V. MULTIVARIATE FUNCTIONS: APPLICATIONS

In this final topic, we shall consider functions of more than one variable. We shall introduce the idea of the *gradient vector*, which encodes the rate of change of the function along any direction, and see how to locate and classify the local extrema of multivariate functions.

We shall also look at systems of coupled, first-order differential equations where we have multiple dependent variables coupled together.

Finally, we shall briefly introduce the idea of *partial differential equations*, which are differential equations that describe the dynamics of multivariate functions.

1 Directional derivative

Consider a function $f(x, y)$, and an infinitesimal (vector) displacement $d\mathbf{s} = (dx, dy)$. The change in $f(x, y)$ due to the displacement is given straightforwardly by the multivariate chain rule:

$$\begin{aligned} df &= \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy \\ &= (dx, dy) \cdot \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) \\ &= d\mathbf{s} \cdot \boldsymbol{\nabla} f \, . \end{aligned}$$

Here, the vector $\boldsymbol{\nabla} f$ is the *gradient* of f, also called grad$f$, with Cartestian components

$$\boldsymbol{\nabla} f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) \, . \tag{1}$$

If we write the displacement as $d\mathbf{s} = ds\,\hat{\mathbf{s}}$, with $|\hat{\mathbf{s}}| = 1$, so that $\hat{\mathbf{s}}$ is the direction and ds is the distance moved, then

$$df = ds\,(\hat{\mathbf{s}} \cdot \boldsymbol{\nabla} f) \, .$$

This motivates introducing the *directional derivative* as follows.

Definition (Directional derivative). The *directional derivative* of f in the direction of $\hat{\mathbf{s}}$ is

$$\frac{df}{ds} = \hat{\mathbf{s}} \cdot \boldsymbol{\nabla} f \,.$$

It is the rate of change of f with distance along the direction $\hat{\mathbf{s}}$.

The directional derivative can be used to give an alternative, geometric definition of the gradient vector $\boldsymbol{\nabla} f$.

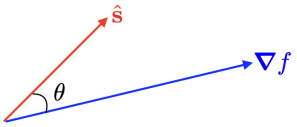
Definition (Gradient vector). The *gradient vector* $\boldsymbol{\nabla} f$ of the function f is defined as the vector that satisfies

$$\frac{df}{ds} = \hat{\mathbf{s}} \cdot \boldsymbol{\nabla} f$$

for all unit vectors $\hat{\mathbf{s}}$.

In Cartesian coordinates, where $d\mathbf{s} = (dx, dy)$, the components of the gradient vector reduce to Eq. (1).

If θ is the angle between $\hat{\mathbf{s}}$ and $\boldsymbol{\nabla} f$ (see figure to the right), we have



$$\frac{df}{ds} = \cos\theta \, |\boldsymbol{\nabla} f| \,.$$

We note from this result the following properties of the gradient vector.

1. The *direction* of $\boldsymbol{\nabla} f$ is the direction in which f *increases* most rapidly.
2. The *magnitude* of $\boldsymbol{\nabla} f$ is the maximum rate of change of f:
$$|\boldsymbol{\nabla} f| = \max_{\forall \theta} \left(\frac{df}{ds} \right) \,.$$
3. If $\hat{\mathbf{s}}$ is parallel to contours of f, then
$$0 = \frac{df}{ds} = \hat{\mathbf{s}} \cdot \boldsymbol{\nabla} f \,.$$
Hence, $\boldsymbol{\nabla} f$ is perpendicular to contours of $f(x, y)$.

2 Stationary points

There is always at least one direction in which $df/ds = 0$, i.e., tangent to the local contour of f.

Stationary points have $df/ds = 0$ for *all* directions. Since

$$\frac{df}{ds} = \hat{\mathbf{s}} \cdot \boldsymbol{\nabla} f \,,$$

we must have

$$\boldsymbol{\nabla} f = 0 \quad \text{at stationary points.}$$

Stationary points may be local maxima, local minima or saddle points.

Near *local maxima*, the contours of f are locally elliptical (see the top row of Fig. 1). The gradient vector points towards a local maximum.

Near *local minima*, the contours of f are also locally elliptical (see the middle row of Fig. 1). The gradient vector points away from a local minimum.

Saddle points are stationary points that are neither local maxima nor minima. Near saddle points, the contours of f are locally hyperbolic (see the bottom row of Fig. 1). Also, the contour lines of f cross at and only at saddle points.

3 Classification of stationary points

To determine whether a stationary point is a maximum, minimum or saddle point, we consider the behaviour of the function in the vicinity of the point. To do so, it is useful first to consider how to generalise Taylor expansions to multivariate functions.

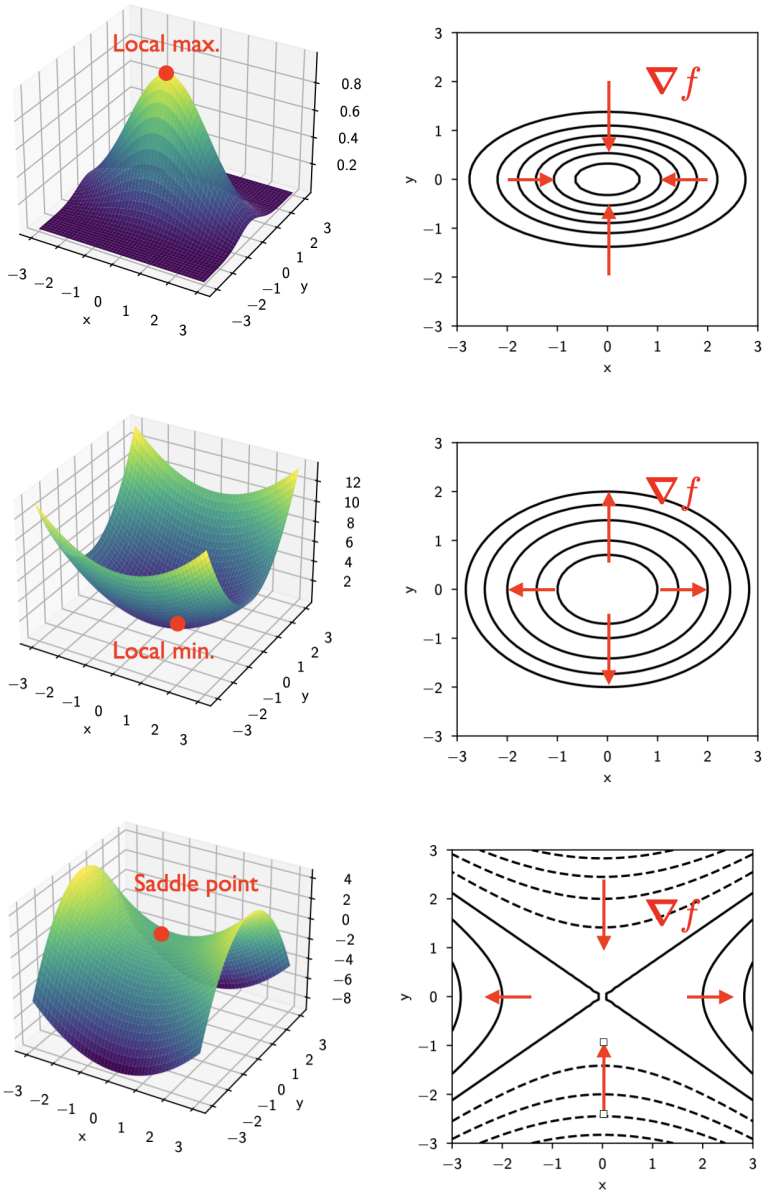
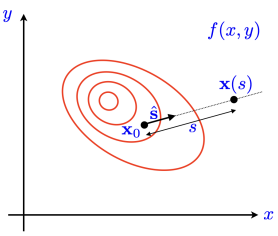


