragraphs can be summarised as follows.

Theorem 5.32. *Given any $n \leq 5$, consider the corresponding value of m from the above analysis, i.e. $m = 1$ for $n = 1$, $m = 2$ for $n = 2$ or 3 and $m = 4$ for $n = 4$ or 5. Let M be a Lagrangian manifold of dimension n. Then at any $x_0 \in \Omega$, if K is a local bound on M at x_0, then K^m is a local Lipschitz constant for W at x_0, where W is the function defined in (17) and (18). Recall that a local bound on M at x_0 is a value $K \in (1, \infty)$ such that $|p| \leq K$ for all $(x, p) \in M$ with $x \in B_\delta(x_0)$ for some δ.*

6. Examples and further work

We defined, after Hypotheses 5.2 in the previous section, two mains classes of examples of Lagrangian manifolds. The first arose from the evolution of an initial Lagrangian manifold under the phase flow corresponding to a finite time variational problem. We saw that all the requirements of Hypotheses 2.1 and 5.2 are natural in this case. In particular, the requirements that the form pdx be globally exact on M and that the Maslov class be zero on M follow from the fact that M is simply connected. The results of the previous section then clearly explain in geometrical terms the often observed fact that a viscosity solution to the associated HJB equation is obtained by taking the minimum of the generating function S over all branches of M. However, as already noted, this fact can be explained in the finite time case by well-known variational arguments.

The second class of examples is more interesting. These arise as stable or unstable Lagrangian manifolds corresponding to infinite time variational problems. The description of this class of examples in the previous section showed how to construct a simply connected submanifold of a stable or unstable manifold. Thus the requirements that the form pdx be globally exact on M and that the Maslov class be zero on M are satisfied, while the other requirements of Hypotheses 2.1 and 5.2 are reasonable to ask for.

We give two simple examples in this second class. These illustrate, firstly, the types of conditions required to guarantee that a stable or unstable Lagrangian manifold exists, secondly how the orientation introduced in the previous section allows a consistent construction of a viscosity solution both upstream and downstream and, thirdly, the types of extra conditions required to deduce that this solution is in fact the value function for the infinite time variational problem. These two examples also motivate a brief discussion of how to deal with non-convex Hamiltonians.

The first example of L_2-gain problems in non-linear systems theory has been dealt with in detail by Day in Section 5 of [11]. This involves solving $-H(x, \partial\varphi/\partial x) = 0$ where H is the Hamiltonian

$$H(x, p) = \max_{w \in \mathbb{R}^m} \left\{ p^T \left(f(x) + g(x)w \right) + \frac{1}{2} \left| h(x) \right|^2 - \frac{1}{2} \gamma^2 \left| w \right|^2 \right\}. \tag{28}$$

Appropriate local controllability and observability conditions guarantee a hyperbolic equilibrium for the Hamiltonian dynamics and thus the existence of a stable Lagrangian manifold M^+. Day then considers a simply connected submanifold M of M^+ lying over a region Ω in state space which is forward invariant for the phase flow corresponding to H, as described after Hypotheses 5.2 in the previous section. If S is the function defined on M by $dS = pdx$ then by Corollary 5.31 above, $V(x) = \sup\{S(x,p) : p$ such that $(x,p) \in M\}$ is a viscosity solution of $-H(x, \partial\varphi/\partial x) = 0$. Note that the Hamiltonian here is concave and all of M is upstream of U, so formula (18) applies with inf replaced by sup and S replaced by $-S$ and with upstream and downstream swapped round in Corollary 5.31. Note also that in [11], the local Lipschitz continuity required to deduce that V is a viscosity solution was assumed as a hypothesis. Here, we obtain it from the results of the previous section using the fact that M is simply connected and under the assumption that it satisfies the rest of Hypotheses 2.1, 5.2 and 5.22 and that $\dim M \leq 5$.

The value function for this problem is called the available storage ϕ_a and is characterised as the minimal non-negative viscosity supersolution of (28). Day addresses the question of whether $V = \phi_a$ by restricting the set of disturbance inputs over which the cost functional is evaluated to those for which corresponding controlled trajectories remain in Ω if they start in Ω. With this restriction he shows, in Theorem 4 of [11], that $V = \phi_a$ if and only if $V(x) \geq 0$ for all $x \in \Omega$.

