



UNIVERSITÀ DEGLI STUDI ROMA TRE

Dipartimento di Matematica e Fisica  
Corso di Laurea Magistrale in Matematica

Tesi di Laurea Magistrale in Matematica

# **Aspetti algebro-geometrici e stabilità nei modelli epidemiologici**

Relatore  
Prof. Angelo Felice Lopez

Candidata  
Valeria Tarditi

Anno Accademico 2019/2020



*A mia madre e mio padre.  
Senza di voi non sarei mai potuta diventare chi sono oggi.*

# Indice

<b>Introduzione</b>	<b>i</b>
<b>Notazione</b>	<b>xi</b>
<b>1 Le varietà algebriche</b>	<b>1</b>
1.1 Varietà e spazi affini . . . . .	1
1.2 Ideali . . . . .	6
1.3 Spazi irriducibili e componenti irriducibili . . . . .	13
1.4 Dimensione di una varietà . . . . .	16
<b>2 Le basi di Groebner</b>	<b>17</b>
2.1 Ordinamenti monomiali in $R_n$ . . . . .	21
2.2 Algoritmo della divisione in $R_n$ . . . . .	27
2.3 Ideali monomiali e lemma di Dickson . . . . .	34
2.4 Teorema della base di Hilbert e basi di Groebner . . . . .	37
2.5 Proprietà delle basi di Groebner e criterio di Buchberger . . . . .	41
2.6 Basi di Groebner minimali e ridotte . . . . .	51
2.7 Zeri di un insieme di polinomi . . . . .	56
Risoluzione di equazioni polinomiali . . . . .	56
Forma implicita . . . . .	57
Varietà finite . . . . .	58
<b>3 Applicazione ai modelli epidemiologici</b>	<b>61</b>
3.1 Cenni di analisi . . . . .	64
Equazioni differenziali ordinarie . . . . .	64
Stabilità e punti di equilibrio . . . . .	66
3.2 Modelli epidemiologici . . . . .	68
Modelli SIR . . . . .	70
Modello SEIR . . . . .	104

# Introduzione

La geometria algebrica è un campo della matematica che, come il nome stesso suggerisce, unisce l'algebra astratta (soprattutto l'algebra commutativa) alla geometria. Gli oggetti di studio della Geometria Algebrica sono le varietà algebriche che sono oggetti geometrici definiti come soluzioni di sistemi di equazioni polinomiali in una o più indeterminate.

Lo scopo della tesi è quello di studiare alcuni argomenti di Geometria Algebrica applicata. In particolare, ci occuperemo di modelli epidemiologici, utilizzati anche per studiare l'andamento dello sviluppo del COVID-19.

Da questo studio sono emersi alcuni *risultati originali*, quali Teor. 0.1 - Teor. 0.10. Tali risultati sono sia di tipo algebro-geometrico (per ciascun modello è stata calcolata la varietà algebrica dei punti di equilibrio), sia di tipo analitico (è stata determinata la stabilità dei punti di equilibrio di ciascun modello epidemiologico).

L'epidemiologia è la scienza che ha per oggetto il fenomeno dell'insorgenza delle malattie nelle popolazioni, con particolare riguardo allo studio delle condizioni e dei fattori che le determinano. E' noto che esistono numerosi modelli matematici capaci di descrivere in modo quantitativo il diffondersi di un fenomeno infettivo. I modelli matematici costituiscono uno strumento importante in questo ambito e vengono usati sia come guida al controllo di infezioni endemiche sia come punto di partenza per affrontare nuove emergenze.

I modelli epidemiologici sono in grado di proporre parametri significativi per l'analisi e la classificazione delle malattie. La comprensione delle caratteristiche di trasmissione di malattie infettive può essere utile per determinare le strategie migliori per farne diminuire la trasmissione.

Per poterci occupare dei vari modelli epidemiologici abbiamo bisogno di studiare particolari basi di un ideale in un anello polinomiale, dette *basi di Groebner*, e le loro principali applicazioni. Nei primi capitoli della tesi forniremo un'introduzione ai concetti algebrici rilevanti. Per cominciare analizzeremo le varietà affini, ovvero punti (come nel caso di alcuni modelli epidemiologici), curve e superfici (e oggetti

di dimensione superiore) e, come vedremo, il modo migliore di definirle sarà usando gli ideali. Per comprenderle, dunque, avremo bisogno di un po' di algebra, e in particolare, avremo bisogno di studiare gli ideali nell'anello polinomiale  $\mathbb{K}[x_1, \dots, x_n]$ . Successivamente, studieremo il *metodo delle basi di Groebner*, che ci permetterà di risolvere problemi riguardanti anelli di polinomi in maniera algoritmica. Tale metodo viene utilizzato anche in molti programmi informatici di algebra per studiare particolari ideali polinomiali che si incontrano nelle applicazioni.

Nel secondo capitolo, in particolare, daremo le basi necessarie per poter risolvere i seguenti problemi.

- i. *Descrizione dell'ideale*: possiamo trovare un insieme finito di generatori per ogni ideale  $I \subset \mathbb{K}[x_1, \dots, x_n]$ ? O, equivalentemente, possiamo scrivere  $I = \langle f_1, \dots, f_s \rangle$  per qualche  $f_1, \dots, f_s \in \mathbb{K}[x_1, \dots, x_n]$  ?
- ii. *Problema di appartenenza all'ideale*: dati  $f \in \mathbb{K}[x_1, \dots, x_n]$  e  $I = \langle f_1, \dots, f_s \rangle$  in  $\mathbb{K}[x_1, \dots, x_n]$ , determinare se  $f \in I$ .
- iii. *Risoluzione di equazioni polinomiali*: siano  $f_1, \dots, f_s \in \mathbb{K}[x_1, \dots, x_n]$ . Possiamo determinare se esiste una soluzione comune a tutte le equazioni  $f_i=0$ ? Questo problema equivale a trovare tutti i punti  $p \in V(f_1, \dots, f_s)$ .
- iv. Possiamo determinare se  $V(f_1, \dots, f_n)$  è finita?
- v. *Forma Implicita*: siano  $f_1, \dots, f_n$  polinomi nelle variabili  $t_1, \dots, t_m$  e sia  $V$  il sottoinsieme di  $\mathbb{K}^n$  dato tramite la seguente parametrizzazione:

$$\begin{cases} x_1 = f_1(t_1, \dots, t_m) \\ \vdots \\ x_n = f_n(t_1, \dots, t_m) \end{cases}$$

Vogliamo trovare un sistema di equazioni polinomiali nelle variabili  $x_i$ , per  $1 \leq i \leq n$ , che definiscano la varietà.

I risultati originali della tesi emergono nel terzo e ultimo capitolo, nel quale ci occuperemo di descrivere i vari modelli epidemiologici. In particolare, analizzeremo le equazioni differenziali, il cui secondo membro è polinomiale, che li rappresentano. Per ogni modello epidemiologico studieremo:

- i. le *varietà dei punti di equilibrio*: gli zeri del sistema ottenuto annullando il secondo membro di ciascuna equazione differenziale sono chiamati punti di equilibrio. Per determinarli utilizzeremo i risultati ottenuti dallo studio del problema “*risoluzione di equazioni polinomiali*”. In particolare, come vedremo, sarà conveniente determinare la base di Groebner ridotta  $G_{rid}$  dell’ideale associato al sistema e calcolare  $V(G_{rid})$ .
- ii. la *stabilità dei punti di equilibrio* in funzione del tasso di riproduzione della malattia: attraverso l’analisi di queste equazioni è possibile determinare il *tasso di riproduzione* della malattia  $R_0$ , ovvero il numero medio di persone che ogni individuo infetta. Una volta trovato il valore  $R_0$  è possibile capire l’andamento della diffusione del virus, quindi se l’epidemia tenderà a diffondersi in tutta la popolazione o se tenderà a scomparire.

Il metodo di studio di ciascun modello epidemiologico prevede i seguenti passi.

- i. Presentiamo il modello matematico mediante la descrizione del corrispondente sistema di equazioni differenziali ordinarie.
- ii. Dal sistema di equazioni differenziali determiniamo l’ideale  $J$  associato a ciascun sistema.
- iii. Calcoliamo la base di Groebner, in particolare la base di Groebner ridotta  $G_{rid}$ , associata all’ideale  $J$ .
- iv. Calcolando  $V(G_{rid})$  otteniamo i punti di equilibrio di ciascun sistema. Nello specifico, dato che nei modelli epidemiologici le variabili sono sempre non negative, studieremo solo i punti di equilibrio con coordinate non negative.
- v. Calcoliamo la matrice Jacobiana associata al sistema in ciascun punto di equilibrio e i suoi rispettivi autovalori.
- vi. Definito  $R_0$ , applicando i teoremi di stabilità è possibile determinare se ciascun punto di equilibrio è stabile o instabile in funzione del valore  $R_0$ . In questo modo sapremo se l’epidemia si espanderà o meno.

Nei modelli epidemiologici la popolazione viene suddivisa in sottogruppi, ognuno dei quali rappresenta uno stadio specifico dell'epidemia.

Le categorie principali sono:

- I *suscettibili*, indicati con **S**, sono le persone non infette, ma che possono contrarre la malattia.
- Gli *infetti*, indicati con **I**, sono le persone infette, quindi quelle che possono diffondere il contagio per contatto.
- I *rimossi*, indicati con **R**: pazienti già infettati che non possono più diffondere la malattia perché sono guariti o deceduti.
- Gli *esposti*, indicati con **E**: soggetti contagiati ma non ancor infettivi.

Possiamo distinguere diversi modelli epidemiologici a seconda delle differenti categorie in cui vengono suddivisi gli individui della popolazione. Inoltre, i modelli epidemiologici che studiamo possono avere due tipi di struttura differente: *epidemica* o *endemica*. La prima fa riferimento ad un modello in cui la malattia infettiva è localizzata nel tempo, ed è tanto rapida da rendere trascurabili la nascita e le morti sulla popolazione (causate da motivi quali non la malattia infettiva in questione). La seconda invece descrive un modello in cui la malattia permane su tempi lunghi e quindi bisogna considerare le nascite e le morti sulla popolazione.

Lo studio dei modelli epidemiologici in base al metodo descritto sopra verrà esplicitato nei seguenti teoremi, tutti originali.

## MODELLI EPIDEMICI

### • MODELLO SIR

Il modello SIR prevede di suddividere la popolazione in 3 categorie: i Suscettibili (S), gli Infetti (I) e i Rimossi (R). Nel tempo, una parte crescente della popolazione suscettibile diventa infetta e gli infetti diventano rimossi. Il modo più semplice per descrivere questo tipo di processo infettivo è il seguente:

$$\begin{cases} \frac{dS}{dt} = -\beta SI \\ \frac{dI}{dt} = \beta SI - \gamma I \\ \frac{dR}{dt} = \gamma I \end{cases} \quad (1)$$

dove i parametri  $\beta > 0$  e  $\gamma > 0$ , rappresentano rispettivamente il *tasso di infezione* della malattia e il *tasso di recupero* dalla categoria degli infetti,

con condizioni iniziali  $S(0) = S_0 > 0$ ,  $I(0) = I_0 > 0$  e  $R(0) = 0$ .

**Teorema 0.1** (aspetto algebro-geometrico)

Sia  $J_{SIR_{ep}} = \langle -\beta SI, \beta SI - \gamma I, \gamma I \rangle$  l'ideale associato al modello SIR epidemico (1) nell'anello  $\mathbb{R}[S, I]$  con l'ordine monomiale  $lex S > I$ . Allora:

i. la base di Groebner ridotta di  $J_{SIR_{ep}}$  è

$$G_{rid} = \{I\};$$

ii. l'insieme dei punti di equilibrio è

$$V_{SIR_{ep}} = \{(S_e, 0, R_e), \forall S_e, R_e \in \mathbb{R}\}.$$

**Teorema 0.2** (stabilità)

Sia  $V_{SIR_{ep}} = \{(S_e, 0, R_e), \forall S_e, R_e \in \mathbb{R}\}$  la varietà associata al modello SIR epidemico (vedasi Teor. 0.1). Allora il punto di equilibrio  $(S_e, 0, R_e)$  con  $S_e \geq 0, R_e \geq 0$  è:

i. instabile se  $S_e > \frac{\gamma}{\beta}$ ;

ii. stabile (ma non asintoticamente stabile) se  $S_e \leq \frac{\gamma}{\beta}$ .

- MODELLO SIRD

Può succedere che l'infezione che si sta diffondendo potrebbe causare la morte di alcuni individui della popolazione. In tal caso è conveniente utilizzare un modello matematico che tenga in considerazione tale fatto. Indichiamo con  $D(t)$  il numero di individui morti a causa dell'infezione al tempo  $t$  con un tasso di morte  $\phi > 0$ . Il modello SIRD è descritto dal seguente sistema:

$$\begin{cases} \frac{dS}{dt} = -\beta SI \\ \frac{dI}{dt} = \beta SI - \phi I - \gamma I \\ \frac{dR}{dt} = \gamma I \\ \frac{dD}{dt} = \phi I, \end{cases} \quad (2)$$

dove i parametri  $\beta > 0$  e  $\gamma > 0$  rappresentano rispettivamente il *tasso di infezione* della malattia e il *tasso di recupero* dalla categoria degli infetti, con condizioni iniziali  $S(0) = S_0 > 0$ ,  $I(0) = I_0 > 0$ ,  $R(0) = 0$  e  $D(0) = 0$ .

**Teorema 0.3** (aspetto algebro-geometrico)

Sia  $J_{SIRD} = \langle -\beta SI, \beta SI - \phi I - \gamma I, \gamma I \rangle$  l'ideale associato al modello SIRD (2) nell'anello  $\mathbb{R}[S, I]$  con l'ordine monomiale  $\text{lex } S > I$ . Allora:

i. la base di Groebner ridotta di  $J_{SIRD}$  è

$$G_{rid} = \{I\};$$

ii. l'insieme dei punti di equilibrio è

$$V_{SIRD} = \{(S_e, 0, R_e, D_e), \forall S_e, R_e, D_e \in \mathbb{R}\}.$$

**Teorema 0.4** (stabilità)

Sia  $V_{SIRD} = \{(S_e, 0, R_e, D_e), \forall S_e, R_e, D_e \in \mathbb{R}\}$  la varietà associata al modello SIRD (vedasi Teor. 0.3). Allora il punto di equilibrio  $(S_e, 0, R_e, D_e)$ , con  $S_e \geq 0, R_e \geq 0, D_e \geq 0$  è:

i. instabile se  $S_e > \frac{\gamma + \phi}{\beta}$ ;

ii. stabile (ma non asintoticamente stabile) se  $S_e < \frac{\gamma + \phi}{\beta}$ .

## MODELLI ENDEMICI

Studieremo due modelli SIR endemici: nel primo, il modello SIR con vaccinazione, una parte dei suscettibili passa direttamente nella categoria dei rimossi con un *tasso di vaccinazione*  $\rho \geq 0$ , dunque il numero di persone suscettibili è  $(1 - \rho)N$ ; nel secondo invece, tutti i suscettibili sono predisposti ad essere infettati, dunque il tasso di vaccinazione è  $\rho = 0$ . Studieremo quindi il caso più generale in cui  $\rho \geq 0$  e applicheremo i risultati ottenuti al caso  $\rho = 0$ .

### • MODELLO SIR CON VACCINAZIONE

Tale modello può essere espresso dal sistema:

$$\begin{cases} \frac{dS}{dt} = -\beta SI - \mu S + (1 - \rho)\mu N \\ \frac{dI}{dt} = \beta SI - \gamma I - \mu I \\ \frac{dR}{dt} = \gamma I - \mu R + \rho\mu N, \end{cases} \quad (3)$$

dove i parametri  $\beta > 0$ ,  $\gamma > 0$  e  $\mu > 0$ , rappresentano rispettivamente il *tasso di infezione* della malattia, il *tasso di recupero* dalla categoria degli infetti e il *tasso di natalità e mortalità*, con condizioni iniziali  $S(0) = S_0 > 0$ ,  $I(0) = I_0 > 0$  e  $R(0) = 0$ .

**Teorema 0.5** (aspetto algebro-geometrico)

Sia  $J_{SIR_{vac}} = \langle -\beta SI - \mu S + (1 - \rho)\mu N, \beta SI - \gamma I - \mu I \rangle$  l'ideale associato al modello SIR con vaccinazione (3) nell'anello  $\mathbb{R}[S, I]$ , con l'ordine monomiale  $lex S > I$ . Sia

$$R_0 = \frac{(1 - \rho)\beta N}{\gamma + \mu}.$$

Allora:

i. la base di Groebner ridotta di  $J_{SIR_{vac}}$  è

$$G_{rid} = \left\{ S + \frac{(1 - \rho)\beta N}{\mu R_0} I - (1 - \rho)N, I^2 - \frac{\mu}{\beta}(R_0 - 1)I \right\};$$

ii. l'insieme dei punti di equilibrio è  $V_{SIR_{vac}} = \{P_1, P_2\}$  dove

$$P_1 = ((1 - \rho)N, 0, \rho N)$$

e

$$P_2 = \left( \frac{(1 - \rho)N}{R_0}, \frac{\mu}{\beta}(R_0 - 1), \frac{\gamma}{\beta}(R_0 - 1) + \rho N \right).$$

**Teorema 0.6** (stabilità)

Sia  $V_{SIR_{vac}} = \{P_1, P_2\}$  la varietà associata al modello SIR con vaccinazione (vedasi Teor. 0.5). Allora, nel caso di coordinate non negative:

- i. se  $R_0 < 1$ ,  $P_1$  è un punto di equilibrio asintoticamente stabile. Invece, se  $R_0 > 1$ ,  $P_1$  è un punto di equilibrio instabile;
- ii.  $P_2$  è un punto di equilibrio asintoticamente stabile.

## • MODELLO SIR

Nel modello SIR endemico, tutte le persone possono diventare infette. Applicando i risultati ottenuti nel modello SIR con vaccinazione (3) al caso  $\rho = 0$ , otteniamo le equazioni che rappresentano il modello SIR endemico:

$$\begin{cases} \frac{dS}{dt} = -\beta SI - \mu S + \mu N \\ \frac{dI}{dt} = \beta SI - \gamma I - \mu I \\ \frac{dR}{dt} = \gamma I - \mu R, \end{cases} \quad (4)$$

con condizioni iniziali  $S(0) = S_0 > 0$ ,  $I(0) = I_0 > 0$  e  $R(0) = 0$ , dove  $\mu > 0$ ,  $\beta > 0$  e  $\gamma > 0$  sono rispettivamente il tasso di natalità e mortalità, il tasso di infezione della malattia e il tasso di recupero dalla categoria degli infetti.