Figure 1: Illustrations of a local maximum (top), minimum (middle) and saddle point (bottom). The plots on the right show the contours near the stationary point and the gradient vector.

3.1 Taylor series for multivariate functions

Consider how a function $f(x, y)$ varies in the vicinity of the point $\mathbf{x}_0 = (x_0, y_0)$ as we move along the straight line through $\mathbf{x}_0$ in the direction of $\hat{\mathbf{s}}$. At distance s along this line from $\mathbf{x}_0$, we are at position

$$\mathbf{x}(s) = \mathbf{x}_0 + s\hat{\mathbf{s}}\,;$$



see the figure to the right.

Along this line, the function can be thought of as a function of s and the usual single-variable Taylor series holds, with the derivatives replaced by the directional derivative

$$\frac{df}{ds} = \hat{\mathbf{s}} \cdot \boldsymbol{\nabla} f\,.$$

It follows that

$$\begin{aligned} f(\mathbf{x}_0 + s\hat{\mathbf{s}}) &= f(\mathbf{x}_0) + s\left.\frac{df}{ds}\right|_{\mathbf{x}_0} + \frac{1}{2}s^2\left.\frac{d^2 f}{ds^2}\right|_{\mathbf{x}_0} + \cdots \\ &= f(\mathbf{x}_0) + s\,\hat{\mathbf{s}} \cdot \boldsymbol{\nabla} f|_{\mathbf{x}_0} + \frac{1}{2}s^2\,(\hat{\mathbf{s}} \cdot \boldsymbol{\nabla})(\hat{\mathbf{s}} \cdot \boldsymbol{\nabla}) f|_{\mathbf{x}_0} \\ &\qquad + \cdots\,. \end{aligned}$$

Let us write the finite displacement

$$\delta\mathbf{x} = s\hat{\mathbf{s}}\,,$$

with components $\delta x = x(s) - x_0$ and $\delta y = y(s) - y_0$. Then

$$s\hat{\mathbf{s}} \cdot \boldsymbol{\nabla} f = (\delta\mathbf{x}) \cdot \boldsymbol{\nabla} f = (\delta x)\frac{\partial f}{\partial x} + (\delta y)\frac{\partial f}{\partial y}$$

and

$$\begin{aligned} s^2(\hat{\mathbf{s}} \cdot \boldsymbol{\nabla})(\hat{\mathbf{s}} \cdot \boldsymbol{\nabla}) f &= (\delta\mathbf{x} \cdot \boldsymbol{\nabla})(\delta\mathbf{x} \cdot \boldsymbol{\nabla}) f \\ &= \left(\delta x\frac{\partial}{\partial x} + \delta y\frac{\partial}{\partial y}\right)\left(\delta x\frac{\partial f}{\partial x} + \delta y\frac{\partial f}{\partial y}\right) \\ &= (\delta x)^2\frac{\partial^2 f}{\partial x^2} + 2\delta x\delta y\frac{\partial^2 f}{\partial x \partial y} + (\delta y)^2\frac{\partial^2 f}{\partial y^2} \\ &= (\delta x, \delta y)\begin{pmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{pmatrix}\begin{pmatrix} \delta x \\ \delta y \end{pmatrix}. \end{aligned}$$

The matrix that appears here in the final line is the *Hessian matrix*.

Definition (Hessian matrix). The *Hessian matrix* is the matrix of second derivatives

$$\mathsf{H} \equiv \begin{pmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{pmatrix} = \boldsymbol{\nabla}\boldsymbol{\nabla} f \,.$$

The Hessian is a symmetric matrix since partial derivatives commute, i.e., $f_{xy} = f_{yx}$.

Putting these results together, we have the multivariate Taylor series

$$f(x_0+\delta x, y_0+\delta y) = f(x_0,y_0) + \left(\delta x \frac{\partial f}{\partial x} + \delta y \frac{\partial f}{\partial y}\right)\Bigg|_{x_0,y_0} + \frac{1}{2}\left((\delta x)^2 \frac{\partial^2 f}{\partial x^2} + 2\delta x \delta y \frac{\partial^2 f}{\partial x \partial y} + (\delta y)^2 \frac{\partial^2 f}{\partial y^2}\right)\Bigg|_{x_0,y_0} + \cdots .$$

We can also write this in coordinate-free form as

$$f(\mathbf{x}_0+\delta\mathbf{x}) = f(\mathbf{x}_0) + \delta\mathbf{x}\cdot(\boldsymbol{\nabla} f)|_{\mathbf{x}_0} + \frac{1}{2}\delta\mathbf{x}\,(\boldsymbol{\nabla}\boldsymbol{\nabla} f)|_{\mathbf{x}_0}\,\delta\mathbf{x}^T + \cdots .$$

3.2 Nature of stationary points and the Hessian

Recall that for functions of one variable, e.g., $f(x)$, if the second derivative $d^2f/dx^2 > 0$ at a stationary point then it is a minimum, while if $d^2f/dx^2 < 0$ it is a maximum. In this section, we develop the equivalent results for multivariate functions.

At a stationary point $\mathbf{x}_0$, we know that $\boldsymbol{\nabla} f = 0$. It follows that in the vicinity of $\mathbf{x}_0$, we have

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \frac{1}{2}\delta\mathbf{x}\,\mathsf{H}\,\delta\mathbf{x}^T + \cdots , \qquad (2)$$

where $\delta\mathbf{x} = \mathbf{x}-\mathbf{x}_0$ and the Hessian matrix H is evaluated at $\mathbf{x}_0$.