The second example can be thought of as either the other half of an H_∞ control problem with null disturbance or the limit of an H_∞ problem as the attenuation bound tends to infinity. It is the infinite time optimal regulator problem

$$\hat{V}(\xi) = \inf_{u(.)\in U} \sup_{T>0} \int_0^T \frac{1}{2} \left(|h(x(t))|^2 + u(t)^T r(x(t)) u(t) \right) dt \tag{29}$$

subject to $\dot{x} = f(x) + g(x)u$, $x(0) = \xi$, $\lim_{t\to\infty} x(t) = 0$. The value function $\hat{V}$ satisfies the HJB equation $H(x, -\partial\hat{V}/\partial x) = 0$, where H is given by

$$H(x,y) = \max_{u\in\mathbb{R}^m} \left\{ y^T (f(x) + g(x)u) - \frac{1}{2} |h(x)|^2 - \frac{1}{2} u^T r(x) u \right\}. \tag{30}$$

Suppose there is an equilibrium at $x = 0$ and that the linearisation of (29) at $x = 0$ is stabilizable and detectable. Then the equilibrium is hyperbolic and so there exists a stable Lagrangian manifold M^+. As above, take a simply connected submanifold M of M^+ lying over a region Ω in state space which is forward invariant for the phase flow corresponding to H. Note that H in (30) is convex and all of M is upstream of U. Thus if we assume that M satisfies Hypotheses 2.1, 5.2 and 5.22 and that $\dim M \leq 5$, then by Corollary 5.31, the function $V(x) = \inf\{-S(x,y) : y$ such that $(x,y) \in M\}$ given by (18) is a viscosity solution of $H(x, -\partial V/\partial x) = 0$, where $S(x,y)$ is the function defined on M by $dS = ydx$. Note, as above, that most of these hypotheses follow from the fact that M is simply connected by construction.

To prove that $V = \hat{V}$ requires a proof that the set U of admissable controls is non-empty. An appealing candidate is the 'feedback' control

$$\hat{u}(x) = r^{-1}(x) g^T(x) \hat{y}$$

where $\hat{y}$ is such that $(x, \hat{y}) \in M$ is a minimising point for $-S(x, .)$ on M. A proof that this control is asymptotically stable has to deal with the fact that it is potentially multi-valued, since there may exist more than one minimising point over a given x. We defer discussion of this to a future paper for reasons of space. However, given this then similar arguments to those used by Day in the L_2-gain case show that, if $V(x) > 0$ for $0 \neq x \in \Omega$, then $V = \hat{V}$ and, furthermore, the above feedback gives the optimal control.

To finish we make two general comments about the results of the previous section and directions for further work. First, the restriction to Hamiltonians which are either convex or concave excludes many questions of current interest in the field of non-linear H_∞ control and differential games. Further, when studying the evolution of viscosity or idempotent solutions to HJB equations, the combined effect of Hypotheses 5.2 is, in a sense, to restrict attention to neighbouring regions of state space which are separated by no more than one set of caustics. A useful way of studying a non-convex Hamiltonian may be to restrict attention to charts on the associated Lagrangian manifold on which H is either convex or concave and then use an analysis similar to that of the previous section of this paper to study the evolution of the solution from one chart to another. What we have in mind is the logarithmic limit of a canonical tunnelling operator as described at the end of Section 4, with the transition between charts achieved by the log limit of a $1/h$-Laplace transform. The switch from convex to concave is related to the orientation given in the previous section and to the correct choice of side of the Laplace transform. A non-convex H can be viewed as a linear PDE with non-constant coefficients in a (min,+) or (max,+) sense and the restriction to charts on which H is either convex or concave suggests a form of localisation over either (min,+) or (max,+). The problem is then to glue together these local linear germs of solutions. This is analogous to the solution of conventional linear PDEs with non-constant coefficients by localisation in a ring of analytic functions. The second Maslov quantisation condition enters into the global analysis of the Maslov canonical operator but not so far, in the literature, into the global analysis of canonical tunnelling operators, perhaps because the focus in this case has been on convex Hamiltonians. The results of the previous section indicate that the second quantisation condition is the one required to ensure that these local (max,+) and (min,+) linear germs can be glued together in the log limit in a way which is consistent with the choice of side of the Laplace transform and the orientation of M and such that the resulting generalised solution is locally Lipschitz. To illustrate how this would work, consider the following simple type of nonconvex Hamiltonian which arises in nonlinear H_∞ control