**Teorema 0.7** (aspetto algebro-geometrico)

Sia  $J_{SIR_{end}} = \langle -\beta SI - \mu S + \mu N, \beta SI - \gamma I - \mu I \rangle$  l'ideale associato al modello SIR endemico (4) nell'anello  $\mathbb{R}[S, I]$ , con l'ordine monomiale  $\text{lex } S > I$ . Sia

$$R_0 = \frac{\beta N}{\gamma + \mu}.$$

Allora:

i. la base di Groebner ridotta di  $J_{SIR_{end}}$  è

$$G_{rid} = \left\{ S + \frac{\beta N}{\mu R_0} I - N, I^2 - \frac{\mu}{\beta} (R_0 - 1) I \right\};$$

ii. l'insieme dei punti di equilibrio è  $V_{SIR_{end}} = \{P_1, P_2\}$  dove

$$P_1 = (N, 0, 0)$$

e

$$P_2 = \left( \frac{N}{R_0}, \frac{\mu}{\beta} (R_0 - 1), \frac{\gamma}{\beta} (R_0 - 1) \right).$$

**Teorema 0.8** (stabilità)

Sia  $V_{SIR_{end}} = \{P_1, P_2\}$  la varietà associata al modello SIR endemico vedasi Teor. 0.7). Allora, nel caso di coordinate non negative:

i. se  $R_0 < 1$ ,  $P_1$  è un punto di equilibrio asintoticamente stabile. Invece, se  $R_0 > 1$ ,  $P_1$  è un punto di equilibrio instabile;

ii.  $P_2$  è un punto di equilibrio asintoticamente stabile.

- MODELLO SEIR

In questo modello l'individuo suscettibile nel momento in cui è stato infettato, non viene spostato direttamente nella classe degli infetti, perchè potrebbe ancora non essere infettivo. Durante questo periodo l'individuo si trova nella classe degli esposti (E). Solo nel momento in cui diventa un vero e proprio infetto allora entra a far parte della popolazione della classe I con un *tasso di contagio*  $\sigma > 0$ .

Le equazioni che rappresentano il modello SEIR endemico sono:

$$\begin{cases} \frac{dS}{dt} = -\beta SI - \mu S + \mu N \\ \frac{dE}{dt} = \beta SI - \sigma E - \mu E \\ \frac{dI}{dt} = \sigma E - \gamma I - \mu I \\ \frac{dR}{dt} = \gamma I - \mu R, \end{cases} \quad (5)$$

dove i parametri  $\beta > 0$ ,  $\gamma > 0$  e  $\mu > 0$ , rappresentano rispettivamente il *tasso di infezione* della malattia, il *tasso di recupero* dalla categoria degli infetti e il *tasso di natalità e mortalità*, con condizioni iniziali  $S(0) = S_0 > 0$ ,  $E(0) = E_0 > 0$ ,  $I(0) = I_0 > 0$  e  $R(0) = 0$ .

**Teorema 0.9** (aspetto algebro-geometrico)

Sia  $J_{SEIR} = \langle -\frac{\beta\sigma}{\gamma+\mu}SE - \mu S + \mu N, \frac{\beta\sigma}{\gamma+\mu}SE - (\sigma + \mu)E \rangle$  l'ideale associato al modello SEIR (5) nell'anello  $\mathbb{R}[S, E]$ , con l'ordine monomiale  $\text{lex } S > E$ . Sia

$$R_0 = \frac{N\beta\sigma}{(\sigma + \mu)(\gamma + \mu)}.$$

Allora:

i. la base di Groebner ridotta di  $J_{SEIR}$  è

$$G_{rid} = \left\{ S + \frac{\sigma + \mu}{\mu}E - N, E^2 - \frac{\mu(\gamma + \mu)}{\beta\sigma}(R_0 - 1)E \right\};$$

ii. l'insieme dei punti di equilibrio è  $V_{SEIR} = \{P_1, P_2\}$  dove

$$P_1 = (N, 0, 0, 0) \quad \text{e} \quad P_2 = \left( \frac{N}{R_0}, \frac{\gamma + \mu}{\sigma} \frac{\mu}{\beta}(R_0 - 1), \frac{\mu}{\beta}(R_0 - 1), \frac{\gamma}{\beta}(R_0 - 1) \right).$$

**Teorema 0.10** (stabilità)

Sia  $V_{SEIR} = \{P_1, P_2\}$  la varietà associata al modello SEIR (vedasi Teor. 0.9). Allora, nel caso di coordinate non negative:

i. supponiamo che

$$(\gamma + 2\mu + \sigma)^2 \geq 4(\gamma + \mu)(\sigma + \mu)(1 - R_0). \quad (6)$$

Se  $R_0 < 1$  allora  $P_1$  è un punto di equilibrio asintoticamente stabile; se  $R_0 > 1$  allora  $P_1$  è un punto instabile;

- ii. se la (6) non è soddisfatta, il punto di equilibrio  $P_1$  è asintoticamente stabile;*
- iii.  $P_2$  è un punto di equilibrio asintoticamente stabile.*

In conclusione, per ogni modello epidemiologico sono state trovate le basi di Groebner dell'ideale associato al sistema di riferimento. Una volta calcolate le basi di Groebner si ottengono molte informazioni sulla varietà e sull'ideale associato. In particolare, a partire dalle basi di Groebner di ciascun ideale è possibile determinare la dimensione di Hilbert, il polinomio di Hilbert e la serie di Hilbert dell'ideale. Quest'ultimi sono importanti nella geometria algebrica computazionale, in quanto determinano la complessità dell'ideale e sono il modo più semplice noto per calcolare la dimensione e il grado di una varietà algebrica definita da equazioni polinomiali esplicite.

Inoltre, i modelli matematici epidemiologici e i rispettivi dati che sono stati trovati possono essere usati per pianificare, implementare e ottimizzare i programmi di prevenzione, terapia e controllo di ciascuna epidemia.

# Notazione

Sia  $\mathbb{K}$  un campo. Denotiamo con  $\mathbb{K}^* := \mathbb{K} \setminus \{0\}$  e con  $R_n = \mathbb{K}[x_1, \dots, x_n]$  **l'anello dei polinomi** in  $n$  variabili a coefficienti in  $\mathbb{K}$ .

# Capitolo 1

## Le varietà algebriche

Come anticipato nell'introduzione, in questi primi capitoli ci occuperemo di alcuni concetti algebro-geometrici. Per cominciare analizzeremo le varietà affini, ovvero punti (come nel caso di alcuni modelli epidemiologici), curve e superfici (e oggetti di dimensione superiore). Come vedremo, il modo migliore di definirle sarà usando gli ideali in  $R_n$ .

Tutto il capitolo 1 è stato scritto con riferimento le note della Prof.ssa L.Caporaso [C].

### 1.1 Varietà e spazi affini

#### **Definizione 1.1** (varietà affine)

*Dato un insieme  $T \subseteq R_n$ , chiamiamo varietà affine definita da  $T$ , l'insieme di tutti i punti di  $\mathbb{K}^n$  che annullano ogni polinomio di  $T$ :*

$$V(T) := \{p = (a_1, \dots, a_n) \in \mathbb{K}^n \mid f(p) = 0 \forall f \in T\}$$

**Osservazione 1.2** Sia  $f \in R_n$ .

$$V(f) := \{p \in \mathbb{K}^n \mid f(p) = 0\}$$

consiste al più di  $d$  punti dove  $d$  è il grado di  $f$ .

Inoltre, siccome  $R_n$  è un dominio di integrità, si ha  $V(f) = V(af) = V(f^m)$ ,  $\forall a \in \mathbb{K}^*$  e  $m \geq 1$ .

Consideriamo l'insieme  $\mathbb{K}^n$ , con  $n \geq 1$ ; possiamo definire su di esso una topologia per la quale i sottoinsiemi chiusi sono dati dalle soluzioni di sistemi di equazioni polinomiali a coefficienti in  $\mathbb{K}$ , detta *topologia di Zariski* su  $\mathbb{K}^n$ .

**Definizione 1.3 (spazi affini e topologia di Zariski)**

- La topologia di Zariski su  $\mathbb{K}^n$  è la topologia che ha per chiusi le varietà affini. Indichiamo la classe degli insiemi chiusi  $C$  in due modi equivalenti come segue:

$$C := \{V(T) \mid \forall T \subseteq R_n\} = \{V(T) \mid T \subset R_n : T \text{ finito}\}$$

- Lo spazio topologico definito su  $\mathbb{K}^n$  dalla topologia di Zariski si denota con  $\mathbb{A}_{\mathbb{K}}^n$  (o semplicemente  $\mathbb{A}^n$ ) e viene chiamato *n-spazio affine su  $\mathbb{K}$* . I suoi sottoinsiemi chiusi vengono chiamati *chiusi affini*.

In questo capitolo daremo le basi per dimostrare che la topologia di Zariski è una topologia. Si veda Prop. 1.15.

**Esempio 1.4** Diamo alcuni esempi di varietà:

- $V((0)) = \mathbb{A}^n(\mathbb{K})$  (dove  $(0)$  è l'insieme costituito dal polinomio nullo);
- $V(x^2 + y^2 - 1) \subset \mathbb{A}^2(\mathbb{R})$  è la circonferenza di raggio 1;

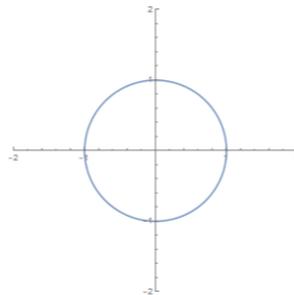


Figura 1.1: Circonferenza di raggio 1.  
Rappresentata con Wolfram Mathematica:  
`ParametricPlot[{Cos[u], Sin[u]}, {u, 0, 2Pi}]`

- $V(x^2 + y^2 - 1) \subset \mathbb{A}^3(\mathbb{R})$  è il cilindro di raggio di base 1;

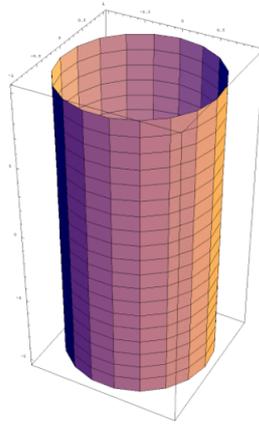


Figura 1.2: Cilindro di raggio 1 e altezza 4.

Rappresentato con Wolfram Mathematica:

```
ParametricPlot3D[{Sin[t], Cos[t], h}, {t, 0, 2Pi}, {h, -2, 2}]
```

- $V(x^2 + y^2 + z^2 - 1) \subset \mathbb{A}^3(\mathbb{R})$  è la sfera di raggio 1;

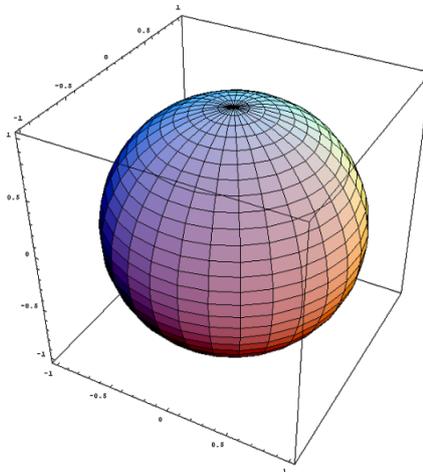


Figura 1.3: Sfera di raggio 1.

Rappresentata con Wolfram Mathematica:

```
ParametricPlot3D[{Sin[t]Cos[u], Cos[t]Cos[u], Sin[u]}, {t, 0, 2Pi}, {u, -Pi/2, Pi/2}]
```

- $V(\mathbb{R}_n) = \emptyset$  (basta pensare ad un polinomio  $f$  costante);
- $V(x^2 + y^2 + 1) = \emptyset \subset \mathbb{A}^2(\mathbb{R})$  perché l'equazione  $x^2 + y^2 = -1$  non è mai verificata nel campo dei reali;
- $V(xy, xy - 1) = \emptyset$  in qualsiasi campo perché non si possono annullare in contemporanea  $xy$  e  $xy - 1$ .

**Proposizione 1.5** *Siano  $T_1 \subseteq R_n$  e  $T_2 \subseteq R_n$ :*

- $T_1 \subseteq T_2 \implies V(T_1) \supseteq V(T_2)$
- Denotiamo  $T_1 \cdot T_2 = \{fg \mid f \in T_1 \text{ e } g \in T_2\}$   
allora  $V(T_1 \cdot T_2) = V(T_1) \cup V(T_2)$
- $\bigcap_{j \in J} V(T_j) = V(\bigcup_{j \in J} T_j)$ , dove  $J$  è un insieme qualsiasi di indici.

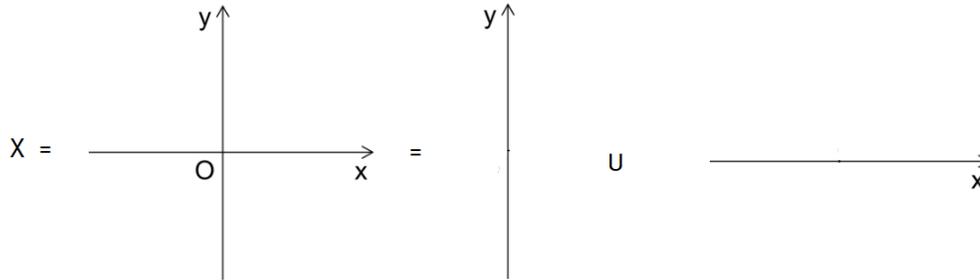
*Dimostrazione.*

- E' ovvia.
- Dimostriamo le due inclusioni:  
Sia  $p \in V(T_1 \cdot T_2)$ ; se  $p \notin V(T_1)$  allora esiste  $f \in T_1$  tale che  $f(p) \neq 0$ . Ora,  $\forall g \in T_2$  abbiamo  $(fg)(p) = 0$ , quindi  $f(p) \cdot g(p) = 0$  e dunque  $g(p) = 0$ . Se ne deduce che  $p \in V(T_2)$ .  
  
Sia ora  $p \in V(T_1) \cup V(T_2)$ ; se  $p \in V(T_1)$ , sicuramente  $p \in V(T_1 \cdot T_2)$  poichè  $(fg)(p) = f(p) \cdot g(p) = 0 \cdot g(p) = 0$ .  
(Identico il caso in cui  $p \in V(T_2)$ , basta scambiare il ruolo di  $f$  e  $g$ ).
- $V(\bigcup_{j \in J} T_j) = \{p \in \mathbb{K}^n \mid f(p) = 0 \forall f \in \bigcup_{j \in J} T_j\} =$   
 $\bigcap_{j \in J} \{p \in \mathbb{K}^n \mid f(p) = 0 \forall f \in T_j\} = \bigcap_{j \in J} V(T_j)$ .

□

**Esempio 1.6** Sia  $X = V(xy) \subset \mathbb{A}^2(\mathbb{R})$ .

E' facile verificare che  $V(xy) = V(x) \cup V(y)$ , ossia il luogo degli zeri di  $V(xy)$  è l'unione dei due assi cartesiani.



Gli esempi in questa sezione conducono a interessanti domande:

Supponiamo di avere  $f_1, \dots, f_n \in R_n$ .

- Possiamo determinare se  $V(f_1, \dots, f_n)$  è non vuota? Ovvero, esiste una soluzione comune a tutte le equazioni  $f_i=0$ ?
- Possiamo determinare se  $V(f_1, \dots, f_n)$  è finita?
- Possiamo determinare la dimensione <sup>1</sup> di  $V(f_1, \dots, f_n)$ ?

In questa tesi, daremo delle basi per rispondere a queste domande.

---

<sup>1</sup>Definiremo in seguito, si veda par.1.4.

## 1.2 Ideali

Come è stato anticipato nell'introduzione, ci occuperemo di studiare varietà affini definite da ideali. In particolare, in questo capitolo forniremo alcune nozioni fondamentali per quanto riguarda gli ideali dell'anello  $R_n$ .

**Definizione 1.7** *Sia  $I$  un ideale di  $R_n$ . Diremo che  $I$  è finitamente generato se esistono  $f_1, \dots, f_s \in R_n$  tali che  $I = \langle f_1, \dots, f_s \rangle$ , e, in tal caso,  $\{f_1, \dots, f_s\}$  costituiscono una base per  $I$ .*

**Definizione 1.8** *Un anello  $A$  è noetheriano se ogni suo ideale ammette un insieme finito di generatori.*

Dimostreremo in seguito (si veda Teor. 2.28) che  $R_n$  è un anello noetheriano (questo fatto è conosciuto come teorema della base di Hilbert).

Notiamo che un dato ideale può avere basi differenti e, mostreremo più avanti, si può scegliere un tipo di base particolarmente utile, detta base di Groebner.

**Osservazione 1.9** Sia  $T \subseteq R_n$ . Allora esistono  $f_1, \dots, f_s \in R_n$  tali che:

$$V(T) = V((T)) = V(\langle f_1, \dots, f_s \rangle) = V(f_1, \dots, f_s).$$

*Dimostrazione.* Dimostriamo le tre uguaglianze cominciando dalla prima e poi a seguire.  $V(T) = V((T))$  infatti un'inclusione è una conseguenza della Prop. 1.5, poichè  $T \subseteq (T)$  e quindi  $V((T)) \subseteq V(T)$ . Viceversa, sia  $p \in V(T)$ ;  $f(p) = 0 \forall f \in T$ . Ogni polinomio  $h \in (T)$  è combinazione lineare di polinomi di  $T$  quindi  $h(p) = 0$ . Se ne deduce che  $p \in V((T))$ .

Dato che  $R_n$  è un anello noetheriano, possiamo scrivere ogni ideale  $(T)$  di  $R_n$  nella forma  $\langle f_1, \dots, f_s \rangle$  con  $f_i \in R_n$ . Dunque  $V((T)) = V(\langle f_1, \dots, f_s \rangle)$ .

Inoltre, affinché un punto di  $\mathbb{K}^n$  sia in  $V(\langle f_1, \dots, f_s \rangle)$ , basta che annulli ciascun  $f_i$ : infatti, se  $p \in V(f_1, \dots, f_s)$ ,  $f_i(p) = 0 \forall 1 \leq i \leq s$ . In particolare,  $\forall f \in \langle f_1, \dots, f_s \rangle$ , esistono  $h_1, \dots, h_s \in R_n$  tali che  $f(p) = h_1(p)f_1(p) + \dots + h_s(p)f_s(p) = h_1 \cdot 0 + \dots + h_s \cdot 0 = 0$ ; ossia,  $p \in V(\langle f_1, \dots, f_s \rangle)$ .

Viceversa, se  $p \in V(\langle f_1, \dots, f_s \rangle)$ , ogni polinomio  $f \in \langle f_1, \dots, f_s \rangle$  è nullo in  $p$ , quindi, in particolare  $f_i(p) = 0 \forall i$ . Di conseguenza  $p \in V(f_1, \dots, f_s)$  e, dunque, possiamo concludere che  $V(\langle f_1, \dots, f_s \rangle) = V(f_1, \dots, f_s)$ .

□

Nella Def. 1.7 abbiamo introdotto il concetto di base di un ideale e la possibilità di avere più di una base.

- Che relazione c'è fra le basi di uno stesso ideale?
- Modificando la base di un ideale, la varietà si modifica?