Definition (Positive-definite and negative-definite matrices). A (real) symmetric matrix $\mathbf{H}$ is *positive definite* if

$$\mathbf{x}\,\mathbf{H}\,\mathbf{x}^T > 0$$

for all non-zero (real) row vectors $\mathbf{x}$. Similarly, $\mathbf{H}$ is *negative definite* if

$$\mathbf{x}\,\mathbf{H}\,\mathbf{x}^T < 0$$

for all such $\mathbf{x}$. A matrix that is neither postive definite nor negative definite is sometimes called *indefinite*.

It follows that if the Hessian matrix is positive definite at a stationary point, then $\delta\mathbf{x}\,\mathbf{H}\,\delta\mathbf{x}^T > 0$ for all non-zero $\delta\mathbf{x}$. Equation (2) then implies that $f(\mathbf{x}) > f(\mathbf{x}_0)$ for all $\mathbf{x}$ sufficiently close to $\mathbf{x}_0$. This is just the definition of a local minimum, so we see that

$$\mathbf{H} \text{ positive definite} \Rightarrow \text{local minimum}.$$

Similarly, if $\mathbf{H}$ is negative definite, then $\delta\mathbf{x}\,\mathbf{H}\,\delta\mathbf{x}^T < 0$ for all non-zero $\delta\mathbf{x}$ and so $f(\mathbf{x}) < f(\mathbf{x}_0)$ in the vicinity of $\mathbf{x}_0$. The stationary point is therefore a local maximum:

$$\mathbf{H} \text{ negative definite} \Rightarrow \text{local maximum}.$$

If the matrix is indefinite, the stationary point may be a maximum, minimum or saddle (see below).

3.2.1 Definiteness and the eigenvalues

How can we determine whether a symmetric matrix is positive definite, negative definite or indefinite? As you know from *Vectors and Matrices*, any real symmetric matrix can be diagonalised by a suitable orthogonal transformation. Using coordinates along the principal axes, we have

$$\delta\mathbf{x}\,\mathbf{H}\,\delta\mathbf{x}^T = (\delta x_1, \delta x_2, \ldots, \delta x_n) \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} \begin{pmatrix} \delta x_1 \\ \delta x_2 \\ \vdots \\ \delta x_n \end{pmatrix},$$

where we have generalised to a function of n variables. The eigenvalues $\{\lambda_i\}$ are real since the matrix $\mathbf{H}$ is real symmetric.

If $\delta\mathbf{x}\,\mathbf{H}\,\delta\mathbf{x}^T > 0$ for all non-zero $\delta\mathbf{x}$, we see that we need all the eigenvalues to be positive. It follows that

$$\mathbf{H} \text{ positive definite} \Leftrightarrow \text{all } \lambda_i > 0.$$

Similarly,

$$\mathbf{H} \text{ negative definite} \Leftrightarrow \text{all } \lambda_i < 0.$$

On the other hand, if all the eigenvalues are non-zero but have mixed signs, then $\delta\mathbf{x}\,\mathbf{H}\,\delta\mathbf{x}^T$ can be positive, negative or zero depending on the direction. This case corresponds to the stationary point being a saddle point.

If any of the eigenvalues of the Hessian are zero, further analysis (e.g., higher terms in the Taylor series) is needed to determine the nature of the stationary point. For example, the function

$$f(x, y) = x^2 + y^4$$

has a (global) minimum at $x = 0$, $y = 0$. The Hessian is

$$\mathbf{H}(x, y) = \begin{pmatrix} 2 & 0 \\ 0 & 12y^2 \end{pmatrix}$$

and so reduces to $\mathrm{diag}(2, 0)$ at the stationary point. The eigenvalues there are 2 and 0.

3.2.2 Definiteness and the signature of the Hessian

There is an alternative way to establish if the Hessian matrix is positive or negative definite, using what is called the *signature* of the matrix. This avoids having to calculate the eigenvalues directly.

Definition (Signature of Hessian matrix). The *signature* of $\mathbf{H}$ is the pattern of the signs of the ordered subdeterminants of its leading principal minors. For a function of n variables, $f(x_1, x_2, \ldots, x_n)$, these subdeterminants are

$$\underbrace{f_{x_1x_1}}_{|\mathbf{H}_1|}, \underbrace{\begin{vmatrix} f_{x_1x_1} & f_{x_1x_2} \\ f_{x_2x_1} & f_{x_2x_2} \end{vmatrix}}_{|\mathbf{H}_2|}, \underbrace{\begin{vmatrix} f_{x_1x_1} & f_{x_1x_2} & f_{x_1x_3} \\ f_{x_2x_1} & f_{x_2x_2} & f_{x_2x_3} \\ f_{x_3x_1} & f_{x_3x_2} & f_{x_3x_3} \end{vmatrix}}_{|\mathbf{H}_3|}, \ldots, |\mathbf{H}_n| = |\mathbf{H}|.$$

It can be shown (*Sylvester's criterion*) that

$$\mathbf{H} \text{ positive definite} \Leftrightarrow \text{signature is } +, +, \ldots, +$$

and

$$\mathbf{H} \text{ negative definite} \Leftrightarrow \text{signature is } -, +, \ldots, (-1)^n.$$

It is straightforward to establish the forward implications here, for example that $\mathbf{H}$ being positive definite implies the $+, +, \ldots, +$ signature. If $\mathbf{H}$ is positive definite, then so too are all its principal minors. This follows from considering the quadratic form $\mathbf{x}\,\mathbf{H}\,\mathbf{x}^T > 0$ for vectors of the form $\mathbf{x} = (x_1, x_2, 0, \ldots, 0)$, for example. In this case, only the leading principal minor $\mathbf{H}_2$ is involved and so it must also be positive definite.[1] As all the principal minors are positive definite, they all have only positive eigenvalues and hence each has positive determinant. This establishes that the signature is $+, +, \ldots, +$. Similarly, if $\mathbf{H}$ is negative definite, so too are all its leading principal minors. It follows that all have only negative eigenvalues, and so the sign of $|\mathbf{H}_m|$ for $m = 1, \ldots, n$ is $(-1)^m$.

It takes more work to prove the converses in Sylvester's criterion (see non-examinable section below).

[1] This result means that if the quadratic function $\mathbf{x}\,\mathbf{H}\,\mathbf{x}^T$ has a minimum at $\mathbf{x} = 0$, it is also a minimum when the function is restricted to any lower-dimensional subspace that includes the origin. (In two dimensions, these would be straight lines through the origin.)