$$\begin{aligned} H(x,p) &= \max_u \min_w \{p^T(f(x) + g(x)u + h(x)w) \\ &\qquad\qquad -\frac{1}{2}|l(x)|^2 - \frac{1}{2}u^T r(x) u + \frac{1}{2}\gamma^2 |w|^2\} \\ &= \frac{1}{2}p^T g(x) r(x)^{-1} g(x)^T p - \frac{1}{2\gamma^2} p^T h(x) h(x)^T p + p^T f(x) - \frac{1}{2}|l(x)|^2 . \end{aligned}$$

This case is simple because the max and min separate. We can divide the region U of Hypotheses 5.2 into two sub-regions $U^\vee$ and $U^\wedge$ where $x_0 \in \pi(U^\vee)$ if $H(x_0, p)$ is convex in p and $x_0 \in \pi(U^\wedge)$ if $H(x_0, p)$ is concave in p. Consider any $x \in \Omega$ and any $(x, p) \in M$ lying over x and let $\gamma_{(x,p)}$ denote the phase curve corresponding to H which connects (x, p) to U. If we assume that $U^\vee$ and $U^\wedge$ are phase disjoint with respect to H in the sense that phase curves which start in $U^\vee$ do not pass through $U^\wedge$ and vice versa, then the various branches of M lying over x can be divided into two groups $M_x^{\vee}$ and $M_x^{\wedge}$ as follows. If $\gamma_{(x,p)}$ comes from $U^{\vee}$ then $(x,p) \in M_x^{\vee}$ while if $\gamma_{(x,p)}$ comes from $U^{\wedge}$ then $(x,p) \in M_x^{\wedge}$. Then we can define

$$V(x) = \inf_{(p:(x,p)\in M_x^{\vee})} \sup_{(p:(x,p)\in M_x^{\wedge})} S(x,p).$$

An analysis similar to that of the previous section shows that V is a locally Lipschitz viscosity solution of $H(x, \partial V/\partial x) = 0$. The case where H does not separate is clearly more involved.

The second comment is to note that, apart from the proof of Conjecture 5.29 for dim$M \leq 5$, none of the results of the previous section relied on the dimension of the Lagrangian manifold M. This conjecture seems reasonable for higher dimensions. The most important item of future work is probably to find a proof of this.

References

[1] V. I. Arnold: Characteristic class entering in quantization conditions, Functional Analysis and its Applications 1 (1967) 1–13.

[2] V. I. Arnold: Normal forms for functions near degenerate critical points, the Weyl groups of A_k, D_k, E_k and Lagrangian singularities, Functional Analysis and its Applications 6(4) (1972) 3–25.

[3] V. I. Arnold: Mathematical Methods of Classical Mechanics, 2nd ed., Springer-Verlag, New York (1989).

[4] P. Brunovsky: On the optimal stabilization of nonlinear systems, Czech. Math. J. 18 (1968) 278–293.

[5] C. Byrnes, H. Frankowska: Unicité des solutions optimales et absence de chocs pour les equations d'Hamilton-Jacobi-Bellman et de Riccati, Comptes Rendues Acad. Sci. Paris 315, Sér. I (1992) 427–431.

[6] C. Byrnes: On the Riccati partial differential equation for nonlinear Bolza and Lagrange problems, Journal of Mathematical Systems, Estimation and Control 8 (1998) 1–54.

[7] P. Cannarsa, H. Frankowska: Some characterizations of optimal trajectories in control theory, SIAM J. Control Optimisation 29(6) (1991) 1322–1347.

[8] N. Caroff, H. Frankowska: Conjugate points and shocks in nonlinear optimal control, Transactions of the American Mathematical Society 348(8) (1996) 3133–3153.