**Proposizione 1.10** *Siano  $\{f_1, \dots, f_s\}$  e  $\{g_1, \dots, g_t\}$  basi di uno stesso ideale in  $R_n$ , allora  $V(f_1, \dots, f_s) = V(g_1, \dots, g_t)$ .*

*Dimostrazione.*  $V(f_1, \dots, f_s) = \{p \in \mathbb{K}^n \text{ tale che } f(p) = 0 \forall f \in \langle f_1, \dots, f_s \rangle\} =$   
 $= \{p \in \mathbb{K}^n \text{ tale che } f(p) = 0 \forall f \in \langle g_1, \dots, g_t \rangle\} = V(g_1, \dots, g_t)$ .

□

**Esempio 1.11** Per la Prop. 2.1,  $\langle x^4 - 1, x^6 - 1 \rangle = \langle x^2 - 1 \rangle$  essendo  $x^2 - 1$  l'MCD in  $\mathbb{K}[x]$ . Allora

$$V(x^4 - 1, x^6 - 1) = V(x^2 - 1) = \{\pm 1\}.$$

**Esempio 1.12**  $\langle x + xy, y + xy, x^2, y^2 \rangle = \langle x, y \rangle$  in  $\mathbb{K}[x, y]$ . Allora

$$V(x + xy, y + xy, x^2, y^2) = V(x, y) = \{(0, 0)\}.$$

**Esempio 1.13** Per la Prop. 2.1,  $\langle x^2 + 3x - 4, x^3 - 5x + 4 \rangle = \langle x - 1 \rangle$  essendo  $x - 1$  l'MCD in  $\mathbb{K}[x]$ . Allora

$$V(x^2 + 3x - 4, x^3 - 5x + 4) = V(x - 1) = \{1\}.$$

Nel prossimo capitolo studieremo particolari basi, detti base di Groebner, che insieme alla proposizione precedente, determinano un ottimo strumento per comprendere le varietà affini. In particolare dimostreremo che  $x - 1$  è una base di Groebner ridotta per  $\langle x^2 + 3x - 4, x^3 - 5x + 4 \rangle$ .

**Osservazione 1.14** Il viceversa della Prop. 1.10 non è vero.

Infatti, se  $V(f_1, \dots, f_s) = V(g_1, \dots, g_t)$  non è detto che le basi siano uguali. Consideriamo, ad esempio, gli ideali  $\langle x \rangle$  e  $\langle x^2 \rangle$  di  $\mathbb{K}[x]$ . Sicuramente si ha che  $V(x) = V(x^2)$  ma non è vero che  $\langle x \rangle = \langle x^2 \rangle$ . Infatti, basta considerare il generatore  $x$  e osservare che  $x \notin \langle x^2 \rangle$ .

Nella Def. 1.3 avevamo introdotto la topologia di Zariski. Adesso abbiamo tutte le basi necessarie per poter dimostrare che tale definizione è ben posta.

**Proposizione 1.15** *La topologia di Zariski è una topologia.*

*Dimostrazione.* Ricordiamo che i chiusi della topologia di Zariski sono definiti nel modo seguente:

$$C := \{V(T) \mid \forall T \subseteq R_n\} = \{V(T) \mid T \subseteq R_n : T \text{ finito}\}$$

L'uguaglianza di destra è vera e segue da Oss. 1.9, ricordando che  $R_n$  è un anello noetheriano.

$$V((0)) = \mathbb{A}^n(\mathbb{K}) \text{ e } V(1) = \emptyset.$$

Resta da dimostrare che la classe  $C$  è chiusa all'unione finita e all'intersezione qualsiasi. Segue direttamente dalla Prop. 1.5 che  $V(T_1 \cdot T_2) = V(T_1) \cup V(T_2)$  e che se  $J$  è un insieme qualsiasi di indici,  $\bigcap_{j \in J} V(T_j) = V(\bigcup_{j \in J} T_j)$ .

□

Le varietà affini danno origine ad una classe di ideali particolari:

**Definizione 1.16** *Sia  $X \subseteq \mathbb{K}^n$  un insieme. Allora poniamo*

$$\mathcal{I}(X) := \{f \in R_n \mid f(p) = 0 \forall p \in X\}$$

**Lemma 1.17** *Sia  $X \subseteq \mathbb{K}^n$ .  $\mathcal{I}(X) \subseteq R_n$  è un ideale, e prende il nome di ideale di  $X$ .*

*Dimostrazione.* Dobbiamo dimostrare le tre proprietà dell'ideale.

$0 \in \mathcal{I}(X)$ , poichè il polinomio nullo si annulla su tutto  $\mathbb{K}^n$ , e quindi, in particolare, anche in  $X$ .

Se  $f, g \in \mathcal{I}(X)$  e  $p \in X$ , allora  $(f + g)(p) = f(p) + g(p) = 0 + 0 = 0$  dunque  $f + g \in \mathcal{I}(X)$ .

Se  $f \in \mathcal{I}(X)$ ,  $h \in R_n$  e  $p \in X$ , allora  $(hf)(p) = h(p) \cdot f(p) = h(p) \cdot 0 = 0$  ossia  $hf \in \mathcal{I}(X)$ .

□

**Osservazione 1.18** *Sia  $V$  una varietà affine di  $\mathbb{K}^n$ , l'ideale della varietà affine  $V$  è  $\mathcal{I}(V) \subseteq R_n$ .*

**Esempio 1.19** *Diamo alcuni esempi di ideali di  $V$ :*

- $\mathcal{I}(\emptyset) = R_n$
- Sia  $\mathbb{K}$  algebricamente chiuso, allora  $\mathcal{I}(\mathbb{K}^n) = (0)$ .<sup>2</sup>

---

<sup>2</sup>Dim. Segue dal Teor. 2.28.

**Esempio 1.20**  $\mathcal{I}(\{(0,0)\}) = \langle x, y \rangle \subset \mathbb{K}[x, y]$ .

*Dimostrazione.* Basta dimostrare le due inclusioni.

Sicuramente  $\langle x, y \rangle \subseteq \mathcal{I}(\{(0,0)\})$  poichè i polinomi della forma  $A(x, y)x + B(x, y)y$  si annullano nell'origine di  $\mathbb{K}^2$ .

Viceversa, sia  $f = \sum_{i,j} a_{ij} x^i y^j \in \mathcal{I}(\{(0,0)\})$ , di conseguenza,  $a_{00} = f(0, 0) = 0$ . Dunque,

$$\begin{aligned} f &= a_{00} + \sum_{i,j \neq 0,0} a_{ij} x^i y^j = 0 + \left( \sum_{i>0,j} a_{ij} x^{i-1} y^j \right) x + \left( \sum_{j>0} a_{0j} y^{j-1} \right) y = \\ &= \left( \sum_{i>0,j} a_{ij} x^{i-1} y^j \right) x + \left( \sum_{j>0} a_{0j} y^{j-1} \right) y \end{aligned}$$

pertanto  $f \in \langle x, y \rangle$ . □

**Lemma 1.21** *Siano  $f_1, \dots, f_s \in R_n$ , allora*

$$\langle f_1, \dots, f_s \rangle \subseteq \mathcal{I}(V(f_1, \dots, f_s)),$$

*sebbene l'uguaglianza non sia vera in generale.*

*In particolare se  $X = V(J) \subseteq \mathbb{K}^n$  è un insieme chiuso allora  $J \subseteq \mathcal{I}(X)$ .*

*Dimostrazione.* Sia  $f \in \langle f_1, \dots, f_s \rangle$ . Allora  $f = \sum_{i=1}^s h_i \cdot f_i$  per qualche polinomio  $h_1, \dots, h_n \in R_n$ . Poichè gli  $f_1, \dots, f_s$  si annullano in  $V(f_1, \dots, f_s)$ , si annulla anche  $f = \sum_{i=1}^s h_i \cdot f_i \equiv 0$ . Di conseguenza,  $f \in \mathcal{I}(V(f_1, \dots, f_s))$ .

Per dimostrare che l'uguaglianza non è vera in generale, facciamo vedere che l'inclusione  $\langle x^2, y^2 \rangle \subset \mathcal{I}(V(x^2, y^2))$  è stretta.

Per prima cosa calcoliamo  $\mathcal{I}(V(x^2, y^2))$ .

Osserviamo che da  $x^2 = 0, y^2 = 0$  segue  $V(x^2, y^2) = \{(0,0)\}$ .

Inoltre per l'esempio 1.20,  $\mathcal{I}(\{(0,0)\}) = \langle x, y \rangle$ . Dunque non è un'uguaglianza perché  $\langle x, y \rangle \not\subseteq \langle x^2, y^2 \rangle$ . Infatti,  $x \notin \langle x^2, y^2 \rangle$ : se così non fosse, dovrebbero esistere due polinomi  $p(x, y)$  e  $q(x, y) \in R_n$ , tali che  $x = p(x, y)x^2 + q(x, y)y^2$  ma questo non è possibile in quanto il polinomio  $p(x, y)x^2 + q(x, y)y^2$  o è nullo oppure ha grado almeno 2. □

**Definizione 1.22** *Sia  $I$  un ideale di  $R_n$ . Il radicale di  $I$  è l'ideale*

$$\sqrt{I} := \{f \in R_n : f^m \in I \text{ per qualche } m \geq 1\}.$$

*Un ideale  $I$  si dice ideale radicale se  $\sqrt{I} = I$ .*

**Osservazione 1.23**  $\sqrt{I}$  è un ideale di  $R_n$ .

*Dimostrazione.* Dimostriamo le tre proprietà dell'ideale.

$0 \in \sqrt{I}$  poichè  $0 \in I$ .

Se  $f, g \in \sqrt{I}$ , allora  $f^n$  e  $g^m \in I$ , per qualche  $n, m \geq 1$ . Quindi

$$(f + g)^{n+m} = \sum_{k=0}^{n+m} \binom{n+m}{k} f^k g^{n+m-k}$$

è ancora un elemento di  $I$ : infatti, o  $k \geq n$  e in questo caso  $f^k \in I$ , oppure  $k < n$ , quindi  $n + m - k \geq m$  e, dunque,  $g^{n+m-k} \in I$ . Si conclude che  $f + g \in \sqrt{I}$ .

Ora, siano  $h \in R_n$  ed  $f \in \sqrt{I}$ ,  $(hf)^n = h^n f^n \in I$  poichè  $f^n \in I$  per quale  $n \geq 1$ .

Di conseguenza,  $hf \in \sqrt{I}$ . □

**Osservazione 1.24**  $V(I) = V(\sqrt{I})$ .

*Dimostrazione.* Dimostriamo le due inclusioni:

Osserviamo che ponendo  $m = 1$  si ottiene,  $I \subseteq \sqrt{I}$ , dunque, per la Prop. 1.5,  $V(\sqrt{I}) \subseteq V(I)$ .

Viceversa, sia  $p \in V(I)$ , allora  $p$  annulla tutti i polinomi di  $I$ , in particolare se  $g$  è nel radicale di  $I$ ,  $g^m(p) = 0$  per quale  $m \geq 1$ ; essendo  $\mathbb{K}$  un dominio,  $g(p) = 0$  per ogni  $g \in \sqrt{I}$ . Allora  $p \in V(\sqrt{I})$ . □

**Proposizione 1.25** Siano  $I_1$  e  $I_2$  ideali di  $R_n$ .

$$V(I_1 \cap I_2) = V(I_1) \cup V(I_2)$$

*Dimostrazione.* Dimostriamo le due inclusioni.

Osserviamo che  $I_1 \cap I_2 \subseteq I_i$ , dunque per la Prop. 1.5,  $V(I_i) \subseteq V(I_1 \cap I_2)$  per  $i = 1, 2$ .

Viceversa sia  $p \in V(I_1 \cap I_2)$ ; supponiamo  $p \notin V(I_1)$ , allora esiste  $f \in I_1$  tale che  $f(p) \neq 0$ . Ora,  $\forall g \in I_2$  abbiamo  $fg \in I_1 \cap I_2$ , poichè  $I_1$  e  $I_2$  sono ideali. Allora,  $(fg)(p) = f(p) \cdot g(p) = 0$  e dunque  $g(p) = 0$ . Se ne deduce che  $p \in V(I_2)$ ; in conclusione  $V(I_1 \cap I_2) \subseteq V(I_1) \cup V(I_2)$ . □

**Lemma 1.26** *Sia  $X$  un chiuso di  $\mathbb{K}^n$ . Allora*

$$V(\mathcal{I}(X)) = X.$$

*Dimostrazione.* Dimostriamo le due inclusioni.

Sia  $p \in X$ . Per ogni  $f \in \mathcal{I}(X)$  si ha che  $f(p) = 0$ . Quindi  $p \in V(\mathcal{I}(X))$ . Dunque  $X \subseteq V(\mathcal{I}(X))$ . Viceversa, per ipotesi sappiamo che  $X$  è chiuso, quindi esiste un ideale  $J$  tale che  $X = V(J)$ . Sia  $p \in V(\mathcal{I}(X))$ . Se  $p \notin X$  allora esiste  $f \in J$  tale che  $f(p) \neq 0$ . Ma, per il Lemma 1.21, si ha che  $J \subseteq \mathcal{I}(X)$ , quindi  $f \in \mathcal{I}(X)$  e dunque  $f(p) = 0$ , contraddizione. Ne segue che  $p \in X$  ossia  $V(\mathcal{I}(X)) \subseteq X$ .

□

**Corollario 1.27** *La seguente mappa è iniettiva:*

$$\begin{aligned} \mathcal{I} : \{\text{chiusi di } \mathbb{A}^n(\mathbb{K})\} &\longrightarrow \{\text{ideali radicali di } R_n\} \\ X &\longrightarrow \mathcal{I}(X) \end{aligned}$$

*Dimostrazione.* Se  $X_1$  e  $X_2$  sono chiusi e  $\mathcal{I}(X_1) = \mathcal{I}(X_2)$  allora, per il Lemma 1.26  $X_1 = V(\mathcal{I}(X_1)) = V(\mathcal{I}(X_2)) = X_2$ .

□

Abbiamo osservato nel Lemma 1.21 che, in generale, se  $X = V(J) \subseteq \mathbb{K}^n$  è un insieme chiuso, allora  $J \subseteq \mathcal{I}(X)$ . Abbiamo fatto vedere con un controesempio che non è sempre vera l'uguaglianza.

Quale condizione deve essere imposta affinché l'inclusione diventi un'uguaglianza?

Il risultato seguente risponde a questa domanda.

**Teorema 1.28 (teorema degli zeri di Hilbert)**

*Sia  $\mathbb{K}$  un campo algebricamente chiuso e sia  $J \subseteq R_n$  un ideale. Sia  $f \in R_n$  tale che per ogni  $p \in V(J)$  si abbia  $f(p) = 0$ ; allora  $f$  appartiene al radicale di  $J$ .<sup>3</sup>*

*Equivalentemente:*

$$\mathcal{I}(V(J)) = \sqrt{J}.$$

---

<sup>3</sup>Dim. Vedasi [C, Teor. 9.10.3]

**Teorema 1.29** <sup>4</sup> Sia  $\mathbb{K}$  algebricamente chiuso. Allora la corrispondenza

$$\begin{aligned} V : \{\text{ideali radicali di } R_n\} &\longrightarrow \{\text{chiusi di } \mathbb{A}^n(\mathbb{K})\} \\ I &\longrightarrow V(I) \end{aligned}$$

è biunivoca e la sua inversa è  $\mathcal{I}$ .

$$\begin{aligned} \mathcal{I} : \{\text{chiusi di } \mathbb{A}^n(\mathbb{K})\} &\longrightarrow \{\text{ideali radicali di } R_n\} \\ X &\longrightarrow \mathcal{I}(X) \end{aligned}$$

Inoltre siano  $V$  e  $W$  sottoinsiemi chiusi in  $\mathbb{A}^n(\mathbb{K})$ , allora valgono le seguenti proprietà:

- i.  $V \subseteq W \iff \mathcal{I}(V) \supseteq \mathcal{I}(W)$
- ii.  $\mathcal{I}(V \cup W) = \mathcal{I}(V) \cap \mathcal{I}(W)$
- iii.  $\mathcal{I}(V \cap W) = \sqrt{\mathcal{I}(V) + \mathcal{I}(W)}$ .

---

<sup>4</sup>Dim. Vedasi [C, Teor.2.2.6]

### 1.3 Spazi irriducibili e componenti irriducibili

**Esempio 1.30** Sia  $\mathbb{K}$  un campo algebricamente chiuso .

Si consideri una ipersuperficie  $V(f) \subset \mathbb{A}^n(\mathbb{K})$ , con  $f$  non costante e quindi  $V(f)$  non vuoto.

Se il polinomio  $f$  è riducibile e

$$f = g_1^{d_1} \cdots g_m^{d_m}$$

con  $g_i$  polinomi irriducibili distinti e  $m \geq 2$ , abbiamo una decomposizione di  $V(f)$  come unione finita di sottoinsiemi chiusi propri, e non vuoti

$$V(f) = \bigcup_{i=1}^m V(g_i).$$

Tale  $V(f)$  è un semplice esempio di chiuso “riducibile”, secondo la seguente definizione puramente topologica:

**Definizione 1.31** Sia  $X$  uno spazio topologico.

- i.*  $X$  si dice riducibile se contiene due sottoinsiemi propri chiusi  $X_1$  e  $X_2$ , tali che  $X = X_1 \cup X_2$ .
- ii.* Uno spazio topologico non riducibile si dice irriducibile.
- iii.* Sia  $Y \subset X$  un chiuso irriducibile massimale (ovvero, non propriamente contenuto in alcun chiuso irriducibile di  $X$ ). Allora  $Y$  si dice componente irriducibile di  $X$ .

**Lemma 1.32** Sia  $X \subseteq \mathbb{A}^n$  un insieme chiuso, non vuoto.  $X$  è irriducibile se e solo se  $\mathcal{I}(X)$  è un ideale primo.

*Dimostrazione.* Sicuramente  $\mathcal{I}(X) \neq R_n$ . Infatti se fosse  $\mathcal{I}(X) = R_n$ , per il Lemma 1.26 si avrebbe che  $X = V(\mathcal{I}(X)) = V(R_n) = \emptyset$ , contraddizione.

Sia  $\mathcal{I}(X)$  non primo, esistono  $f$  e  $g$  in  $R_n \setminus \mathcal{I}(X)$  tali che  $f g \in \mathcal{I}(X)$ ; dunque  $X \subset V(fg) = V(f) \cup V(g)$ . Denotiamo con  $X_1 = X \cap V(f)$  e  $X_2 = X \cap V(g)$ ; abbiamo quindi che  $X = X_1 \cup X_2$ .

Inoltre  $X_1$  e  $X_2$  sono chiusi propriamente contenuti in  $X$ , dato che nè  $f$  nè  $g$  si annullano identicamente su  $X$ . Di conseguenza,  $X$  è riducibile.

Viceversa sia  $\mathcal{I}(X)$  primo e  $X = X_1 \cup X_2$  con  $X_i$  chiusi. Mostriamo che almeno

uno dei due  $X_i$  sia uguale a  $X$ .