Sylvester's criterion (non-examinable)

Let us first sketch the proof of the converse in Sylvester's criterion for the positive-definite case. We aim to show that if the subdeterminants of the leading principal minors of a real-symmetric $n \times n$ matrix $\mathbf{H}$ are all positive, then $\mathbf{H}$ is positive definite. We start with $\mathbf{H}_1$, which has a single element h_{11}. If $|\mathbf{H}_1| > 0$, then $h_{11} > 0$ and $\mathbf{H}_1$ is positive definite.

We next show that if $\mathbf{H}_k$ is positive definite, and $|\mathbf{H}_{k+1}| > 0$, then $\mathbf{H}_{k+1}$ is also positive definite. If $|\mathbf{H}_{k+1}| > 0$, then its eigenvalues are either all positive, or all but two, four, etc., are positive and two, four, etc., are negative. We shall prove that it is not possible to have two or more negative eigenvalues by contradiction. Suppose that $\mathbf{H}_{k+1}$ does have two or more negative eigenvalues. Let two of the associated eigenvectors be $\mathbf{u}$ and $\mathbf{v}$, with components u_i and v_i for $i = 1, 2, \ldots, k+1$. Since these are the eigenvectors of a real-symmetric matrix, they may always be chosen to be orthogonal. Consider now the (row) vector

$$\mathbf{w} = v_{k+1}\mathbf{u} - u_{k+1}\mathbf{v}\,,$$

which by construction has no $k+1$ component. It follows that

$$\mathbf{w}\,\mathbf{H}_{k+1}\,\mathbf{w}^T = (v_{k+1})^2\,\mathbf{u}\,\mathbf{H}_{k+1}\,\mathbf{u}^T + (u_{k+1})^2\,\mathbf{v}\,\mathbf{H}_{k+1}\,\mathbf{v}^T < 0\,,$$

since $\mathbf{u}$ and $\mathbf{v}$ are eigenvectors of $\mathbf{H}_{k+1}$ with negative eigenvalues. However, since $\mathbf{w}$ has no $k+1$ component, evaluating $\mathbf{w}\,\mathbf{H}_{k+1}\,\mathbf{w}^T$ amounts to using $(w_1, w_2, \cdots, w_k)$ in the quadratic form constructed from $\mathbf{H}_k$. As $\mathbf{H}_k$ is positive definite, $\mathbf{w}\,\mathbf{H}_{k+1}\,\mathbf{w}^T > 0$ so we have a contradiction. It follows that all the eigenvalues of $\mathbf{H}_{k+1}$ are positive and so it is a positive-definite matrix.

Working through $\mathbf{H}_1$, $\mathbf{H}_2$ up to $\mathbf{H}_n = \mathbf{H}$, we see that if all have positive determinant, then they are all positive definite too. This establishes the converse in Sylvester's criterion for the positive-definite case.

To prove the negative-definite case, suppose that the determinants of the leading principal minors of the real-symmetric $n \times n$ matrix $\mathbf{H}$ have signature $-, +, \ldots, (-1)^n$. Consider the matrix $-\mathbf{H}$. Since multiplying a $k \times k$ matrix by -1 changes the determinant by $(-1)^k$, the matrix $-\mathbf{H}$ will have signature $+, +, \ldots, +$. As we have shown above, such a matrix must be positive definite. As $-\mathbf{H}$ is positive definite, $\mathbf{H}$ must be negative definite.

3.3 Contours near stationary points

Consider a function $f(x, y)$ with a stationary point at (x_0, y_0). Denote the Hessian matrix there by $\mathbf{H}$, and

adopt coordinates with axes aligned with the principal axes of $\mathbf{H}$. In these coordinates,

$$\mathbf{H} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} ,$$

where λ_1 and λ_2 are the eigenvalues, which we shall assume are non-zero. If we write

$$\mathbf{x} = \mathbf{x}_0 + (\xi, \eta) ,$$

then in the vicinity of $\mathbf{x}_0$

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \frac{1}{2} \left(\lambda_1 \xi^2 + \lambda_2 \eta^2 \right) .$$

The contours of f therefore locally satisfy

$$\lambda_1 \xi^2 + \lambda_2 \eta^2 = \text{const.} \qquad (3)$$

At a maximum or minimum, the eigenvaluesall have the same sign so the contours given by Eq. (3) are locally *elliptical*. At a saddle point, the eigenvalues have opposite signs and so the contours are locally *hyperbolic*.

Example. Consider the function

$$f(x, y) = 4x^3 - 12xy + y^2 + 10y + 6 .$$

We have

$$\begin{aligned} f_x &= 12x^2 - 12y , \\ f_y &= -12x + 2y + 10 . \end{aligned}$$

At stationary points, $f_x = 0$ and $f_y = 0$. The first of these gives $y = x^2$ and the second $6x = y + 5$. Substituting for $y = x^2$ we have

$$x^2 - 6x + 5 = 0 \quad \Rightarrow \quad x = 1 \text{ or } x = 5 .$$

The stationary points are therefore $(1, 1)$ and $(5, 25)$.

Evaluating the second derivatives, we have

$$\begin{aligned} f_{xx} &= 24x , \\ f_{xy} &= -12 , \\ f_{yy} &= 2 . \end{aligned}$$

Consider first the point $(1,1)$. The Hessian there is

$$\mathbf{H} = \begin{pmatrix} 24 & -12 \\ -12 & 2 \end{pmatrix} .$$

The subdeterminants of the leading principal minors are $|\mathbf{H}_1| = 24$ and $|\mathbf{H}_2| = |\mathbf{H}| = -96$. The signature is $+,-$ and so $\mathbf{H}$ is neither positive definite nor negative definite. In this two-dimensional case, we know from $|\mathbf{H}| < 0$ that the eigenvalues are opposite in sign and so we have a saddle point.

At the other stationary point, $(5,25)$, we have

$$\mathbf{H} = \begin{pmatrix} 120 & -12 \\ -12 & 2 \end{pmatrix} .$$

The subdeterminants of the leading principal minors are now $|\mathbf{H}_1| = 120$ and $|\mathbf{H}_2| = |\mathbf{H}| = 96$. The signature is $+,+$ and so we know from Sylvester's crterion that $\mathbf{H}$ is positive definite. We see that $(5,25)$ is a local minimum.