[9] M. G. Crandall, P. L. Lions: Viscosity solutions of Hamilton-Jacobi equations, Transactions of the American Mathematical Society 277 (1983) 1–42.

[10] M. G. Crandall, L. C. Evans, P. L. Lions: Some properties of viscosity solutions of Hamilton-Jacobi equations, Transactions of the American Mathematical Society 282 (1984) 487–502.

[11] M. V. Day: On Lagrange manifolds and viscosity solutions, Journal of Mathematical Systems, Estimation and Control 8 (1998) (http://www.math.vt.edu/people/day/research /LMVS.pdf).

[12] S. Yu. Dobrokhotov, V. N. Kolokoltsov, V. P. Maslov: Quantization of the Bellman equation, exponential asymptotics and tunneling, in: Advances in Soviet Mathematics 13, V. P. Maslov, S. N. Samborskii (eds.), American Mathematical Society, Providence, Rhode Island (1992) 1–46.

[13] L. C. Evans: Some min-max methods for the Hamlton-Jacobi equation, Indiana Uni. Math. J. 33 (1984) 31–50.

[14] W. H. Fleming, R. W. Rishel: Deterministic and Stochastic Optimal Control, Springer-Verlag, New York (1975).

[15] W. H Fleming, H. M. Soner: Controlled Markov Processes and Viscosity Solutions, Springer-Verlag, New York (1993).

[16] H. Frankowska: Hamilton-Jacobi equations: viscosity solutions and generalised gradients, Journal of Mathematical Analysis Applications 141 (1989) 21–26.

[17] H. Frankowska: Lower semicontinuous solutions of Hamilton-Jacobi-Bellman equations, SIAM J. Control Optimisation 31 (1993) 257–272.

[18] M. Gondran: Analyse MINPLUS, Comptes Rendues Acad. Sci. Paris 323, Sér. I (1996) 371–375.

[19] J. Gunawardena (ed.): Idempotency, Publications of the Newton Institute, Cambridge University Press, Cambridge, UK (1998).

[20] V. N. Kolokoltsov, V. P. Maslov: Idempotent analysis as a tool of control theory and optimal synthesis I, Functional Analysis Applications 23 (1989) 1–11.

[21] V. N. Kolokoltsov, V. P. Maslov: Idempotent Analysis and its Applications, Kluwer, Dordrecht, Holland (1998).

[22] P. L. Lions: Generalised Solutions of Hamilton-Jacobi Equations, Research Notes in Mathematics 69, Pitman, London (1982).

[23] D. L. Lukes: Optimal regulation of nonlinear dynamical systems, SIAM Journal on Control 7 (1969) 75–100.

[24] E. Hopf: Generalized solutions of non-linear equations of first order, J. Mathematics Mechanics 14 (1965) 951–973.

[25] V. P. Maslov, M. V. Fedoriuk: Semi-Classical Approximation in Quantum Mechanics, D. Reidel, Dordrecht, Holland (1981).

[26] V. P. Maslov: Global exponential asymptotics of solutions of tunnel-type equations, Soviet Math. Dokl. 24 (1981) 655–659.

[27] V. P. Maslov: On a new principle of superposition for optimisation problems, Russian Math. Surveys 42(3) (1987) 43–54.

[28] A. S. Mishchenko, V. E. Shatalov, B. Yu. Sternin: Lagrangian Manifolds and the Maslov Operator, Springer-Verlag, Berlin (1990).

[29] D. L. Russell: Mathematics of Finite Dimensional Control Systems, Dekker, New York (1979).

[30] P. Soravia: H_∞ control of nonlinear systems: differential games and viscosity solutions, SIAM Journal on Control and Optimisation 34 (1996) 1071–1097.

[31] I. Vaisman: Symplectic Geometry and Secondary Characteristic Classes, Progress in Mathematics 72, Birkhäuser, Boston (1987).

[32] A. J. van der Schaft: On a state space approach to nonlinear H_∞ control, Systems and Control Letters 16 (1991) 1–8.

[33] A. J. van der Schaft: L_2 gain analysis of nonlinear systems and nonlinear state feedback H_∞ control, IEEE Transactions on Automatic Control AC-37(6) (1992) 770–784.