Se  $X_1 \neq X$ , o equivalentemente  $X_1 \subsetneq X$ , allora  $\mathcal{I}(X) \subsetneq \mathcal{I}(X_1)$  propriamente. Infatti, se così non fosse, per il Lemma 1.26 applicato ai chiusi  $X$  e  $X_1$ , si avrebbe che  $V(\mathcal{I}(X)) = V(\mathcal{I}(X_1))$ , o equivalentemente,  $X = X_1$ .

Sia  $f \in \mathcal{I}(X_1) \setminus \mathcal{I}(X)$ ,  $\forall g \in \mathcal{I}(X_2)$ ,  $fg \in \mathcal{I}(X_1) \cap \mathcal{I}(X_2) = \mathcal{I}(X_1 \cup X_2) = \mathcal{I}(X)$ . Dunque,  $fg \in \mathcal{I}(X)$ , e in particolare  $g \in \mathcal{I}(X)$  essendo  $\mathcal{I}(X)$  primo. Di conseguenza,  $g \in \mathcal{I}(X_2) \subseteq \mathcal{I}(X)$ , e siccome  $\mathcal{I}(X) \subseteq \mathcal{I}(X_2)$ , si ha  $\mathcal{I}(X_2) = \mathcal{I}(X)$ ; per il Cor. 1.27,  $\mathcal{I}(X)$  è iniettiva, quindi  $X = X_1$ . □

**Osservazione 1.33** Sia  $\mathbb{K}$  algebricamente chiuso. Se  $X$  è un chiuso riducibile, allora, per qualunque ideale  $I$  tale che  $X = V(I)$ ,  $I$  non è primo.

**Esempio 1.34** Consideriamo l'insieme algebrico  $X$  definito nell'esempio 1.6.

$X$  è un insieme riducibile, infatti abbiamo dimostrato che valgono le seguenti uguaglianze:  $X = V(xy) = V(x) \cup V(y) = X_1 \cup X_2$ , con  $X_1 = V(x)$  e  $X_2 = V(y)$  sottoinsiemi chiusi propri di  $X$ . Per il Lemma 1.32 allora possiamo concludere che l'ideale  $I(X) = (xy)$  non è primo. Infatti se consideriamo due polinomi  $f = x$  e  $g = y$  notiamo che  $fg \in I(X)$  ma  $f \notin I(X)$  e  $g \notin I(X)$  in quanto  $x \notin (xy)$  e  $y \notin (xy)$ .

**Definizione 1.35** Uno spazio topologico  $X$  si dice noetheriano se ogni catena discendente di chiusi è stazionaria; ossia, per ogni catena

$$X \supseteq X_1 \supseteq X_2 \dots \supseteq X_m \dots \tag{1.1}$$

con  $X_i$  chiuso in  $X$ , per ogni  $i$ , esiste un  $m$  tale che  $X_m = X_{m+h}$  per ogni  $h$  positivo.

Come abbiamo già detto nel prossimo capitolo dimostreremo che  $R_n$  è un anello noetheriano, per il momento diamo per buono tale fatto e dimostriamo i seguenti lemmi:

**Lemma 1.36** Sia  $X$  un chiuso di  $\mathbb{A}^n$ , allora  $X$  è uno spazio topologico noetheriano.

*Dimostrazione.* Dire che uno spazio topologico  $X$  è noetheriano equivale a dire che soddisfa la condizione della catena discendente sui sottoinsiemi chiusi descritta nella Def. 1.35. Questa condizione per i chiusi di  $\mathbb{A}^n$  segue immediatamente dal fatto che  $R_n$  è un anello noetheriano, e dunque ogni catena ascendente di ideali è stazionaria<sup>5</sup>. Difatti alla catena di chiusi (1.1) si associa la catena di ideali

$$\mathcal{I}(X) \subseteq \mathcal{I}(X_1) \dots \subseteq \mathcal{I}(X_m)$$

---

<sup>5</sup>Vedasi Teor. 2.32.

che è stazionaria. Poichè  $\mathcal{I}(X)$  è iniettiva, anche la catena (1.1) è stazionaria.  $\square$

**Lemma 1.37** *Sia  $X$  uno spazio topologico noetheriano ( per esempio  $X$  un chiuso di  $\mathbb{A}^n$  ). Allora  $X$  ha un insieme finito di componenti irriducibili  $\{X_1, \dots, X_m\}$ . Inoltre  $X$  ammette una decomposizione (unica a meno dell'ordine) come unione delle sue componenti irriducibili*

$$X = X_1 \cup \dots \cup X_m$$

*tale che per ogni  $I \subsetneq \{1, \dots, m\}$  si ha  $X \not\supseteq \bigcup_{i \in I} X_i$ .* <sup>6</sup>

Combinando il Teor. 1.29 e il Lemma 1.37 otteniamo il seguente corollario.

**Corollario 1.38** *Sia  $X \subseteq \mathbb{A}^n(\mathbb{K})$  un chiuso. Allora  $R_n/\mathcal{I}(X)$  è una  $\mathbb{K}$ -algebra ridotta, ossia priva di elementi nilpotenti, e finitamente generata su  $\mathbb{K}$  tale che, se  $X$  è irriducibile allora  $R_n/\mathcal{I}(X)$  è un dominio.*

*Viceversa, ad ogni quoziente  $R_n/I$  privo di nilpotenti resta associato un unico chiuso  $X \subseteq \mathbb{A}^n(\mathbb{K})$  per il quale  $I = \mathcal{I}(X)$  tale che, se  $R_n/I$  è un dominio allora  $X$  è irriducibile.* <sup>7</sup>

---

<sup>6</sup>Dim. Vedasi [C, Lemma 2.3.8].

<sup>7</sup>Dim. Vedasi [C, Cor. 2.3.6].

## 1.4 Dimensione di una varietà

**Definizione 1.39** Data un'algebra  $A$  su un campo  $\mathbb{K}$ , siano  $J$  un insieme di indici,  $a_j \in A$  e  $j \in J$ . Diremo che gli  $a_j \in A$ ,  $j \in J$ , sono algebricamente indipendenti su  $\mathbb{K}$  se non soddisfano alcun polinomio non nullo a coefficienti in  $\mathbb{K}$ . Altrimenti, si dicono algebricamente dipendenti.

**Definizione 1.40** Siano  $\mathbb{K}$  un campo,  $F \supset \mathbb{K}$  una estensione di campi e  $a_j \in F$ ,  $j \in J$ .  $\mathbb{K}(a_j, j \in J)$  è il campo dei quozienti generato dagli  $a_j$  su  $\mathbb{K}$ , i cui elementi si possono scrivere come frazioni di polinomi negli  $a_j$ .

**Definizione 1.41** Data  $F \supset \mathbb{K}$  una estensione di campi e  $a_j \in F$ ,  $j \in J$ . Gli  $a_j$  generano algebricamente  $F$  su  $\mathbb{K}$  se ogni elemento di  $F$  è algebrico sul campo  $\mathbb{K}(a_j, j \in J)$ . Inoltre, se gli  $a_j$  sono algebricamente indipendenti e generano  $F$  su  $\mathbb{K}$ , formano una base di trascendenza di  $F$  su  $\mathbb{K}$ .

### Definizione 1.42 (dimensione di una varietà affine)

Sia  $X \subseteq \mathbb{A}^n$  un chiuso non vuoto. Denotiamo con  $\dim X$  la dimensione di  $X$ , definita nel modo seguente.

- Se  $X$  è irriducibile,  $\dim X$ , è il grado di trascendenza (ovvero il numero di elementi di una base di trascendenza) su  $\mathbb{K}$  del campo dei quozienti di  $R_n/\mathcal{I}(X)$ .
- Sia  $X = X_1 \cup \dots \cup X_m$  la decomposizione in componenti irriducibili di  $X$ , allora

$$\dim X := \max \{ \dim X_i, i = 1, \dots, m \}$$

- Se  $X$  è vuoto poniamo  $\dim X = -\infty$ .

**Osservazione 1.43** Sia  $\mathbb{K}$  algebricamente chiuso. Allora  $\dim \mathbb{A}^n(\mathbb{K}) = n$ .

*Dimostrazione.* Per il Lemma 1.32 lo spazio affine  $\mathbb{A}^n(\mathbb{K})$  è irriducibile. Infatti, per l'esempio 1.19 il suo ideale  $\mathcal{I}(\mathbb{A}^n(\mathbb{K})) = (0)$ , il quale è un ideale primo. Dunque,

$$\dim \mathbb{A}^n(\mathbb{K}) = \dim R_n.$$

La dimensione di  $\mathbb{R}_n$  è uguale al grado di trascendenza su  $\mathbb{K}$  del campo  $\mathbb{K}(x_1, \dots, x_n)$  il quale ha grado di trascendenza  $n$ . Pertanto  $\dim \mathbb{A}^n = n$ .

□

# Capitolo 2

## Le basi di Groebner

Nel Capitolo 1, abbiamo visto come l'algebra degli anelli di polinomi in  $R_n$  sia strettamente connessa con la geometria delle varietà affini. In questo capitolo, studieremo il *metodo delle basi di Groebner*, che ci permetterà di risolvere problemi riguardanti anelli di polinomi in maniera algoritmica. Tale metodo viene utilizzato anche in molti programmi informatici di algebra per studiare particolari ideali polinomiali che si incontrano nelle applicazioni.

Tutto il capitolo 2 è stato scritto con riferimento al libro [CLO] riportato nella bibliografia, ad eccezione del Lemma. 2.55 il quale è un contributo personale.

In questo capitolo daremo le basi necessarie per poter risolvere i seguenti problemi.

- i. *Descrizione dell'ideale*: possiamo trovare un insieme finito di generatori per ogni ideale  $I \subset R_n$ ? O, equivalentemente, possiamo scrivere  $I = \langle f_1, \dots, f_s \rangle$  per qualche  $f_1, \dots, f_s \in R_n$ ?
- ii. *Problema di appartenenza all'ideale*: dati  $f \in R_n$  e  $I = \langle f_1, \dots, f_s \rangle \subset R_n$ , determinare se  $f \in I$ .
- iii. *Risoluzione di equazioni polinomiali*: siano  $f_1, \dots, f_s \in R_n$ . Possiamo determinare se esiste una soluzione comune a tutte le equazioni  $f_i=0$ ? Questo problema equivale a trovare tutti i punti  $p \in V(f_1, \dots, f_s)$ .
- iv. Possiamo determinare se  $V(f_1, \dots, f_n)$  è finita?

v. *Forma Implicita*: siano  $f_1, \dots, f_n$  polinomi nelle variabili  $t_1, \dots, t_m$  e sia  $V$  il sottoinsieme di  $\mathbb{K}^n$  dato tramite la seguente parametrizzazione:

$$\begin{cases} x_1 = f_1(t_1, \dots, t_m) \\ \vdots \\ x_n = f_n(t_1, \dots, t_m) \end{cases}$$

Vogliamo trovare un sistema di equazioni polinomiali nelle variabili  $x_i$ , per  $1 \leq i \leq n$ , che definiscano la varietà.

Innanzitutto, soffermiamoci nel caso  $n = 1$ . Dato un polinomio non nullo  $g$  in  $\mathbb{K}[x]$ , per l'algoritmo della divisione in  $\mathbb{K}[x]$ , ogni polinomio  $f \in \mathbb{K}[x]$  può essere scritto come

$$f = qg + r,$$

dove  $q, r \in \mathbb{K}[x]$  e  $r = 0$  o  $\deg(r) < \deg(g)$ . Inoltre, sia  $f \in \mathbb{K}[x]$  e sia  $I$  un ideale non nullo. Per la Prop. 2.1 esiste  $g \in I$  tale che  $I = \langle g \rangle$ . Allora  $f \in I = \langle g \rangle$  se e solo se  $r = 0$ .

Dunque, se  $n = 1$  i problemi di descrizione di un ideale e di appartenenza all'ideale sono risolti dall'algoritmo della divisione in  $\mathbb{K}[x]$ .

Infatti, per risolvere il problema di descrizione di un ideale basta applicare la seguente proposizione.

**Proposizione 2.1** *Ogni ideale  $I$  di  $\mathbb{K}[x]$  è principale, ossia  $I = \langle g \rangle$  per qualche  $g \in R_n$ . Il polinomio  $g$  è unico a meno di moltiplicazioni per elementi non nulli di  $\mathbb{K}$ . Inoltre, se  $I = \langle f_1, \dots, f_s \rangle$  allora  $I = \langle \text{MCD}(f_1, \dots, f_s) \rangle$ .*

*Dimostrazione.* Sia  $I$  un ideale di  $\mathbb{K}[x]$ . Se  $I = \{0\}$  allora  $I = \langle 0 \rangle$ . Dunque, possiamo supporre  $I$  non nullo. Sia  $g$  un polinomio non nullo di grado minimo appartenente a  $I$ . Vogliamo dimostrare che  $I = \langle g \rangle$ . Sicuramente  $\langle g \rangle \subseteq I$ , dato che  $g \in I$ . Sia ora  $f \in I$ . Per l'algoritmo della divisione in una variabile

$$f = qg + r$$

dove  $r = 0$  oppure  $\deg(r) < \deg(g)$ . Di conseguenza,

$$r = f - qg \in I.$$

Se per assurdo  $r \neq 0$  allora  $\deg(r) < \deg(g)$ , contraddizione. Se ne deduce che  $r = 0$  e  $f = qg \in \langle g \rangle$ . Allora, possiamo concludere che  $I \subseteq \langle g \rangle$  e quindi l'uguaglianza

è verificata. Per dimostrare l'unicità, supponiamo per assurdo che  $\langle f \rangle = \langle g \rangle$  per qualche  $f \in I$ . In particolare osserviamo che  $f \in \langle g \rangle$ , dunque  $f = hg$  per qualche  $h \in R_n$ . Allora

$$\deg(f) = \deg(h) + \deg(g),$$

quindi  $\deg(f) \geq \deg(g)$ . Applicando lo stesso procedimento scambiando l'ordine di  $f$  e  $g$  otteniamo che  $\deg(g) \geq \deg(f)$ . Di conseguenza,  $\deg(f) = \deg(g)$  e quindi  $\deg(h) = 0$ . In conclusione  $h$  è una costante non nulla.

Inoltre se  $I = \langle f_1, \dots, f_s \rangle$ , dato che ogni ideale è principale per quanto appena dimostrato,  $I = \langle f_1, \dots, f_s \rangle = \langle h \rangle$  per qualche  $h \in \mathbb{K}[x]$ . Vogliamo dimostrare che  $h = \text{MCD}(f_1, \dots, f_s)$ . Innanzitutto osserviamo che  $f_1, \dots, f_s \in \langle h \rangle$ , dunque  $h$  divide la  $s$ -upla  $(f_1, \dots, f_s)$ . Inoltre dato che  $h \in \langle f_1, \dots, f_s \rangle$  esistono  $h_1, \dots, h_s \in \mathbb{K}[x]$  tali che

$$h = h_1 f_1 + \dots + h_s f_s. \quad (2.1)$$

Ora, sia  $p \in \mathbb{K}[x]$  tale che  $p$  divide  $f_1, \dots, f_s$ . Quindi esistono  $q_1, \dots, q_s \in \mathbb{K}[x]$  tali che  $f_i = q_i p$  per ogni  $1 \leq i \leq s$ . Dunque la (2.1) diventa

$$h = h_1 q_1 p + \dots + h_s q_s p = p(h_1 q_1 + \dots + h_s q_s).$$

Di conseguenza,  $p$  divide  $h$  e ne segue che  $h = \text{MCD}(f_1, \dots, f_s)$ .

□

**Esempio 2.2** Supponiamo di voler sapere se

$$x^2 - 1 \in \langle x^2 + 3x - 4, x^3 - 5x + 4 \rangle \subset \mathbb{K}[x].$$

Abbiamo osservato nell'esempio 1.13 che

$$\langle x^2 + 3x - 4, x^3 - 5x + 4 \rangle = \langle x - 1 \rangle$$

essendo  $x - 1$  l'MCD in  $\mathbb{K}[x]$ . Dunque, è sufficiente dimostrare  $x^2 - 1 \in \langle x - 1 \rangle$ . Per l'algoritmo della divisione in una variabile,  $x^2 - 1 = (x - 1)(x + 1) + 0$ . Allora, possiamo concludere che  $x^2 - 1 \in \langle x - 1 \rangle$ .

E' ben noto quindi il caso in cui abbiamo una sola variabile. Sorge spontanea una domanda. *Che cosa succede in presenza di più variabili?*

Il nostro obiettivo è quello di estendere tale procedura a sistemi di equazioni polinomiali di qualsiasi grado, in un numero qualsiasi di variabili. In particolare, studieremo un algoritmo che permette di dividere un polinomio  $f \in R_n$  per  $f_1, \dots, f_s \in R_n$ . Lo scopo sarà quello di trovare  $q_1, \dots, q_s, r \in R_n$  tali che

$$f = q_1 f_1 + \dots + q_s f_s + r.$$

*Cosa possiamo dire su  $r$  affinché  $f \in \langle f_1, \dots, f_s \rangle$ ?*

Vedremo che l'algoritmo della divisione per polinomi in  $R_n$  estende l'algoritmo della divisione in  $\mathbb{K}[x]$ .

Per poter proseguire, abbiamo bisogno di introdurre gli *ordinamenti monomiali* che sono di particolare importanza nelle divisioni tra polinomi.

## 2.1 Ordinamenti monomiali in $R_n$

Torniamo a pensare a cosa succede nell'algoritmo della divisione di polinomi in una variabile. Supponiamo di dividere un polinomio  $f$  per  $g$  in  $\mathbb{K}[x]$ . Per prima cosa si cerca il quoziente nella divisione del termine direttivo di  $f$  per il termine direttivo di  $g$  e successivamente si itera su un nuovo dividendo.

*Chi è il termine direttivo nel caso  $n = 1$ ?* Il termine direttivo è semplicemente il termine di grado massimo. Diremo allora che nell'algoritmo della divisione di polinomi in una variabile, lavoriamo con il seguente ordinamento:

$$\dots > x^{m+1} > x^m > \dots > x > 1.$$

In questa sezione daremo alcuni esempi di ordinamenti monomiali che saranno utili in differenti contesti e ci occuperemo di capire quali sono le proprietà che un ordinamento deve soddisfare.

Innanzitutto, possiamo stabilire una corrispondenza biunivoca fra:

$$\{\text{monomi di } R_n\} \longleftrightarrow \{\mathbb{Z}_{\geq 0}^n\}$$

$$x^\alpha = x_1^{\alpha_1} \dots x_n^{\alpha_n} \text{ con } \alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{Z}_{\geq 0}^n.$$

Inoltre, ogni relazione d'ordine  $>$  che stabiliremo su  $\mathbb{Z}_{\geq 0}^n$  ci fornirà un ordinamento per i monomi:

$$\text{se } \alpha > \beta, \text{ diremo che } x^\alpha > x^\beta.$$

Ci sono molti modi differenti di definire relazioni d'ordine in  $\mathbb{Z}_{\geq 0}^n$ . Per i nostri scopi, tuttavia, cercheremo ordinamenti "compatibili" con la struttura algebrica degli anelli dei polinomi.