To determine the orientation of the contours near the stationary points, consider, for example, the saddle point $(1,1)$. Writing

$$(x,y) = (1,1) + (\delta x, \delta y) ,$$

the contours locally have

$$\begin{aligned} & f_{xx}(\delta x)^2 + 2f_{xy}\delta x\delta y + f_{yy}(\delta y)^2 = \text{const.} \\ \Rightarrow \quad & 12(\delta x)^2 - 12\delta x\delta y + (\delta y)^2 = \text{const.} \end{aligned}$$

The intersecting straight-line contours through the saddle point (which are also the asymptotes of the neighbouring hyperbolic contours) are therefore described by

$$12(\delta x)^2 - 12\delta x\delta y + (\delta y)^2 = 0 \quad \Rightarrow \quad \delta y = (6 \pm 2\sqrt{6})\delta x .$$

To sketch the contours, we can draw what they look like near the stationary points and then try to join them togther noting they only cross at the saddle point. The contours are shown in Fig. 2.

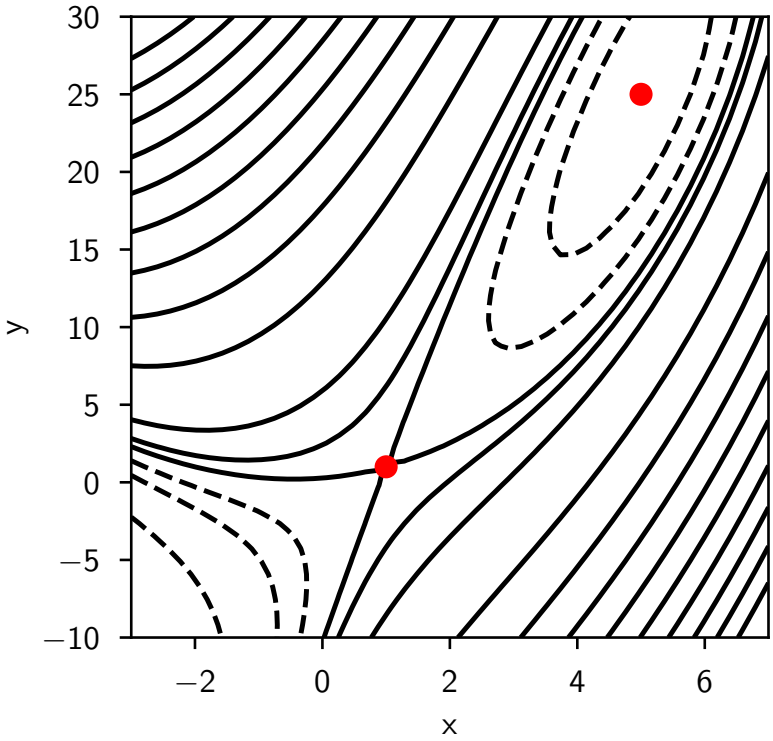


Figure 2: Contours of the function $f(x, y) = 4x^3 - 12xy + y^2 + 10y + 6$. (The contour levels are not equally spaced.) Note the shape of the contours close to the saddle point at $(1, 1)$ and the local minimum at $(5, 25)$.

4 Systems of linear differential equations

In this section we consider the behaviour of systems of first-order linear differential equations, where we have multiple dependent variables that may be coupled to each other.

Consider two functions, $y_1(t)$ and $y_2(t)$, which satisfy

$$\dot{y}_1 = ay_1 + by_2 + f_1(t)\,, \tag{4}$$
$$\dot{y}_2 = cy_1 + dy_2 + f_2(t)\,, \tag{5}$$

where a, b, c and d are constants. We can write these in vector form as

$$\dot{\mathbf{Y}} = \mathsf{M}\mathbf{Y} + \mathbf{F}\,,$$

where

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\,, \qquad \mathsf{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\,, \quad \mathbf{F} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}\,.$$

One way to solve Eqs (4) and (5) is to convert them into a higher-order equation for one of the dependent variables. For example, differentiating Eq. (4), we have

$$\begin{aligned} \ddot{y}_1 &= a\dot{y}_1 + b\dot{y}_2 + \dot{f}_1 \\ &= a\dot{y}_1 + b\left(cy_1 + dy_2 + f_2\right) + \dot{f}_1 \\ &= a\dot{y}_1 + bcy_1 + d\left(\dot{y}_1 - ay_1 - f_1\right) + bf_2 + \dot{f}_1\,, \end{aligned}$$

so that

$$\ddot{y}_1 - (a+d)\dot{y}_1 + (ad-bc)y_1 = bf_2 - df_1 + \dot{f}_1\,.$$

This is a linear, second-order differential equation with constant coefficients, which we know how to solve.

However, it is often more convenient to solve the first-order system of equations directly with the matrix methods developed in the rest of this section, rather than solving the higher-order equation. Indeed, we often go the other way and convert a linear higher-order differential equation to a system of (coupled) linear first-order

equations. This is particularly the case when solving equations numerically.

For example, the second-order equation

$$\ddot{y} + \alpha\dot{y} + \beta y = f\,,$$

where α and β are constant coefficients, can be recast as a first-order system by writing

$$y_1 = y \quad \text{and} \quad y_2 = \dot{y}\,,$$

so that

$$\begin{aligned}\dot{y}_1 &= y_2\,,\\ \dot{y}_2 &= \ddot{y} = -\alpha y_2 - \beta y_1 + f\,.\end{aligned}$$

In matrix form, this is

$$\dot{\mathbf{Y}} = \begin{pmatrix} 0 & 1 \\ -\beta & -\alpha \end{pmatrix} \mathbf{Y} + \begin{pmatrix} 0 \\ f \end{pmatrix}$$

with $\mathbf{Y} = (y_1, y_2)^T$.

4.1 Matrix methods

To solve a linear system of equations of the form

$$\dot{\mathbf{Y}} = \mathsf{M}\mathbf{Y} + \mathbf{F}(t)\,,$$

where the matrix M has constant elements, we proceed as follows.

1. We write $\mathbf{Y} = \mathbf{Y}_c + \mathbf{Y}_p$, where the complementary solution $\mathbf{Y}_c$ satisfies the homogeneous equation

$$\dot{\mathbf{Y}}_c = \mathsf{M}\mathbf{Y}_c\,. \tag{6}$$

2. We look for a complementary solution of the form $\mathbf{Y}_c = \mathbf{v}e^{\lambda t}$, where $\mathbf{v}$ is a constant vector. For this to satisfy Eq. (6), we must have

$$\mathsf{M}\mathbf{v} = \lambda\mathbf{v}\,,$$

i.e., $\mathbf{v}$ must be an eigenvector of M and then λ is the associated eigenvalue. For a system of n equations, there will be n such complementary solutions, and any linear combination of them is a solution of Eq. (6).