### Definizione 2.3 (ordinamento monomiale)

*Un ordinamento monomiale su  $R_n$  è una qualsiasi relazione  $>$  su  $\mathbb{Z}_{\geq 0}^n$ , o equivalentemente, una qualsiasi relazione  $>$  sull'insieme dei monomi  $x^\alpha$ , con  $\alpha \in \mathbb{Z}_{\geq 0}^n$ , che soddisfi le seguenti proprietà.*

- i.  $>$  è una relazione d'ordine totale su  $\mathbb{Z}_{\geq 0}^n$ .*
- ii. Se  $\alpha > \beta$  e  $\gamma \in \mathbb{Z}_{\geq 0}^n$  allora  $\alpha + \gamma > \beta + \gamma$ .*
- iii.  $>$  è una buona relazione d'ordine su  $\mathbb{Z}_{\geq 0}^n$ , ovvero ogni sottoinsieme non vuoto di  $\mathbb{Z}_{\geq 0}^n$  ha un elemento minimo rispetto a  $>$ .*

Poichè un polinomio è somma algebrica di monomi, vorremo poter riordinare i termini di un monomio in ordine discendente o ascendente. Per fare ciò, è necessario poter confrontare ciascuna coppia di monomi per stabilire le posizioni corrette. Per garantire tale fatto, abbiamo richiesto che le relazioni d'ordine  $>$  siano *totali*. In questo modo, per ogni coppia di monomi  $x^\alpha$  e  $x^\beta$  una e una sola tra

$$x^\alpha > x^\beta, \quad x^\alpha = x^\beta, \quad x^\alpha < x^\beta$$

è vera.

Inoltre, la proprietà 2.3 *ii.* è equivalente a richiedere che se  $x^\alpha$  e  $x^\beta$  sono due monomi tali che  $x^\alpha > x^\beta$  e  $x^\gamma$  è un monomio qualunque, deve valere che

$$x^\alpha x^\gamma > x^\beta x^\gamma.$$

**Lemma 2.4** *Una relazione d'ordine  $>$  su  $\mathbb{Z}_{\geq 0}^n$  è una buona relazione d'ordine se e solo se ogni sequenza strettamente decrescente in  $\mathbb{Z}_{\geq 0}^n$*

$$\alpha(1) > \alpha(2) > \alpha(3) > \dots$$

*è finita.*

*Dimostrazione.* La tesi è equivalente a dimostrare che  $>$  non è una buona relazione d'ordine se e solo se esiste in  $\mathbb{Z}_{\geq 0}^n$  una sequenza strettamente decrescente infinita. Se  $>$  non è una buona relazione d'ordine, esiste un sottoinsieme non vuoto  $S \subset \mathbb{Z}_{\geq 0}^n$  senza elemento minimo. Sia  $\alpha(1) \in S$ . Siccome  $\alpha(1)$  non è minimo, esiste  $\alpha(2) \in S$  tale che  $\alpha(1) > \alpha(2)$ . Ragionando in modo analogo per  $\alpha(2)$ , esiste  $\alpha(3) \in S$  tale che  $\alpha(1) > \alpha(2) > \alpha(3)$ . Ripetendo questo procedimento, si costruisce una sequenza strettamente decrescente infinita

$$\alpha(1) > \alpha(2) > \alpha(3) > \dots$$

Viceversa, il sottoinsieme non vuoto  $\{\alpha(1), \alpha(2), \alpha(3), \dots\}$  di  $\mathbb{Z}_{\geq 0}^n$  non ammette elemento minimo perché la sequenza strettamente decrescente  $\alpha(1) > \alpha(2) > \dots$  è infinita. Dunque  $>$  non è una buona relazione d'ordine. □

Diamo alcuni esempi di ordini monomiali.

**Definizione 2.5 (ordine lessicografico)**

Siano  $\alpha = (\alpha_1, \dots, \alpha_n)$  e  $\beta = (\beta_1, \dots, \beta_n)$  due vettori di  $\mathbb{Z}_{\geq 0}^n$ . Diremo che  $\alpha >_{lex} \beta$  se, nel vettore differenza  $\alpha - \beta \in \mathbb{Z}^n$ , il primo valore non nullo da sinistra è positivo.

Diremo che  $x^\alpha >_{lex} x^\beta$  se  $\alpha >_{lex} \beta$ .

Osserviamo che l'ordine  $lex$  controlla la variabile più a sinistra (la maggiore secondo  $>_{lex}$ ) e favorisce la potenza maggiore.

**Proposizione 2.6** *L'ordinamento  $>_{lex}$  su  $\mathbb{Z}_{\geq 0}^n$  è un ordinamento monomiale.*

*Dimostrazione.* Sicuramente  $>_{lex}$  è una relazione d'ordine totale poichè l'ordine numerico usuale su  $\mathbb{Z}_{\geq 0}$  è un ordinamento totale.

Se  $\alpha >_{lex} \beta$ , allora il primo valore da sinistra non nullo in  $\alpha - \beta$  è positivo. Sia  $\alpha_k - \beta_k$  tale valore. Consideriamo ora i vettori  $\alpha + \gamma$  e  $\beta + \gamma$  con  $\gamma \in \mathbb{Z}_{\geq 0}^n$ . Osserviamo che il vettore  $(\alpha + \gamma) - (\beta + \gamma)$  coincide con il vettore  $\alpha - \beta$ . Dunque il primo valore non nullo di  $(\alpha + \gamma) - (\beta + \gamma)$  è  $\alpha_k - \beta_k > 0$ . Allora,  $\alpha + \gamma >_{lex} \beta + \gamma$ . Infine, supponiamo per assurdo che  $>_{lex}$  non sia un buon ordinamento. Per il Lemma 2.4 esiste una sequenza strettamente decrescente infinita

$$\alpha(1) >_{lex} \alpha(2) >_{lex} \dots$$

di elementi di  $\mathbb{Z}_{\geq 0}^n$ . Consideriamo i primi valori dei vettori  $\alpha(i) \in \mathbb{Z}_{\geq 0}^n$ . Per come è definito  $>_{lex}$  questi valori formano una sequenza non crescente di interi non negativi. Dato che  $\mathbb{Z}_{\geq 0}^n$  è ben ordinato, i primi valori degli  $\alpha(i)$  devono stabilizzarsi, ossia esiste un  $k$  tale che tutte le prime componenti degli  $\alpha(i)$  con  $i \geq k$  siano uguali. Allora, a partire da  $\alpha(k)$ , per determinare l'ordine  $lex$  bisogna considerare i secondi valori e successivi. I secondi valori di  $\alpha(k), \alpha(k+1), \dots$  formano una sequenza non crescente e, dunque, si devono anch'essi stabilizzare per lo stesso ragionamento fatto per i primi valori. Ripetendo tale procedimento per i valori successivi, osserviamo che, per qualche  $l$ , gli  $\alpha(l), \alpha(l+1), \dots$  sono tutti uguali, il che contraddice  $\alpha(l) >_{lex} \alpha(l+1)$ . Pertanto,  $>_{lex}$  è una buona relazione d'ordine su  $\mathbb{Z}_{\geq 0}^n$  e quindi per la Def. 2.3  $>_{lex}$  è un ordinamento monomiale su  $\mathbb{Z}_{\geq 0}^n$ .  $\square$

Consideriamo le variabili  $x_1, \dots, x_n$ . Per la Def. 2.5 si ha

$$(1, 0, \dots, 0) >_{lex} (0, 1, \dots, 0) >_{lex} \dots >_{lex} (0, \dots, 0, 1).$$

O, equivalentemente,  $x_1 >_{lex} x_2 >_{lex} \dots >_{lex} x_n$ .

In particolare, quando lavoriamo con polinomi in due o tre variabili, è solito chiamare le variabili  $x, y, z$  invece di  $x_1, x_2, x_3$  rispettivamente. Se non è specificato nel testo, supporremo che l'ordine alfabetico sia utilizzato per definire l'ordine lessicografico, dunque

$$x >_{lex} y >_{lex} z.$$

**Definizione 2.7 (ordine lessicografico graduato)**

Siano  $\alpha = (\alpha_1, \dots, \alpha_n)$  e  $\beta = (\beta_1, \dots, \beta_n)$  due vettori di  $\mathbb{Z}_{\geq 0}^n$ .

Diremo che  $\alpha >_{grlex} \beta$  se

$$\sum_{i=1}^n \alpha_i = |\alpha| > |\beta| = \sum_{i=1}^n \beta_i \text{ oppure } |\alpha| = |\beta| \text{ e } \alpha >_{lex} \beta.$$

Inoltre,  $x^\alpha >_{grlex} x^\beta$  se  $\alpha >_{grlex} \beta$ .

Notiamo che l'ordine  $>_{grlex}$  ordina prima secondo il grado complessivo, successivamente ordina i monomi di grado uguale utilizzando l'ordine  $lex$ .

**Proposizione 2.8** *L'ordinamento  $grlex$  su  $\mathbb{Z}_{\geq 0}^n$  è un ordinamento monomiale.*

La dimostrazione è molto simile alla dimostrazione della Prop. 2.6, pertanto sarà omessa.

**Esempio 2.9** Il polinomio  $f = x^3y^2 - x^2y^3 + x \in \mathbb{R}[x, y]$  è ordinato secondo l'ordine  $grlex$ .

Infatti, consideriamo i vettori  $\alpha = (3, 2)$ ,  $\beta = (2, 3)$  e  $\gamma = (1, 0)$ . Osserviamo che  $|\alpha| = 5 = |\beta|$  e  $|\gamma| = 1$ , quindi  $\alpha >_{grlex} \gamma$  e  $\beta >_{grlex} \gamma$ . Per capire chi è il "maggiore" fra  $\alpha$  e  $\beta$ , a parità di grado totale, dobbiamo usare l'ordine  $lex$ . Dunque, consideriamo il vettore  $\alpha - \beta = (1, -1)$ . Essendo il primo valore non nullo da sinistra positivo allora  $\alpha >_{lex} \beta$ , e di conseguenza,  $\alpha >_{grlex} \beta >_{grlex} \gamma$ . Se ne deduce che  $x^3y^2 >_{grlex} x^2y^3 >_{grlex} x$ .

Come nel caso di una sola variabile, anche nella divisione tra polinomi in  $R_n$  è importante definire i termini principali.

**Definizione 2.10** *Sia  $f = \sum_{\alpha} a_{\alpha} x^{\alpha}$  un polinomio non nullo di  $R_n$  e sia  $>$  un ordinamento monomiale.*

*i. Il multigrado di  $f$  è*

$$\text{multideg}(f) = \max(\alpha \in \mathbb{Z}_{\geq 0}^n \mid a_{\alpha} \neq 0).$$

*dove il massimo è calcolato rispetto all'ordine  $>$ .*

ii. Il coefficiente direttivo di  $f$  è

$$\text{LC}(f) = a_{\text{multideg}(f)} \in \mathbb{K}.$$

iii. Il monomio direttivo di  $f$  è

$$\text{LM}(f) = x^{\text{multideg}(f)}.$$

iv. Il termine direttivo di  $f$  è

$$\text{LT}(f) = \text{LC}(f) \cdot \text{LM}(f).$$

**Esempio 2.11** Sia  $f = 3x^2y + xy^2 + y^2 \in \mathbb{R}[x, y]$  ordinato secondo l'ordine  $lex$ . Allora  $\text{multideg}(f) = (2, 1)$ ,  $\text{LC}(f) = 3$ ,  $\text{LM}(f) = x^2y$  e  $\text{LT}(f) = 3x^2y$ .

**Esempio 2.12** Il polinomio  $f = 2xy^2 + x^2 \in \mathbb{R}[x, y]$  è ordinato secondo l'ordine  $grlex$ . Infatti, consideriamo i vettori  $\alpha = (1, 2)$ ,  $\beta = (2, 0)$ . Osserviamo che  $|\alpha| = 3$  e  $|\beta| = 2$ , quindi  $\alpha >_{grlex} \beta$ .

Allora  $\text{multideg}(f) = (1, 2)$ ,  $\text{LC}(f) = 2$ ,  $\text{LM}(f) = xy^2$  e  $\text{LT}(f) = 2xy^2$ . Osserviamo inoltre che il polinomio  $2xy^2 + x^2$  non è ordinato secondo l'ordine  $lex$ . Infatti il vettore  $\alpha - \beta = (-1, 2)$  ha il primo valore da sinistra negativo.

**Lemma 2.13** Siano  $f, g \in R_n$  polinomi non nulli. Allora:

i.  $\text{multideg}(fg) = \text{multideg}(f) + \text{multideg}(g)$ .

ii. Se  $f + g \neq 0$ , allora  $\text{multideg}(f + g) \leq \max(\text{multideg}(f), \text{multideg}(g))$ .

Inoltre, se  $\text{multideg}(f) \neq \text{multideg}(g)$ , vale l'uguaglianza.

*Dimostrazione.* Siano  $f = \sum_{\alpha} a_{\alpha} x^{\alpha}$  e  $g = \sum_{\beta} b_{\beta} x^{\beta}$  polinomi non nulli di  $R_n$ . Osserviamo che

$$f = a_{\text{multideg}(f)} x^{\text{multideg}(f)} + \sum_{\alpha \neq \text{multideg}(f)} a_{\alpha} x^{\alpha}$$

e

$$g = b_{\text{multideg}(g)} x^{\text{multideg}(g)} + \sum_{\beta \neq \text{multideg}(g)} b_{\beta} x^{\beta}.$$

i. Per come è definito il prodotto fra polinomi, sicuramente

$$fg = a_{\text{multideg}(f)}b_{\text{multideg}(g)}x^{\text{multideg}(f)+\text{multideg}(g)} + \\ + \sum_{(\alpha,\beta) \neq (\text{multideg}(f), \text{multideg}(g))} a_{\alpha}b_{\beta}x^{\alpha+\beta}.$$

Allora  $\text{multideg}(fg) = \text{multideg}(f) + \text{multideg}(g)$ .

ii. Se  $f + g \neq 0$ ,

$$f + g = a_{\text{multideg}(f)}x^{\text{multideg}(f)} + b_{\text{multideg}(g)}x^{\text{multideg}(g)} + \\ + \sum_{\alpha \neq \text{multideg}(f)} a_{\alpha}x^{\alpha} + \sum_{\beta \neq \text{multideg}(g)} b_{\beta}x^{\beta}.$$

Se  $\text{multideg}(f) = \text{multideg}(g)$ , allora o  $\text{LC}(f) + \text{LC}(g) \neq 0$  dunque

$$\text{multideg}(f + g) = \text{multideg}(f) = \text{multideg}(g),$$

oppure  $\text{LC}(f) + \text{LC}(g) = 0$ , allora

$$\text{multideg}(f + g) \leq \text{multideg}(f) = \text{multideg}(g).$$

Infine, se  $\text{multideg}(f) \neq \text{multideg}(g)$ ,

$$\text{multideg}(f + g) = \max(\text{multideg}(f), \text{multideg}(g)).$$

Dunque,

$$\text{multideg}(f + g) \leq \max(\text{multideg}(f), \text{multideg}(g)).$$

□

## 2.2 Algoritmo della divisione in $R_n$

Lo scopo di questa sezione è di risolvere il problema di *appartenenza all'ideale*: dati  $f \in R_n$  e  $I = \langle f_1, \dots, f_s \rangle \subset R_n$ , determinare se  $f \in I$ .

Per risolverlo, l'idea è simile al caso di una variabile ed è una conseguenza del seguente teorema.

**Teorema 2.14** *Sia  $>$  un ordine monomiale su  $\mathbb{Z}_{\geq 0}^n$  fissato, e sia  $(f_1, \dots, f_s)$  una  $s$ -upla ordinata di polinomi in  $R_n$ . Allora ogni polinomio  $f \in R_n$  può essere scritto come*

$$f = q_1 f_1 + \dots + q_s f_s + r,$$

dove  $q_1, \dots, q_s, r \in R_n$  e  $r = 0$  oppure  $r$  è una combinazione  $\mathbb{K}$ -lineare di monomi, nessuno dei quali è divisibile per  $i \text{ LT}(f_1), \dots, \text{LT}(f_s)$ . Diremo che  $r$  è il resto di  $f$  diviso per  $(f_1, \dots, f_s)$ .

Inoltre, se  $q_i f_i \neq 0$  vale  $\text{multideg}(f) \geq \text{multideg}(q_i f_i)$ .

*Dimostrazione.* Dimostriamo l'esistenza di  $q_1, \dots, q_s, r \in R_n$  fornendo un algoritmo per la costruzione e mostrando che esso opera correttamente. L'algoritmo è il seguente.

**Input :**  $f_1; \dots; f_s$

**Output :**  $q_1; \dots; q_s; r$

$q_1 := 0; \dots; q_s := 0; r := 0$

$p := f$

WHILE  $p \neq 0$  DO

$i := 1$

divisione := falso

WHILE  $i \leq s$  AND divisione = falso DO

IF  $\text{LT}(f_i)$  divide  $\text{LT}(p)$  THEN

$q_i := q_i + \frac{\text{LT}(p)}{\text{LT}(f_i)}$

$p := p - \frac{\text{LT}(p)}{\text{LT}(f_i)} f_i$

divisione := vero

ELSE  $i = i + 1$

IF divisione = falso THEN

$r := r + \text{LT}(p)$

$p = p - \text{LT}(p)$

Osserviamo che ogni volta che viene eseguito il ciclo **WHILE** accade una ed una sola tra queste due possibilità.

- (*Fase di Divisione*) Se qualcuno degli  $LT(f_i)$  divide  $LT(p)$ , allora l'algoritmo procede come nel caso ad una variabile.
- (*Fase di Resto*) Se nessuno degli  $LT(f_i)$  divide  $LT(p)$ , allora l'algoritmo aggiunge  $LT(p)$  al resto.

Mostriamo che

$$f = q_1 f_1 + \dots + q_s f_s + p + r \quad (2.2)$$

è vera ad ogni passaggio dell'algoritmo. Inoltre, dal momento che l'algoritmo termina quando  $p = 0$ , dall'equazione (2.2) segue che esistono  $q_1, \dots, q_s, r \in R_n$  tali che

$$f = q_1 f_1 + \dots + q_s f_s + r.$$

Sicuramente la (2.2) è vera per i valori iniziali dell'algoritmo, avendo inizializzato tutte le variabili a 0 e definito  $p$  come  $p := f$ . Supponiamo che (2.2) sia vera ad un passaggio generico dell'algoritmo. Se il passaggio successivo è una *Fase di Divisione* qualcuno degli  $LT(f_i)$  divide  $LT(p)$  e vengono modificati sia i  $q_i$  sia  $p$ . Notiamo però che

$$q_i f_i + p = \left( q_i + \frac{LT(p)}{LT(f_i)} \right) f_i + p - \frac{LT(p)}{LT(f_i)} f_i$$

dunque,  $q_i f_i + p$  non è cambiato. Poichè nessuna delle altre variabili viene modificata, l'equazione (2.2) rimane vera nella *Fase di Divisione*.

Ora, supponiamo che il passaggio successivo sia una *Fase di Resto*. Le variabili che vengono modificate sono  $p$  e  $r$  ma la loro somma rimane uguale. Infatti

$$p + r = r + LT(p) + p - LT(p).$$

Allora, l'equazione (2.2) è verificata anche in questa fase.

Inoltre, dobbiamo dimostrare che l'algoritmo termina in un numero finito di passi. Questo è vero dal momento in cui ogni volta che ridefiniamo la variabile  $p$ , o  $p$  si annulla o il suo multigrado diminuisce. Essendo ciò ovvio nella *Fase di Resto*, supponiamo di trovarci nella *Fase di Divisione* e di ridefinire la variabile  $p$  secondo l'algoritmo.

$$p' := p - \frac{LT(p)}{LT(f_i)} f_i.$$

Per il Lemma 2.13,

$$\text{LT}\left(\frac{\text{LT}(p)}{\text{LT}(f_i)}f_i\right) = \frac{\text{LT}(p)}{\text{LT}(f_i)}\text{LT}(f_i) = \text{LT}(p). \quad (2.3)$$

Di conseguenza, se  $p' \neq 0$  si ha che

$$\text{multideg}(p) > \text{multideg}(p')$$

secondo l'ordinamento monomiale  $>$  fissato. Se per assurdo l'algoritmo non terminasse, si verrebbe a creare una catena strettamente decrescente infinita di multi-gradi, contraddicendo il Lemma 2.4. Quindi necessariamente l'algoritmo termina in un numero finito di passi, in particolare termina quando troviamo un polinomio  $p$  tale che nessuno degli  $\text{LT}(f_i)$  divide ciascun monomio di  $p$ . Di conseguenza aggiungiamo tutti i monomi di  $p$  al resto e otteniamo un nuovo polinomio  $p$  nullo. Allora, l'algoritmo termina quando  $p = 0$ . Poichè vengono aggiunti termini a  $r$  solo quando non sono divisibili per nessuno dei  $\text{LT}(f_i)$ , ne consegue che  $q_1, \dots, q_s$  ed  $r$  hanno le proprietà richieste nel momento in cui l'algoritmo termina.

Infine, resta da dimostrare la relazione che lega  $\text{multideg}(f)$  e  $\text{multideg}(q_i f_i)$ . L'algoritmo inizia quando  $p := f$  e, per quanto detto al paragrafo precedente,  $\text{multideg}(p)$  decresce. Di conseguenza,  $\text{multideg}(p) \leq \text{multideg}(f)$ , o equivalentemente,  $\text{LT}(p) \leq \text{LT}(f)$ . Ogni monomio di  $q_i$  è della forma  $\frac{\text{LT}(p)}{\text{LT}(f_i)}$  quindi dall'equazione (2.3) segue che per ogni monomio di  $q_i$  vale la seguente relazione

$$\text{LT}\left(\frac{\text{LT}(p)}{\text{LT}(f_i)}f_i\right) \leq \text{LT}(f).$$

Dalla Def. 2.3 segue che  $\text{multideg}(f) \geq \text{multideg}(q_i f_i)$ . □

**Corollario 2.15 (condizione sufficiente<sup>1</sup> per l'appartenenza all'ideale)**

*Dati  $f \in R_n$  e  $I = \langle f_1, \dots, f_s \rangle \subset R_n$ . Se nella divisione di  $f$  per la  $s$ -upla ordinata di polinomi  $(f_1, \dots, f_s)$  si ha che  $r = 0$ , allora  $f \in \langle f_1, \dots, f_s \rangle$ .*

*Dimostrazione.* Per il Teor. 2.14, esistono  $q_1, \dots, q_s, r \in R_n$  tali che

$$f = q_1 f_1 + \dots + q_s f_s + r.$$

Se  $r = 0$  allora

$$f = q_1 f_1 + \dots + q_s f_s,$$

di conseguenza  $f \in \langle f_1, \dots, f_s \rangle$ . □

---

<sup>1</sup>Vedasi Oss. 2.18.

Cercheremo di capire con degli esempi le proprietà del resto  $r$  e come funziona la divisione fra polinomi in più variabili e più divisori.

Innanzitutto, proprio come in  $\mathbb{K}[x]$ , la regola base dell'algoritmo è: fissato un ordine monomiale, dobbiamo elidere il termine direttivo di  $f$ , moltiplicando  $f_i$  per un appropriato monomio e poi sottrarre. Questo monomio diventa un termine del corrispondente  $g_i$ .

Ciò che distingue il caso di una divisione fra polinomi di più variabili rispetto ad una divisione tra polinomi di una variabile, è il comportamento del resto della divisione. Se pensiamo infatti alla divisione tra polinomi in una variabile, una volta che il termine direttivo del divisore non divide più il termine direttivo di ciò che resta da dividere, l'algoritmo termina. Nel caso di più variabili invece, non è detto che l'algoritmo termina in quel momento.

Vediamo un esempio che ci permette di capire tale fatto.

**Esempio 2.16** Sia  $f = x^2y + xy^2 + y^2 \in \mathbb{R}[x, y]$  ordinato secondo l'ordine *lex*. Supponiamo di dover dividere  $f$  per due polinomi  $f_1 = xy - 1$  e  $f_2 = y^2 - 1$ . Allora per la Def. 2.10,  $LT(f) = x^2y$ ,  $LT(f_1) = xy$  e  $LT(f_2) = y^2$ . Osserviamo che  $LT(f_1)$  divide  $LT(f)$ , dunque iniziamo a dividere  $f$  per  $f_1$ . Rappresentiamo la divisione graficamente, ma a differenza della divisione fra polinomi in una variabile, ora disponiamo i diversi divisori e quozienti in verticale.

$$\begin{array}{r}
 q_1 : x + y \\
 q_2 : \\
 \\
 \begin{array}{r}
 xy - 1 \qquad x^2y + xy^2 + y^2 \\
 y^2 - 1 \qquad x^2y - x \\
 \hline
 \qquad \qquad \qquad xy^2 + x + y^2 \\
 \qquad \qquad \qquad xy^2 - y \\
 \hline
 \qquad \qquad \qquad x + y^2 + y
 \end{array}
 \end{array}$$

Dividendo  $LT(f) = x^2y$  per  $LT(f_1) = xy$  otteniamo il primo monomio  $x$  che fa parte del quoziente  $q_1$ . Proprio come nella divisione fra polinomi in una variabile, moltiplichiamo ora  $x$  per  $f_1$  e lo sottraiamo a  $f$ . Il risultato è  $xy^2 + x + y^2$ , che ha termine direttivo  $LT(xy^2 + x + y^2) = xy^2$  ancora divisibile per  $LT(f_1)$ . Dunque, ripetiamo lo stesso procedimento dividendo  $xy^2 + x + y^2$  per  $f_1$ . Al quoziente  $q_1$  va aggiunto allora il monomio  $y$ , ottenuto dividendo  $LT(xy^2 + x + y^2) = xy^2$  per  $LT(f_1) = xy$ . Ora, sottraendo  $y \cdot f_1$  da  $xy^2 + x + y^2$  otteniamo  $x + y^2 + y$ . Ora, osserviamo che né  $LT(f_1) = xy$  né  $LT(f_2) = y^2$  dividono  $LT(x + y^2 + y) = x$ ,

tuttavia il resto della divisione non può essere  $x + y^2 + y$ . Infatti,  $\text{LT}(f_2) = y^2$  divide un monomio di  $x + y^2 + y$  e per il Teor. 2.14 il polinomio  $x + y^2 + y$  non può essere un resto. Notiamo che il polinomio  $y^2 + y$  ottenuto da  $x + y^2 + y$  eliminando la  $x$  ha coefficiente direttivo divisibile per  $\text{LT}(f_2) = y^2$ . Dunque aggiungiamo la  $x$  al resto, in modo tale da poter continuare a dividere per  $f_2$ .

$$\begin{aligned} q_1 &: x + y \\ q_2 &: 1 \end{aligned}$$

$$\begin{array}{r} xy - 1 \qquad x^2y + xy^2 + y^2 \qquad r \\ y^2 - 1 \qquad x^2y - x \\ \hline \qquad \qquad \qquad xy^2 + x + y^2 \\ \qquad \qquad \qquad xy^2 - y \\ \hline \qquad \qquad \qquad x + y^2 + y \\ \hline \qquad \qquad \qquad y^2 + y \qquad \longrightarrow \qquad x \end{array}$$

Possiamo dire che  $q_1 = x + y$  e procedendo con la divisione tra  $y^2 + y$  e  $y^2 - 1$  otteniamo

$$\begin{aligned} q_1 &: x + y \\ q_2 &: 1 \end{aligned}$$

$$\begin{array}{r} xy - 1 \qquad x^2y + xy^2 + y^2 \qquad r \\ y^2 - 1 \qquad x^2y - x \\ \hline \qquad \qquad \qquad xy^2 + x + y^2 \\ \qquad \qquad \qquad xy^2 - y \\ \hline \qquad \qquad \qquad x + y^2 + y \\ \hline \qquad \qquad \qquad y^2 + y \qquad \longrightarrow \qquad x \\ \qquad \qquad \qquad y^2 - 1 \\ \hline \qquad \qquad \qquad y + 1 \\ \hline \qquad \qquad \qquad 1 \qquad \longrightarrow \qquad x + y \\ \hline \qquad \qquad \qquad 0 \qquad \longrightarrow \qquad x + y + 1 \end{array}$$

dove  $q_2 = 1$ , e al monomio  $x$  del resto dobbiamo aggiungere  $y + 1$ . Allora

$$x^2y + xy^2 + y^2 = (x + y)(xy - 1) + 1(y^2 - 1) + (x + y + 1).$$

Osserviamo che il resto è somma di monomi, nessuno dei quali è divisibile per i termini direttivi di  $f_1$  e  $f_2$ .

Nell'algoritmo della divisione in  $\mathbb{K}[x]$  il resto è univocamente determinato. Cosa possiamo dire per  $r \in R_n$ ?

**Osservazione 2.17** Siano  $q_1, \dots, q_s, r \in R_n$  tali che  $f = q_1f_1 + \dots + q_sf_s + r$ . Il resto  $r$  non è univocamente determinato.

*Dimostrazione.* Utilizzeremo un controesempio per dimostrare tale fatto.

Sia  $f = x^2y + xy^2 + y^2 \in \mathbb{R}[x, y]$  ordinato secondo l'ordine *lex*. Dividiamo  $f$  per due polinomi  $f_1 = y^2 - 1$  e  $f_2 = xy - 1$ .

Notiamo che i dati sono equivalenti all'esempio 2.16, scambiando l'ordine dei divisori  $f_1$  e  $f_2$ .

$$\begin{aligned} q_1 &: x + 1 \\ q_2 &: x \end{aligned}$$

$$\begin{array}{r} \begin{array}{r} y^2 - 1 \\ xy - 1 \end{array} \quad \begin{array}{r} x^2y + xy^2 + y^2 \\ x^2y - x \end{array} \quad r \\ \hline \quad \quad \quad \begin{array}{r} xy^2 + x + y^2 \\ xy^2 - x \end{array} \\ \hline \quad \quad \quad \quad \quad \begin{array}{r} 2x + y^2 \end{array} \\ \hline \quad \quad \quad \quad \quad \quad \begin{array}{r} y^2 \quad \longrightarrow \quad 2x \\ y^2 - 1 \end{array} \\ \hline \quad \quad \quad \quad \quad \quad \quad \quad \begin{array}{r} 1 \end{array} \\ \hline \quad \quad \quad \quad \quad \quad \quad \quad \quad \begin{array}{r} 0 \quad \longrightarrow \quad 2x + 1 \end{array} \end{array}$$

Questo mostra che

$$x^2y + xy^2 + y^2 = (x + 1)(y^2 - 1) + x(xy - 1) + (2x + 1).$$

Osserviamo che se dividiamo  $f$  per  $f_1 = y^2 - 1$  e  $f_2 = xy - 1$  il resto è  $r = 2x + 1$ ; scambiando l'ordine di  $f_1$  e  $f_2$ , come nell'esempio 2.16, il resto è  $r = x + y + 1$ . In conclusione, nel caso di una divisione fra polinomi in più variabili, il resto  $r$  non è univocamente determinato. Infatti, gli  $q_i$  e  $r$  possono cambiare riordinando gli  $f_i$ .

□

Abbiamo osservato nel Cor. 2.15 che se  $r = 0$  allora  $f \in \langle f_1, \dots, f_s \rangle$ . Ma cosa possiamo dire su  $r$  se  $f \in \langle f_1, \dots, f_s \rangle$ ?

**Osservazione 2.18** La condizione definita nel Cor. 2.15 è una condizione sufficiente ma non necessaria.

*Dimostrazione.* Per dimostrare che non è una condizione necessaria utilizzeremo un controesempio.

Siano  $f = x^2y + xy + x + 1$ ,  $f_1 = xy + 1$  e  $f_2 = x^2 + 2 \in \mathbb{K}[x, y]$  con ordine *lex*. Se dividiamo  $f$  per  $(f_1, f_2)$  otteniamo

$$x^2y + xy + x + 1 = (x + 1)(xy + 1) + 0 \cdot (x^2 + 1) + 0$$

quindi  $f \in \langle f_1, f_2 \rangle$ . Scambiando l'ordine dei divisori, ossia dividendo  $f$  per  $(f_2, f_1)$  otteniamo

$$x^2y + xy + x + 1 = y(x^2 + 2) + 1 \cdot (xy + 1) + (x - 2y).$$

In particolare osserviamo che  $f \in \langle f_1, f_2 \rangle$  ma non è detto che  $r = 0$ ; infatti, dividendo  $f$  per  $(f_2, f_1)$  otteniamo  $r = x - 2y$ .

□

Dato un ideale  $I = \langle f_1, \dots, f_s \rangle \subset R_n$ , d'ora in poi consideriamo il polinomio  $f_i$  quello nella posizione  $i$ -esima.

In conclusione, abbiamo riscontrato due problemi dell'algoritmo della divisione in  $R_n$ :

- il resto  $r$  non è univocamente determinato;
- la condizione  $r = 0$  non è una condizione necessaria per l'appartenenza all'ideale.

Nella Sezione §2.5 definiremo un metodo per cercare polinomi  $g_1, \dots, g_s \in R_n$  che generano particolari basi, dette basi di Groebner, grazie alle quali questi due problemi saranno risolti.

## 2.3 Ideali monomiali e lemma di Dickson

Nella sezione §1.2 abbiamo introdotto gli ideali. Adesso studieremo particolari ideali, detti ideali monomiali.

### Definizione 2.19 (ideale monomiale)

Siano  $I \subset R_n$  un ideale non nullo e  $A \subseteq \mathbb{Z}_{\geq 0}^n$ . Diremo che  $I$  è un ideale monomiale se è formato da tutti i polinomi che sono somme finite nella forma  $f = \sum_{\alpha \in A} h_\alpha x^\alpha$ , dove al più un numero finito di  $h_\alpha \in R_n$  sono non nulli. In questo caso scriveremo  $I = \langle x^\alpha, \alpha \in A \rangle$ .

Per gli ideali monomiali il problema dell'appartenenza è molto semplice. Infatti basta applicare il seguente lemma.

**Lemma 2.20** *Sia  $I = \langle x^\alpha, \alpha \in A \rangle$  un ideale monomiale. Allora un monomio  $x^\beta$  appartiene a  $I$  se e solo se  $x^\beta$  è divisibile per  $x^\alpha$  per qualche  $\alpha \in A$ .*

*Dimostrazione.* Sicuramente se  $x^\beta$  è divisibile per  $x^\alpha$  per qualche  $\alpha \in A$ , allora  $x^\beta$  è un multiplo di  $x^\alpha$  per qualche  $\alpha \in A$ . Dunque,  $x^\beta \in I$ .

Viceversa, se  $x^\beta \in I$ , allora  $x^\beta = \sum_{i=1}^s h_i x^{\alpha(i)}$  dove  $h_i \in R_n$  e  $\alpha(i) \in A$ . Se riscriviamo ogni  $h_i$  come combinazione lineare di monomi notiamo che i termini non nulli al secondo membro dell'equazione sono della forma  $c_{ij} x^{\gamma(i,j)} x^{\alpha(i)}$  con  $\gamma(i,j) \in A$  e  $c_{ij} \neq 0$ . Dunque, esistono  $i$  e  $j$  tali che  $x^\beta = c_{ij} x^{\gamma(i,j)} x^{\alpha(i)}$ . Allora  $x^{\alpha(i)}$  divide  $x^\beta$ , per qualche  $\alpha(i)$ . □

**Lemma 2.21** *Siano  $I$  un ideale monomiale e sia  $f \in R_n$ . Allora le seguenti proprietà sono equivalenti.*

- i.  $f \in I$ .
- ii. Ogni termine di  $f$  appartiene ad  $I$ .
- iii.  $f$  è una combinazione  $\mathbb{K}$ -lineare di monomi di  $I$ .

*Dimostrazione.*  $I$  è un ideale monomiale, o equivalentemente,  $I = \langle x^\alpha, \alpha \in A \rangle$ .

Osserviamo che  $i. \Rightarrow ii.$  dato che se  $f \in I$  allora  $f = \sum_{i=1}^s h_i x^{\alpha(i)}$ . In particolare, ciascun termine di  $f$  è un multiplo di qualche  $x^{\alpha(i)}$ , dunque appartiene ad  $I$ .

Inoltre  $ii. \Rightarrow iii.$  poichè se ogni termine di  $f$  appartiene ad  $I$ , anche i corrispondenti monomi sono in  $I$ , ed  $f$  è somma di quei termini, ossia  $f$  è una combinazione  $\mathbb{K}$ -lineare di monomi di  $I$ .

Infine,  $iii. \Rightarrow i.$  poichè, essendo  $I$  un ideale, ciascuna combinazione di elementi di  $I$  appartiene a  $I$ . □

**Corollario 2.22** *Due ideali monomiali coincidono se e solo se contengono gli stessi monomi.*

*Dimostrazione.* Siano  $I_1$  e  $I_2$  due ideali monomiali, ossia  $I_1 = \langle x^\alpha, \alpha \in A \rangle$  e  $I_2 = \langle x^\beta, \beta \in B \rangle$ , con  $A, B \subseteq \mathbb{Z}_{\geq 0}^n$ .

Sicuramente se  $I_1$  e  $I_2$  coincidono contengono gli stessi monomi. Viceversa, per la proprietà *iii.* del Lemma 2.21 gli ideali monomiali sono determinati in modo univoco dai propri monomi, dunque se  $I_1$  e  $I_2$  contengono gli stessi monomi, allora  $I_1 = I_2$ . □

Il seguente teorema è un risultato fondamentale perché dimostra che tutti gli ideali monomiali di  $R_n$  sono finitamente generati.

**Teorema 2.23 (lemma di Dickson)**

*Un ideale monomiale  $I = \langle x^\alpha, \alpha \in A \rangle \subseteq R_n$  può essere scritto nella forma  $\langle x^{\alpha(1)}, \dots, x^{\alpha(s)} \rangle$ , dove  $\alpha(1), \dots, \alpha(s) \in A$ . In particolare,  $I$  ha una base finita.*

*Dimostrazione.* Dimostriamo per induzione sul numero delle variabili  $n$ .