3. Finally, we find a particular solution, $\mathbf{Y}_p$, which satisfies the full system of forced equations. Its form will depend on the forcing vector $\mathbf{F}(t)$.

Example. Consider the linear system

$$\dot{\mathbf{Y}}_c = \underbrace{\begin{pmatrix} -4 & 24 \\ 1 & -2 \end{pmatrix}}_{\mathsf{M}} \mathbf{Y}_c + \begin{pmatrix} 4 \\ 1 \end{pmatrix} e^t \,. \tag{7}$$

We look for a complementary solution of the form $\mathbf{Y} = \mathbf{v}e^{\lambda t}$. The eigenvalues λ of the matrix M follow from $\det(\mathsf{M} - \lambda \mathsf{I}) = 0$, which gives

$$(\lambda + 8)(\lambda - 2) = 0 \quad \Rightarrow \quad \lambda = 2, -8 \,.$$

The associated eigenvectors are

$$\mathbf{v}_1 = \begin{pmatrix} 4 \\ 1 \end{pmatrix} \quad \text{for } \lambda_1 = 2$$

and

$$\mathbf{v}_2 = \begin{pmatrix} -6 \\ 1 \end{pmatrix} \quad \text{for } \lambda_2 = -8$$

The general complementary function is therefore

$$\mathbf{Y}_c = A \begin{pmatrix} 4 \\ 1 \end{pmatrix} e^{2t} + B \begin{pmatrix} -6 \\ 1 \end{pmatrix} e^{-8t} \,,$$

where A and B are constants.

For the particular solution, we try $\mathbf{Y}_p = \mathbf{u}e^t$, inspired by the time dependence of the forcing term. We require

$$\mathbf{u} = \mathsf{M}\mathbf{u} + \begin{pmatrix} 4 \\ 1 \end{pmatrix}$$

$$\Rightarrow \quad \begin{pmatrix} 5 & -24 \\ -1 & 3 \end{pmatrix} \mathbf{u} = \begin{pmatrix} 4 \\ 1 \end{pmatrix}$$

$$\Rightarrow \quad \mathbf{u} = -\frac{1}{9} \begin{pmatrix} 3 & 24 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} 4 \\ 1 \end{pmatrix} = - \begin{pmatrix} 4 \\ 1 \end{pmatrix} .$$

It follows that the general solution of the full system is

$$\mathbf{Y} = A \begin{pmatrix} 4 \\ 1 \end{pmatrix} e^{2t} + B \begin{pmatrix} -6 \\ 1 \end{pmatrix} e^{-8t} - \begin{pmatrix} 4 \\ 1 \end{pmatrix} e^{t} \, .$$

Note that if the time dependence of the forcing is $e^{\lambda t}$, where λ is an eigenvalue of M, then we should instead look for a particular solution

$$\mathbf{Y}_p = \mathbf{u} t e^{\lambda t} \, .$$

4.2 Non-degenerate phase portraits

The phase space for a system of n first-order differential equations is the n-dimensional space with points $\mathbf{Y} = (y_1, y_2, \ldots, y_n)^T$.

A phase portrait shows the solution trajectories in this space.

We shall consider the homogeneous equation

$$\dot{\mathbf{Y}} = \mathsf{M}\mathbf{Y} \, ,$$

which clearly has a fixed point at $\mathbf{Y} = 0$. For $n = 2$, the general solution of the equation is

$$\mathbf{Y}(t) = \mathbf{v}_1 e^{\lambda_1 t} + \mathbf{v}_2 e^{\lambda_2 t} \, , \tag{8}$$

where $\mathbf{v}_1$ and $\mathbf{v}_2$ are eigenvectors of M and λ_1 and λ_2 are the associated eigenvalues.

We shall only consider the possible forms of the phase portraits in the *non-degenerate* cases $\lambda_1 \neq 0$, $\lambda_2 \neq 0$, and $\lambda_1 \neq \lambda_2$.[2]

[2]For the degenerate case $\lambda_1 = \lambda_2 = \lambda$ (with λ real), in general $\mathsf{M} \neq \lambda \mathsf{I}$ and there is only a single eigenvector $\mathbf{v}$. The second solution is then of the form

$$\mathbf{Y}(t) = e^{\lambda t} \left(t\mathbf{v} + \mathbf{w} \right) \, ,$$

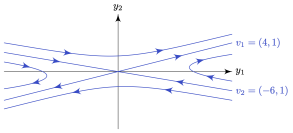
where the vector $\mathbf{w}$ satisfies $(\mathsf{M} - \lambda \mathsf{I})\mathbf{w} = \mathbf{v}$. Note that $\mathbf{w}$ is uniquely determined by this equation up to addition of multiples of $\mathbf{v}$, but such additional terms simply replicate the first solution. For the degenerate case when one of the eigenvalues vanishes, $\lambda_1 = 0$ say, the general solution is of the form

$$\mathbf{Y}(t) = \mathbf{v}_1 + \mathbf{v}_2 e^{\lambda_2 t} \, .$$

In the non-degenerate case, there are three distinct behaviours depending on the eigenvalues λ_1 and λ_2.

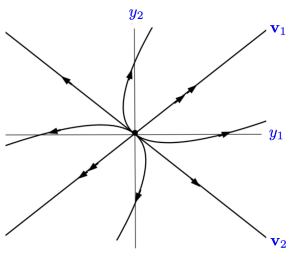
Case 1: λ_1 and λ_2 real and of opposite signs. Without loss of generality, we can take $\lambda_1 > 0$ and $\lambda_2 < 0$. The eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are necessarily real in this case. If $\mathbf{Y}$ starts out displaced from the origin along $\mathbf{v}_1$, it remains so and moves outwards as t increases (since $\lambda_1 > 0$). On the other hand, if $\mathbf{Y}$ starts out displaced along $\mathbf{v}_2$, it will move inwards along this direction approaching $\mathbf{Y} = 0$ as $t \to \infty$.

This case corresponds to a *saddle node*. An example phase portrait is shown to the right, corresponding to Eq. (7) with no forcing term. The arrows show the direction of evolution with increasing t. The curved trajectories in this figure can be added based on the flow direction along the eigenvectors. As $t \to \infty$, these curved lines become parallel to the eigenvector $\mathbf{v}_1$ with positive eigenvalue, while as $t \to -\infty$ they become parallel to $\mathbf{v}_2$.



Case 2: λ_1 and λ_2 real and of the same sign. Without loss of generality, we can take $|\lambda_1| > |\lambda_2|$. Again, the eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are necessarily real and if $\mathbf{Y}$ starts out displaced along these it will continue so, moving outwards as t increases for $\lambda_1 > 0$ and inwards for $\lambda_1 < 0$.