Sia  $n = 1$ , allora  $I$  è generato dai monomi  $x^\alpha$ , dove  $\alpha \in \mathbb{Z}_{\geq 0}$ . Sia  $\beta$  il più piccolo elemento di  $A \subseteq \mathbb{Z}_{\geq 0}$ ; allora  $x^\beta$  divide tutti gli altri generatori. Dunque  $I = \langle x^\beta \rangle$ .

Supponiamo ora  $n > 1$  e che il teorema sia vero per  $n - 1$ . Scriveremo le variabili come  $x_1, \dots, x_{n-1}, y$  in modo tale da poter scrivere i polinomi di  $\mathbb{K}[x_1, \dots, x_{n-1}, y]$  nella forma  $x^\alpha y^m$ , dove  $\alpha = (\alpha_1, \dots, \alpha_{n-1}) \in \mathbb{Z}_{\geq 0}^{n-1}$  e  $m \in \mathbb{Z}_{\geq 0}$ .

Supponiamo che  $I \subset \mathbb{K}[x_1, \dots, x_{n-1}, y]$  sia un ideale monomiale. Definiamo

$$J := \langle x^\alpha \text{ tale che } x^\alpha y^m \in I \text{ per qualche } m \geq 0 \rangle \subseteq R_{n-1}.$$

Poiché  $J$  è un ideale monomiale in  $R_{n-1}$ , vale l'ipotesi induttiva. Di conseguenza,

$$J = \langle x^{\alpha(1)}, \dots, x^{\alpha(s)} \rangle \subseteq R_{n-1}$$

e lo immaginiamo come la proiezione di  $I$  su  $R_{n-1}$ .

Per ogni  $1 \leq i \leq s$ ,  $x^{\alpha(i)} y^{m_i} \in I$  per qualche  $m_i \geq 0$ . Sia  $m$  il massimo degli  $m_i$ , allora per ogni  $0 \leq k \leq m - 1$  consideriamo l'ideale  $J_k \subset R_{n-1}$  generato dai monomi  $x^\beta$  tali che  $x^\beta y^k \in I$ . Utilizzando nuovamente l'ipotesi di induzione,  $J_k$  ha un insieme finito di generatori, dunque poniamo  $J_k = \langle x^{\alpha_k(1)}, \dots, x^{\alpha_k(s_k)} \rangle$ .

Vogliamo mostrare che  $I$  è generato dai monomi del seguente elenco:

- da  $J : x^{\alpha(1)} y^m, \dots, x^{\alpha(s)} y^m$ ,
- da  $J_0 : x^{\alpha_0(1)}, \dots, x^{\alpha_0(s_0)}$ ,
- ⋮
- da  $J_{m-1} : x^{\alpha_{m-1}(1)} y^{m-1}, \dots, x^{\alpha_{m-1}(s_{m-1})} y^{m-1}$ .

Prima di tutto osserviamo che  $J_k \subseteq I$  per ogni  $k$ . Viceversa ogni monomio in  $I$  è divisibile per un elemento dell'elenco. Infatti, sia  $x^\alpha y^p \in I$ . Se  $p \geq m$ ,  $x^\alpha \in J$  per definizione di  $J$ . Per il Lemma 2.20  $x^\alpha$  è divisibile da qualche  $x^{\alpha^{(i)}}$  e quindi  $x^\alpha y^p$  è divisibile da  $x^{\alpha^{(i)}} y^m$ . Al contrario, se  $p \leq m - 1$ , dal Lemma 2.20 segue che  $x^\alpha y^p$  è divisibile per qualche  $x^{\alpha^{(j)}} y^p$  per costruzione di  $J_p$ . Per il Lemma 2.20, i monomi dell'elenco generano un ideale avente gli stessi monomi di  $I$ ; di conseguenza, per il Cor. 2.22 gli ideali coincidono.

Per completare la dimostrazione del teorema, dobbiamo dimostrare che l'insieme finito dei generatori può essere scelto da un insieme di generatori per l'ideale. Tornando a indicare le variabili con  $x_1, \dots, x_n$  il nostro ideale monomiale è  $I = \langle x^\alpha, \alpha \in A \rangle \subseteq R_n$ . Dobbiamo dimostrare che  $I$  è generato da un numero finito di  $x^\alpha$ , dove  $\alpha \in A$ . Per il paragrafo precedente, sappiamo che  $I = \langle x^{\beta^{(1)}}, \dots, x^{\beta^{(s)}} \rangle$  per certi monomi  $x^{\beta^{(i)}} \in I$ . Poichè  $x^{\beta^{(i)}} \in I = \langle x^\alpha, \alpha \in A \rangle$ , per il Lemma 2.20 per ogni  $\beta^{(i)}$  esiste almeno un  $\alpha^{(i)} \in A$  tale che  $x^{\beta^{(i)}}$  sia divisibile per  $x^{\alpha^{(i)}}$ . In conclusione  $I = \langle x^{\alpha^{(1)}}, \dots, x^{\alpha^{(s)}} \rangle$ . □

Nel Lemma 2.55 dimostreremo che la base trovata nella dimostrazione del lemma di Dickson è una base di Groebner ridotta.

Dal lemma di Dickson segue la prossima osservazione che permette di risolvere il problema di appartenenza all'ideale per gli ideali monomiali.

**Osservazione 2.24** Sia  $I = \langle x^{\alpha^{(1)}}, \dots, x^{\alpha^{(s)}} \rangle$ , con  $\alpha^{(1)}, \dots, \alpha^{(s)} \in \mathbb{Z}_{\geq 0}^n$  e sia  $f \in R_n$ . Allora  $f \in I$  se e solo se il resto della divisione di  $f$  per  $(x^{\alpha^{(1)}}, \dots, x^{\alpha^{(s)}})$  è  $r = 0$ .

*Dimostrazione.* Per il Cor. 2.15, se il resto della divisione di  $f$  per  $(x^{\alpha^{(1)}}, \dots, x^{\alpha^{(s)}})$  è  $r = 0$ , allora  $f \in I = \langle x^{\alpha^{(1)}}, \dots, x^{\alpha^{(s)}} \rangle$ .

Viceversa, sia  $f \in I$ . Per il Teor. 2.14

$$f = q_1 x^{\alpha^{(1)}} + \dots + q_s x^{\alpha^{(s)}} + r,$$

dove  $q_1, \dots, q_s, r \in R_n$  e  $r = 0$  oppure  $r$  è una combinazione  $\mathbb{K}$ -lineare di monomi, nessuno dei quali è divisibile per  $\text{LT}(x^{\alpha^{(1)}}), \dots, \text{LT}(x^{\alpha^{(s)}})$ . Osserviamo che  $\text{LT}(x^{\alpha^{(i)}}) = x^{\alpha^{(i)}}$  per ogni  $1 \leq i \leq s$ . Supponiamo per assurdo che  $r \neq 0$ . In particolare

$$r = f - q_1 x^{\alpha^{(1)}} - \dots - q_s x^{\alpha^{(s)}} \in I = \langle x^{\alpha^{(1)}}, \dots, x^{\alpha^{(s)}} \rangle.$$

Dunque, per il Lemma 2.20,  $r$  è divisibile per  $x^{\alpha^{(i)}}$ , per qualche  $1 \leq i \leq s$ , contraddizione. Se ne deduce che il resto della divisione di  $f$  per  $(x^{\alpha^{(1)}}, \dots, x^{\alpha^{(s)}})$  è  $r = 0$ . □

## 2.4 Teorema della base di Hilbert e basi di Groebner

Per gli ideali monomiali abbiamo risolto sia il *problema di descrizione di un ideale* sia il *problema di appartenenza all'ideale*. Che cosa succede per gli ideali non monomiali di  $R_n$ ?

Quando abbiamo introdotto il capitolo 2 ci siamo domandati se potevamo trovare un insieme finito di generatori per *ogni* ideale  $I \subset R_n$ . Sicuramente il lemma di Dickson garantisce una risposta positiva per gli ideali monomiali. In questa sezione ci occuperemo di estendere questo fatto ad *ogni* ideale di  $R_n$ .

Dato un polinomio  $f \in R_n$  non nullo, nella Def. 2.10 abbiamo definito  $\text{LT}(f)$ . Allora per ogni ideale  $I \subset R_n$  possiamo definire il seguente ideale.

**Definizione 2.25** *Sia  $I \subset R_n$  un ideale non nullo.*

*i. Denotiamo con  $\text{LT}(I)$  l'insieme dei termini direttivi degli elementi in  $I$ , ovvero*

$$\text{LT}(I) = \{cx^\alpha \mid \exists f \in I \text{ con } \text{LT}(f) = cx^\alpha\}.$$

*ii. Denotiamo con  $\langle \text{LT}(I) \rangle$  l'ideale generato dagli elementi di  $\text{LT}(I)$ .*

**Proposizione 2.26** *Sia  $I = \langle f_1, \dots, f_s \rangle$ . Allora*

$$\langle \text{LT}(f_1), \dots, \text{LT}(f_s) \rangle \subseteq \langle \text{LT}(I) \rangle$$

*sebbene l'uguaglianza non sia vera in generale.*

*Dimostrazione.* Sia  $f_i \in I$ . Sicuramente,  $\text{LT}(f_i) \in \text{LT}(I) \subseteq \langle \text{LT}(I) \rangle$ , per ogni  $1 \leq i \leq s$ . In particolare  $\langle \text{LT}(f_1), \dots, \text{LT}(f_s) \rangle \subseteq \langle \text{LT}(I) \rangle$ .

Useremo un controesempio per dimostrare che, in generale, non è sempre vera l'uguaglianza.

Consideriamo l'ideale  $I = \langle xy^2 - y + 1, x^2y - x \rangle$ , con ordine *lex* sui monomi in  $\mathbb{R}[x, y]$ . Poniamo  $f_1 = xy^2 - y + 1$  e  $f_2 = x^2y - x$ . Sicuramente  $x \in I$  perché

$$x = x(xy^2 - y + 1) - y(x^2y - x)$$

dunque  $x \in \text{LT}(I) \subseteq \langle \text{LT}(I) \rangle$ . Ma,  $x \notin \langle xy^2, x^2y \rangle = \langle \text{LT}(f_1), \text{LT}(f_2) \rangle$  per il Lemma 2.20. Di conseguenza,  $\langle \text{LT}(f_1), \text{LT}(f_2) \rangle \neq \langle \text{LT}(I) \rangle$ . □

La seguente proposizione garantisce che  $\langle \text{LT}(I) \rangle$  è un ideale monomiale. Applicando il lemma di Dickson (si veda Teor. 2.23) possiamo concludere che  $\langle \text{LT}(I) \rangle$  ammette una base finita.

**Proposizione 2.27** *Sia  $I \subset R_n$  un ideale. Allora:*

*i.  $\langle \text{LT}(I) \rangle$  è un ideale monomiale.*

*ii. Esistono  $g_1, \dots, g_t \in I$  tali che  $\langle \text{LT}(I) \rangle = \langle \text{LT}(g_1), \dots, \text{LT}(g_t) \rangle$ .*

*Dimostrazione.* *i.* I monomi direttivi  $\text{LM}(g)$  degli elementi  $g \in I \setminus \{0\}$  generano l'ideale monomiale  $\langle \text{LM}(g), g \in I \setminus \{0\} \rangle$ . Poichè  $\text{LM}(g)$  e  $\text{LT}(g)$  sono uno multiplo dell'altro per una costante non nulla,  $\langle \text{LM}(g), g \in I \setminus \{0\} \rangle = \langle \text{LT}(g), g \in I \setminus \{0\} \rangle$ . Se ne deduce che  $\langle \text{LT}(I) \rangle$  è un ideale monomiale.

*ii.* Poichè  $\langle \text{LT}(I) \rangle$  è generato dai monomi  $\text{LM}(g)$  per  $g \in I \setminus \{0\}$ , il lemma di Dickson (vedasi Teor. 2.23) ci dice che  $\langle \text{LT}(I) \rangle = \langle \text{LM}(g_1), \dots, \text{LM}(g_t) \rangle$  per un numero finito di  $g_1, \dots, g_t \in I$ . Poichè  $\text{LM}(g_i)$  e  $\text{LT}(g_i)$  sono multipli per una costante non nulla, ne segue che  $\langle \text{LT}(I) \rangle = \langle \text{LM}(g_1), \dots, \text{LM}(g_t) \rangle = \langle \text{LT}(g_1), \dots, \text{LT}(g_t) \rangle$ .  $\square$

Nel primo capitolo della tesi abbiamo accennato il fatto che l'anello dei polinomi  $R_n$  sia un *anello noetheriano*<sup>2</sup>. Tale fatto è garantito dal seguente teorema.

**Teorema 2.28 (teorema della base di Hilbert)**

*Ogni ideale  $I \subset R_n$  ha un insieme finito di generatori, ovvero*

$$I = \langle g_1, \dots, g_t \rangle$$

*per certi  $g_1, \dots, g_t \in I$ .*

*Dimostrazione.* Se  $I = \{0\}$ , allora  $I$  è finitamente generato dal polinomio nullo. Ora, sia  $I$  non nullo. Per la Prop. 2.27 esistono  $g_1, \dots, g_t \in I$  tali che

$$\langle \text{LT}(I) \rangle = \langle \text{LT}(g_1), \dots, \text{LT}(g_t) \rangle.$$

Dimostreremo che  $I = \langle g_1, \dots, g_t \rangle$ , facendo vedere che valgono le due inclusioni. Sicuramente  $\langle g_1, \dots, g_t \rangle \subseteq I$  per ogni  $g_i \in I$ . Viceversa, Sia  $f \in I$ . Dividendo  $f$  per  $g_1, \dots, g_t \in I$ , per il Teor. 2.14 esistono  $q_1, \dots, q_t, r \in R_n$  tali che:

$$f = q_1 g_1 + \dots + q_t g_t + r \tag{2.4}$$

---

<sup>2</sup>Si veda Def. 1.8.

dove nessun termine di  $r$  è divisibile per  $\text{LT}(g_1), \dots, \text{LT}(g_t)$ .  
Dall'equazione (2.4) segue che

$$r = f - q_1g_1 - \dots - q_tg_t \in I.$$

Se per assurdo  $r \neq 0$ , allora  $\text{LT}(r) \in \langle \text{LT}(I) \rangle = \langle \text{LT}(g_1), \dots, \text{LT}(g_t) \rangle$  e per il Lemma 2.20 si ha che  $\text{LT}(r)$  deve essere divisibile per  $\text{LT}(g_i)$ , per qualche  $1 \leq i \leq t$ , contraddizione. Allora,  $r = 0$  e per il Cor. 2.15 si ha che  $f \in \langle g_1, \dots, g_t \rangle$ . Dunque,  $I \subseteq \langle g_1, \dots, g_t \rangle$  e, di conseguenza, vale la tesi.  $\square$

La base utilizzata nella dimostrazione del Teor. 2.28 ha una particolare proprietà:

$$\langle \text{LT}(I) \rangle = \langle \text{LT}(g_1), \dots, \text{LT}(g_t) \rangle.$$

Nella Prop. 2.26 abbiamo fatto vedere che l'uguaglianza non è vera in generale. Allora, daremo un nome specifico a tutte le basi  $\{g_1, \dots, g_t\} \subset I$  per le quali l'uguaglianza è verificata.

**Definizione 2.29 (base di Groebner)**

*Fissato un ordine monomiale, un sottoinsieme finito*

$G = \{g_1, \dots, g_t\}$ , con  $g_i \neq 0$  per  $1 \leq i \leq t$ , di un ideale  $I$  è detto base di Groebner se

$$\langle \text{LT}(I) \rangle = \langle \text{LT}(g_1), \dots, \text{LT}(g_t) \rangle.$$

Per il Lemma 2.20 la definizione di base di Groebner è equivalente alla seguente definizione.

**Definizione 2.30** *Un insieme  $\{g_1, \dots, g_t\}$  di  $I$ , con  $g_i \neq 0$  per  $1 \leq i \leq t$ , è una base di Groebner per  $I$  se e solo se il termine direttivo di ogni elemento di  $I$  è divisibile per uno degli  $\text{LT}(g_i)$ .*

**Lemma 2.31** *Fissato un ordine monomiale, ogni ideale  $I \subset R_n$ , non nullo, ammette una base di Groebner. Inoltre, ogni base di Groebner di un ideale  $I$  è una base di  $I$ .*

*Dimostrazione.* Dato un ideale  $I$  non nullo, l'insieme  $G = \{g_1, \dots, g_t\} \subset R_n$  costruito nella dimostrazione del Teor. 2.28 è una base di Groebner per definizione. Inoltre, sia  $\{g_1, \dots, g_t\}$  una base di Groebner di  $I$ . Allora  $\langle g_1, \dots, g_t \rangle \subseteq I$ . Utilizzando l'algoritmo di divisione, come nella dimostrazione del Teor. 2.28, si deduce che  $I \subseteq \langle g_1, \dots, g_t \rangle$ . Dunque,  $I = \langle g_1, \dots, g_t \rangle$ , pertanto  $G$  è una base di  $I$ .  $\square$

Concludiamo questa sezione con un'applicazione algebrica del teorema della base di Hilbert (vedasi Teor. 2.28) riguardante gli ideali di  $R_n$ .

**Teorema 2.32 (condizione della catena ascendente)**

*Sia*

$$I_1 \subseteq I_2 \subseteq I_3 \subseteq \dots$$

*una catena ascendente di ideali in  $R_n$ . Allora la catena si stabilizza, ovvero esiste un  $N \geq 1$  tale che*

$$I_N = I_{N+1} = I_{N+2} = \dots$$

*Dimostrazione.* Data la catena ascendente  $I_1 \subseteq I_2 \subseteq I_3 \subseteq \dots$ , consideriamo l'insieme  $I = \bigcup_{i \geq 1} I_i$ . Innanzitutto,  $I$  è un ideale in  $R_n$ . Infatti,  $0 \in I$  dato che  $0 \in I_i$  per ogni  $i$ . Inoltre se  $f, g \in I$  allora  $f \in I_i$  e  $g \in I_j$ , per certi indici  $i, j$ . Poichè gli ideali  $I_i$  formano una catena ascendente, se riordiniamo gli indici in modo tale che  $i \leq j$ , allora  $f, g \in I_j$ . Poichè  $I_j$  è un ideale allora  $f + g \in I_j$  dunque  $f + g \in I$ . Analogamente, se  $f \in I$  e  $h \in R_n$ ,  $f \in I_i$  per qualche  $i$ , pertanto  $hf \in I_i$ . Di conseguenza  $hf \in I$ . Allora  $I$  è un ideale.