This case corresponds to a *stable node* if λ_1 and $\lambda_2 < 0$, and an *unstable node* if λ_1 and $\lambda_2 > 0$. An unstable node is illustrated in the figure to the right. Here, a generic trajectory is parallel to $\mathbf{v}_1$ as $t \to \infty$ (as $\lambda_1 > \lambda_2$) and approaches the origin along $\mathbf{v}_2$ as $t \to -\infty$. In the unstable case, the directions of the arrows are reversed.



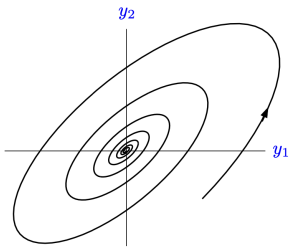
Case 3: λ_1 and λ_2 complex conjugate pairs. In this case, the eigenvectors are necessarily complex (as the matrix $\mathbf{M}$ is real) and are complex conjugates of each other: $\mathbf{v}_2 = \mathbf{v}_1^*$. Straight-line trajectrories are not possible.

The general solution, Eq. (8), for real $\mathbf{Y}$ can be written as

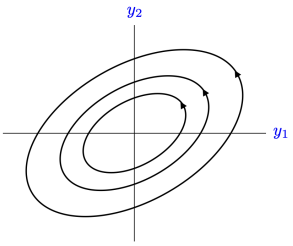
$$\begin{aligned}\mathbf{Y}(t) &= c\mathbf{v}_1 e^{\mathrm{Re}(\lambda_1)t} e^{i\mathrm{Im}(\lambda_1)t} + c^*\mathbf{v}_1^* e^{\mathrm{Re}(\lambda_1)t} e^{-i\mathrm{Im}(\lambda_1)t} \\ &= 2e^{\mathrm{Re}(\lambda_1)t}\{[c_1\mathrm{Re}(\mathbf{v}_1) - c_2\mathrm{Im}(\mathbf{v}_1)]\cos[\mathrm{Im}(\lambda_1)t] \\ &\qquad - [c_1\mathrm{Im}(\mathbf{v}_1) + c_2\mathrm{Re}(\mathbf{v}_1)]\sin[\mathrm{Im}(\lambda_1)t]\}\,,\end{aligned}$$

where the complex constant $c = c_1 + ic_2$, with c_1 and c_2 real.

Trajectories generally spiral around the origin. If $\mathrm{Re}(\lambda_1) < 0$, we have a *stable spiral*, whereby the trajectories spiral into the origin as $t \to \infty$ (see the figure to the right for an example). For $\mathrm{Re}(\lambda_1) > 0$, we have an *unstable spiral* and the trajectories spiral outwards with increasing t.



However, if $\mathrm{Re}(\lambda_1) = 0$ we have a *centre* and the solutions are periodic giving closed trajectories in phase space. These are generally elliptical and have common centres at the origin (see figure to the right).



To find the sense of rotation, it is sufficient to determine $\dot{\mathbf{Y}}$ at one point. For example, if we evaluate $\dot{\mathbf{Y}}$ at $\mathbf{Y} = (0, 1)^T$, then $\dot{y}_2 > 0$ there implies counter-clockwise rotation.

5 Nonlinear dynamical systems

In this section, we briefly introduce systems of nonlinear differential equations. In particular, we shall see how the techniques of the previous section for systems of linear equations can be used to investigate the stability of equilibrium points of the nonlinear system.

Consider an *autonomous system* of two nonlinear, first-order differential equations:

$$\begin{aligned} \dot{x} &= f(x, y)\,, \\ \dot{y} &= g(x, y)\,. \end{aligned} \tag{9}$$

The functions $f(x, y)$ and $g(x, y)$ are general, nonlinear functions of the dependent variables x and y but are independent of time t (hence the system is autonomous).

Solving such systems of equations can be very difficult. However, we can learn a lot about the phase-space trajectories of the solutions of these equations by studying the equilibrium points and their stability.

5.1 Equilibrium points

Definition (Equilibrium point). An *equilibrium point* (or fixed point) of the system of equations (9) is a point at which $\dot{x} = \dot{y} = 0$.

If (x_0, y_0) is a fixed point of Eq. (9), this requires

$$f(x_0, y_0) = 0 \quad \text{and} \quad g(x_0, y_0) = 0\,.$$

We must solve these equations simultaneously to determine (x_0, y_0).

To determine the stability of an equilibrium point, we conduct a perturbation analysis. Let

$$x(t) = x_0 + \xi(t) \quad \text{and} \quad y(t) = y_0 + \eta(t)\,,$$

where ξ and η are small perturbations around the fixed point. Substituting into Eq. (9), we have, for example,

$$\begin{aligned} \dot{\xi} &= f(x_0 + \xi, y_0 + \eta) \\ &\approx f(x_0, y_0) + \xi \frac{\partial f}{\partial x}(x_0, y_0) + \eta \frac{\partial f}{\partial y}(x_0, y_0) \\ &\approx \xi \frac{\partial f}{\partial x}(x_0, y_0) + \eta \frac{\partial f}{\partial y}(x_0, y_0)\,. \end{aligned}$$

Here, we have performed a multivariate Taylor expansion and dropped higher-order terms. Similarly,

$$\dot{\eta} \approx \xi \frac{\partial g}{\partial x}(x_0, y_0) + \eta \frac{\partial g}{\partial y}(x_0, y_0) \,.$$

We can combine these linear equations into the vector equation

$$\begin{pmatrix} \dot{\xi} \\ \dot{\eta} \end{pmatrix} = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} , \tag{10}$$

where the matrix of first derivatives is evaluated at (x_0, y_0). This is a linear system of first-order differential equations and we can apply the techniques of Sec. 4 to determine the nature of the equilibrium point. In particular, the stability is determined by the eigenvalues of the matrix of first derivatives.

Example (Population dynamics: predator–prey system). Consider an ecosytem with predators and prey. Let the number of prey at time t be $x(t)$ and the number of predators be $y(t)$. We model the dynamics of the prey as

$$\dot{x} = \alpha x - \beta x^2 - \gamma xy \,, \tag{11}$$

where α, β and γ are positive constants. In the absence of predators ($y = 0$), this is the logistic differential equation of Topic III, where, recall, α describes the excess rate of births over natural deaths and the term $-\beta x^2$ increases the death rate at high x to account for competition over some scarce resource. The term $-\gamma xy$ in Eq. (11) accounts for the prey being killed by the predators; we assume that the predators have infinite appetite, so consume all prey that they encounter.

We model the dynamics of the predators as

$$\dot{y} = \epsilon xy - \delta y \,,$$

where ϵ and δ are further positive constants. The first term on the right is the birth rate of predators, which increases if more prey is available to sustain the population. The final term is the natural death rate of the

predators. If there are no prey ($x = 0$), the number of predators decays exponentially.

We shall consider the following specific example:

$$\begin{aligned} \dot{x} &= 8x - 2x^2 - 2xy\,, \\ \dot{y} &= xy - y\,. \end{aligned} \tag{12}$$

The equilibrium points of this nonlinear, first-order autonomous system are where

$$2x(4 - x - y) = 0 \quad \text{and} \quad y(x - 1) = 0\,.$$

The first equation requires either $x = 0$ or $x = 4 - y$. In the former case, the second equation then requires $y = 0$ so we have an equilibrium point at $(0, 0)$. On the other hand, if $x = 4 - y$, the second equation reduces to

$$y(3 - y) = 0\,,$$

so either $y = 0$ or $y = 3$. We thus have two further equilibrium points: $(4, 0)$ and $(1, 3)$.

We consider the stability of these in turn using Eq. (10). Noting that

$$f(x, y) = 8x - 2x^2 - 2xy \quad \text{and} \quad g(x, y) = xy - y\,,$$

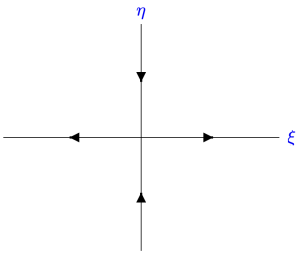
the required derivatives evaluate to

$$\begin{aligned} f_x &= 8 - 4x - 2y & \qquad f_y &= -2x\,, \\ g_x &= y & \qquad g_y &= x - 1\,. \end{aligned}$$

$(\mathbf{0}, \mathbf{0})$. Perturbations around this point evolve as

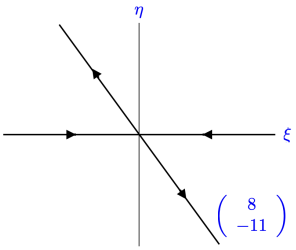
$$\begin{pmatrix} \dot{\xi} \\ \dot{\eta} \end{pmatrix} = \begin{pmatrix} 8 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix}\,.$$

The eigenvalues of the matrix are clearly 8 and -1 and associated eigenvectors are $(1, 0)^T$ and $(0, 1)^T$. As the eigenvalues are real and of opposite sign, the equilibrium point is a saddle node. Perturbations along the x-direction move away from $(0, 0)$, while those along the y-direction move towards it (see figure to the right). Note that if motion is restricted to $y = 0$, we recover the unstable nature of $x = 0$ for the logistic differential equation .



$(\mathbf{4}, \mathbf{0})$. Perturbations around this point evolve as

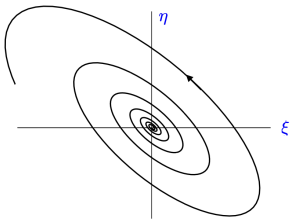
$$\begin{pmatrix} \dot{\xi} \\ \dot{\eta} \end{pmatrix} = \begin{pmatrix} -8 & -8 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} .$$



The eigenvalues of the matrix are -8 and 3 and the eigenvectors are $(1, 0)^T$ and $(8, -11)^T$, respectively. We see that this is also a saddle node, with displacements along the x-direction moving back towards the equilibrium point, but those along $(8, -11)^T$ moving away. For motion restricted to $y = 0$, we recover the stable nature of the equilibrium point $x = 4$ for the logistic differential equation.

$(\mathbf{1}, \mathbf{3})$. Finally, perturbations around this point evolve as

$$\begin{pmatrix} \dot{\xi} \\ \dot{\eta} \end{pmatrix} = \begin{pmatrix} -2 & -2 \\ 3 & 0 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} .$$



The eigenvalues of the matrix are $-1 \pm i\sqrt{5}$. Since these are a complex-conjugate pair with negative real part, the equilibrium point is a stable spiral. We can determine the sense of rotation by considering $(\xi, \eta) = (1, 0)$. For such a displacement, $\dot{\eta} = 3$. Since this is positive, the spiral is traversed anti-clockwise (see figure to the right).

The full phase portrait is shown in Fig. 3. The equilibrium (saddle) points at $(0, 0)$ and $(4, 0)$ are unstable and the introduction of any predators around these points will drive the system to spiral towards the stable equilibrium point at $(1, 3)$.

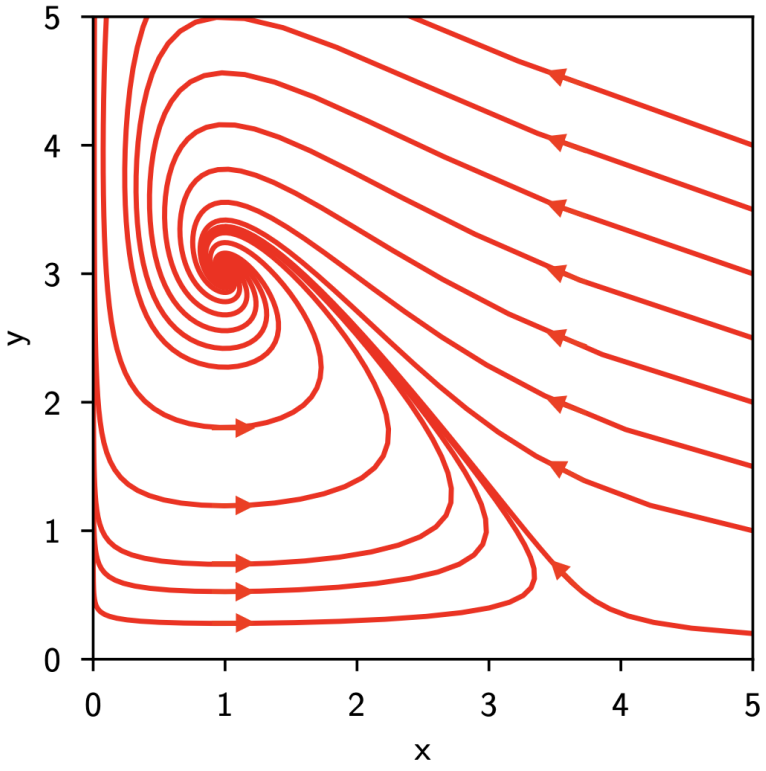


Figure 3: Phase portrait for the predator–prey system described by Eq. (12). Note the stable spiral equilibrium point at $(1, 3)$, where the number of prey $x = 1$ and predators $y = 3$.