Per il Teor. 2.28 esistono  $f_1, \dots, f_s \in I$  tali che  $I = \langle f_1, \dots, f_s \rangle$ . Ciascun generatore  $f_i \in I_{j_i}$  per certi  $j_i$ , con  $i = 1, \dots, s$ . Sia  $N$  il massimo dei  $j_i$ : per la definizione di catena ascendente  $f_i \in I_N$  per ogni  $i$ . Se ne deduce che

$$I = \langle f_1, \dots, f_s \rangle \subseteq I_N \subseteq I_{N+1} \subseteq \dots \subseteq I.$$

Di conseguenza, la catena ascendente si stabilizza con  $I_N$  e tutti gli ideali successivi nella catena risultano uguali.

□

## 2.5 Proprietà delle basi di Groebner e criterio di Buchberger

Dato un ideale  $I \subset R_n$ , a volte è facile verificare che una base  $F = \{f_1, \dots, f_s\}$  non è di Groebner per  $I$ . Infatti, come nella dimostrazione della Prop. 2.26, basta trovare un polinomio  $p \in I$ , non nullo, tale che  $\text{LT}(p) \in \langle \text{LT}(I) \rangle$  ma  $\text{LT}(p)$  non è divisibile per nessuno degli  $\text{LT}(f_i)$ . Ma, come possiamo dimostrare che una data base di polinomi  $\{g_1, \dots, g_s\} \subset R_n$  è una base di Groebner?

In questa sezione daremo delle proprietà di tali basi necessarie per poter rispondere a questa domanda.

Abbiamo accennato che le basi di Groebner sono una “buona” base: i problemi riscontrati nell’algoritmo della divisione fra polinomi in più variabili vengono risolti. In particolare, il resto  $r$  ottenuto dalla divisione di  $f$  per una qualsiasi base di Groebner  $G = \{g_1, \dots, g_t\}$  è indipendente dall’ordine in cui gli elementi di  $G$  vengono considerati nell’esecuzione dell’algoritmo della divisione. Inoltre, come per gli ideali monomiali <sup>3</sup>, data una base di Groebner, la condizione  $r = 0$  è una condizione necessaria di appartenenza all’ideale. Questi due fatti sono una conseguenza della seguente proposizione.

**Proposizione 2.33** *Sia  $G = \{g_1, \dots, g_t\}$  una base di Groebner per un’ideale  $I$  in  $R_n$  non nullo e sia  $f \in R_n$ . Allora esiste un unico resto  $r \in R_n$  con le seguenti proprietà.*

*i. Nessun termine di  $r$  è divisibile per alcuno dei  $\text{LT}(g_1), \dots, \text{LT}(g_t)$ .*

*ii. Esiste  $g \in I$  tale che  $f = g + r$ .*

*Dimostrazione.* Le due proprietà sono garantite dal Teor. 2.14, dove nella proprietà *ii.* abbiamo posto  $g = q_1g_1 + \dots + q_tg_t \in I$ .

Per dimostrare l’unicità, supponiamo per assurdo che esistano  $r_1, r_2 \in R_n$  che soddisfino le due proprietà *i.* e *ii.* Allora  $g_1 + r_1 = f = g_2 + r_2$  o, equivalentemente,  $r_2 - r_1 = g_1 - g_2 \in I$ . Se  $r_2 \neq r_1$ ,  $\text{LT}(r_2 - r_1) \in \langle \text{LT}(I) \rangle = \langle \text{LT}(g_1), \dots, \text{LT}(g_t) \rangle$  essendo  $G$  una base di Groebner. Dal Lemma 2.20 segue che  $\text{LT}(r_2 - r_1)$  deve essere divisibile per qualche  $\text{LT}(g_i)$ , contraddicendo la proprietà *i.* Di conseguenza,  $r_2 = r_1$ . □

---

<sup>3</sup>Vedasi Oss. 2.24.

**Corollario 2.34 (condizione necessaria per l'appartenenza all'ideale)**

Sia  $G = \{g_1, \dots, g_t\}$  una base di Groebner per un ideale  $I \subset R_n$  non nullo e sia  $f \in R_n$ . Allora  $f \in I$  se e solo se il resto della divisione di  $f$  per  $G$  è zero.

*Dimostrazione.* Se il resto è zero, allora per il Cor. 2.15,  $f \in I$ . Viceversa, dato  $f \in I$  abbiamo che  $f = f + 0$  soddisfa le due condizioni della Prop. 2.33. Ne segue che  $r = 0$  è il resto della divisione di  $f$  per  $G$ . □

**Definizione 2.35** Indicheremo con  $\bar{f}^F$  il resto della divisione di  $f$  per la  $s$ -upla ordinata  $F = (f_1, \dots, f_s)$ .

**Osservazione 2.36** Sia  $F = \{f_1, \dots, f_s\}$  con  $f_j$  non nullo per ogni  $1 \leq j \leq s$ . Per ogni  $\alpha \in \mathbb{Z}_{\geq 0}^n$  si ha che  $\overline{x^\alpha f_i}^F = 0$  per ogni  $i$  tale che  $1 \leq i \leq s$ .

*Dimostrazione.* Per calcolare il resto della divisione di  $x^\alpha f_i$  per  $F$  applichiamo l'algoritmo descritto nella dimostrazione del Teor. 2.14. Osserviamo che  $LT(f_i)$  divide  $LT(x^\alpha f_i)$  e quindi  $p = 0$  e  $r = 0$ . Si noti che i quozienti della divisione sono tutti nulli ad eccezione del quoziente  $q_i = x^\alpha$ . □

La base  $F = \{xy^2 - y + 1, x^2y - x\}$  usata nella dimostrazione della Prop. 2.26 non è una base di Groebner proprio per quanto visto nella medesima proposizione. Conosciamo ora un altro metodo per dimostrare tale fatto. Per il Cor. 2.34 basta dimostrare che prendendo un qualsiasi polinomio  $f \in I = \langle xy^2 - y + 1, x^2y - x \rangle$ , il resto della divisione di  $f$  per  $F$ ,  $\bar{f}^F$ , non è zero.

Sia  $f = x$ . Allora

$$x = x(xy^2 - y + 1) - y(x^2y - x) \in I$$

ma,  $\bar{f}^F = x$ . Per il Cor. 2.34 possiamo concludere che  $F = \{xy^2 - y + 1, x^2y - x\}$  non è una base di Groebner.

Nel Teor. 2.28 abbiamo dimostrato che l'anello dei polinomi  $R_n$  è noetheriano, dunque ogni ideale  $I \subset R_n$  è finitamente generato. In particolare, ogni ideale ammette una base di Groebner per il Lemma 2.31. Dunque, per dimostrare che un polinomio  $f \in I$ , dobbiamo calcolare la base di Groebner che genera l'ideale, e per il Cor. 2.34 verificare che  $\bar{f}^G = 0$ .

Nella Def. 2.29 abbiamo definito una base di Groebner come un sottoinsieme finito  $G = \{g_1, \dots, g_t\}$  di un ideale  $I$ , non nullo, tale che

$$\langle LT(I) \rangle = \langle LT(g_1), \dots, LT(g_t) \rangle.$$

In particolare osserviamo che una data base  $F = \{f_1, \dots, f_s\}$  non è di Groebner se esiste qualche combinazione polinomiale delle  $f_i$  i cui termini direttivi non appartengono all'ideale  $\langle \text{LT}(I) \rangle$ . Possiamo quindi trovare un metodo che permette di elidere i termini direttivi e fare in modo che rimangono termini più piccoli, secondo l'ordine monomiale stabilito. Definiamo allora un polinomio ottenuto dalla combinazione lineare di elementi della base che permette di cancellare i termini direttivi.

**Definizione 2.37** ( $\mathcal{S}$  – polinomio)

Siano  $f, g \in R_n$  polinomi non nulli.

- i. Se  $\text{multideg}(f) = \alpha$  e  $\text{multideg}(g) = \beta$ , allora sia  $\gamma = (\gamma_1, \dots, \gamma_n)$  dove  $\gamma_i = \max(\alpha_i, \beta_i)$  per  $1 \leq i \leq n$ . Chiamiamo  $x^\gamma$  il minimo comune multiplo di  $\text{LM}(f)$  e  $\text{LM}(g)$  e lo indichiamo  $x^\gamma = \text{mcm}(\text{LM}(f), \text{LM}(g))$ .
- ii. L'  $\mathcal{S}$ -polinomio di  $f$  e  $g$  è la combinazione

$$\mathcal{S}(f, g) = \frac{x^\gamma}{\text{LT}(f)} f - \frac{x^\gamma}{\text{LT}(g)} g.$$

**Esempio 2.38** Siano  $f = 4x^2z - 7y^2$  e  $g = xyz^2 + 3xz^4$  in  $\mathbb{R}[x, y, z]$ . Allora  $\alpha = (2, 0, 1)$  e  $\beta = (1, 1, 2)$ , dunque  $\gamma = (2, 1, 2)$ . In particolare

$$\mathcal{S}(f, g) = \frac{x^2yz^2}{4x^2z}(4x^2z - 7y^2) - \frac{x^2yz^2}{xyz^2}(xyz^2 + 3xz^4) = -\frac{7}{4}y^3z - 3x^2z^4.$$

**Proposizione 2.39** Siano  $\alpha(1), \dots, \alpha(s) \in \mathbb{Z}_{\geq 0}^n$  e siano  $f_1, \dots, f_s \in R_n$ . Allora

$$\mathcal{S}(x^{\alpha(j)} f_j, x^{\alpha(k)} f_k) = x^{\delta - \gamma_{jk}} \mathcal{S}(f_j, f_k)$$

con  $x^\delta = \text{mcm}(x^{\alpha(j)} \text{LM}(f_j), x^{\alpha(k)} \text{LM}(f_k))$  e  $x^{\gamma_{jk}} = \text{mcm}(\text{LM}(f_j), \text{LM}(f_k))$ .

*Dimostrazione.*

$$\begin{aligned} \mathcal{S}(x^{\alpha(j)} f_j, x^{\alpha(k)} f_k) &= \frac{x^\delta}{x^{\alpha(j)} \text{LT}(f_j)} x^{\alpha(j)} f_j - \frac{x^\delta}{x^{\alpha(k)} \text{LT}(f_k)} x^{\alpha(k)} f_k = \\ &= \frac{x^\delta}{\text{LT}(f_j)} f_j - \frac{x^\delta}{\text{LT}(f_k)} f_k = \frac{x^\delta}{x^{\gamma_{jk}}} \frac{x^{\gamma_{jk}}}{\text{LT}(f_j)} f_j - \frac{x^\delta}{x^{\gamma_{jk}}} \frac{x^{\gamma_{jk}}}{\text{LT}(f_k)} f_k = \\ &= \frac{x^\delta}{x^{\gamma_{jk}}} \left( \frac{x^{\gamma_{jk}}}{\text{LT}(f_j)} f_j - \frac{x^{\gamma_{jk}}}{\text{LT}(f_k)} f_k \right) = x^{\delta - \gamma_{jk}} \mathcal{S}(f_j, f_k). \end{aligned}$$

□

**Lemma 2.40** *Siano  $f_1, \dots, f_s \in R_n$  non nulli e sia  $\text{multideg}(f_i) = \delta \in \mathbb{Z}_{\geq 0}^n$  per ogni  $1 \leq i \leq s$ . Supponiamo di avere una somma, non nulla, nella forma  $\sum_{i=1}^s c_i f_i$  dove i  $c_i$  sono costanti in  $\mathbb{K}$ . Se  $\text{multideg}(\sum_{i=1}^s c_i f_i) \preceq \delta$ , allora esistono costanti  $c_{jk}$  in  $\mathbb{K}$  con  $1 \leq j, k \leq s$  tali che*

$$\sum_{i=1}^s c_i f_i = \sum_{j,k} c_{jk} \mathcal{S}(f_j, f_k).$$

*Inoltre, ogni  $\mathcal{S}(f_j, f_k)$  non nullo ha multigrado strettamente minore di  $\delta$ .*

*Dimostrazione.* Siano  $f_1, \dots, f_s \in R_n$  non nulli. Per  $1 \leq i \leq s$  sia  $d_i = \text{LC}(f_i)$ , dunque  $c_i d_i = \text{LC}(c_i f_i)$ . Sia  $\sum_{i=1}^s c_i f_i$  una somma non nulla, dove i  $c_i$  sono costanti in  $\mathbb{K}$ . Per ipotesi  $\text{multideg}(f_i) = \delta$  e  $\text{multideg}(\sum_{i=1}^s c_i f_i) \preceq \delta$  allora

$$\sum_{i=1}^s c_i d_i = 0. \quad (2.5)$$

Definiamo  $p_i = \frac{f_i}{d_i}$  e osserviamo che  $\text{LC}(p_i) = 1$ . Consideriamo la somma telescopica

$$\begin{aligned} \sum_{i=1}^s c_i f_i &= \sum_{i=1}^s c_i d_i p_i = c_1 d_1 (p_1 - p_2) + (c_1 d_1 + c_2 d_2) (p_2 - p_3) + \dots + \\ &+ (c_1 d_1 + \dots + c_{s-1} d_{s-1}) (p_{s-1} - p_s) + (c_1 d_1 + \dots + c_s d_s) p_s. \end{aligned} \quad (2.6)$$

Dalla (2.5) segue che la (2.6) è equivalente a

$$\begin{aligned} \sum_{i=1}^s c_i f_i &= c_1 d_1 (p_1 - p_2) + (c_1 d_1 + c_2 d_2) (p_2 - p_3) + \dots + \\ &+ \dots + (c_1 d_1 + \dots + c_{s-1} d_{s-1}) (p_{s-1} - p_s). \end{aligned} \quad (2.7)$$

Ora,  $\text{LT}(f_i) = d_i x^\delta$  per ogni  $1 \leq i \leq s$  dal momento che  $\text{multideg}(f_i) = \delta \in \mathbb{Z}_{\geq 0}^n$  per ogni  $1 \leq i \leq s$ . Se ne deduce che  $x^\delta = \text{mcm}(\text{LM}(f_j), \text{LM}(f_k))$  e vale la seguente uguaglianza.

$$\mathcal{S}(f_j, f_k) = \frac{x^\delta}{\text{LT}(f_j)} f_j - \frac{x^\delta}{\text{LT}(f_k)} f_k = \frac{x^\delta}{d_j x^\delta} f_j - \frac{x^\delta}{d_k x^\delta} f_k = p_j - p_k. \quad (2.8)$$

Dalla (2.8) otteniamo che la (2.7) diventa

$$\begin{aligned} \sum_{i=1}^s c_i f_i &= c_1 d_1 \mathcal{S}(f_1, f_2) + (c_1 d_1 + c_2 d_2) \mathcal{S}(f_2, f_3) + \dots + \\ &+ \dots + (c_1 d_1 + \dots + c_{s-1} d_{s-1}) \mathcal{S}(f_{s-1}, f_s). \end{aligned}$$

Infine, dato che  $\text{LT}(p_i) = \text{LT}(\frac{f_i}{d_i}) = x^\delta$ , il  $\text{multideg}(p_j - p_k)$ , o per la (2.8), equivalentemente il  $\text{multideg}(\mathcal{S}(f_j, f_k))$ , è strettamente minore di  $\delta$ . □

**Teorema 2.41 (criterio di Buchberger)**

Sia  $I$  un ideale non nullo in  $R_n$ . Allora una sua base  $G = \{g_1, \dots, g_t\}$  è una base di Groebner per  $I$  se e solo se il resto di  $\mathcal{S}(g_i, g_j)$  diviso  $G$  (elencata in un qualsiasi ordine) è zero, per ogni coppia  $1 \leq i \neq j \leq t$ .

*Dimostrazione.* Se  $G$  è una base di Groebner, siccome  $\mathcal{S}(g_i, g_j) \in I$ , per il Cor. 2.34, il resto della divisione di  $\mathcal{S}(g_i, g_j)$  per  $G$  è zero.

Viceversa, sia  $G = \{g_1, \dots, g_t\}$  una base per  $I$  tale che il resto di  $\mathcal{S}(g_i, g_j)$  diviso  $G$  è zero, per ogni coppia  $1 \leq i \neq j \leq t$ . Ora, dobbiamo dimostrare che

$$\langle \text{LT}(I) \rangle = \langle \text{LT}(g_1), \dots, \text{LT}(g_t) \rangle.$$

Un'inclusione segue dal fatto che  $g_i \in I$ . Per il viceversa, sia  $f \in I = \langle g_1, \dots, g_t \rangle$  non nullo. Dimosteremo che  $\text{LT}(f)$  appartiene a  $\langle \text{LT}(g_1), \dots, \text{LT}(g_t) \rangle$ .

Dato  $f \in I$ , esistono polinomi  $h_i \in R_n$  tali che

$$f = \sum_{i=1}^t h_i g_i. \quad (2.9)$$

Dal Lemma 2.13, segue

$$\text{multideg}(f) \leq \max_{1 \leq i \leq t} (\text{multideg}(h_i g_i)). \quad (2.10)$$

Sia  $m(i) = \text{multideg}(h_i g_i)$  e definiamo  $\delta = \max(m(1), \dots, m(t))$ . Allora la (2.10) diventa

$$\text{multideg}(f) \leq \delta.$$

Per ottenere la tesi basta dimostrare che  $\text{multideg}(f) = \delta$ , in modo tale che

$$\text{multideg}(f) = \text{multideg}(h_i g_i)$$

per qualche  $i$  tale che  $1 \leq i \leq t$ . Infatti dal Lemma 2.13 segue

$$x^{\text{multideg}(f)} = x^{\text{multideg}(h_i g_i)} = x^{\text{multideg}(h_i) + \text{multideg}(g_i)} = x^{\text{multideg}(h_i)} x^{\text{multideg}(g_i)}.$$

Dunque  $\text{LM}(f) = \text{LM}(h_i) \text{LM}(g_i)$  e, dato che  $\text{LM}(f)$  è un multiplo di  $\text{LT}(f)$  per una costante non nulla, allora  $\text{LT}(f)$  è divisibile per  $\text{LT}(g_i)$ . Allora  $\text{LT}(f) \in \langle \text{LT}(g_1), \dots, \text{LT}(g_t) \rangle$ .

Ora dimostriamo che si può avere  $\text{multideg}(f) = \delta$ . Consideriamo tutti i possibili modi in cui  $f$  può essere scritta nella forma (2.9). Per ognuna di queste espressioni otteniamo un  $\delta$ , eventualmente differente. Siccome un ordinamento monomiale è un buon ordinamento (si veda Def. 2.3), possiamo scegliere un'espressione (2.9)

tale che  $\delta$  sia minimo. Supponiamo per assurdo  $\text{multideg}(f) \not\leq \delta$ . Riscriviamo  $f$  nella seguente forma:

$$f = \sum_{m(i)=\delta} h_i g_i + \sum_{m(i) \leq \delta} h_i g_i = \sum_{m(i)=\delta} \text{LT}(h_i) g_i + \sum_{m(i)=\delta} (h_i - \text{LT}(h_i)) g_i + \sum_{m(i) \leq \delta} h_i g_i. \quad (2.11)$$

Siccome le ultime due somme hanno tutti i monomi con multigrado strettamente minore di  $\delta$ , affinché  $\text{multideg}(f) \leq \delta$  deve valere che

$$\text{multideg}\left(\sum_{m(i)=\delta} \text{LT}(h_i) g_i\right) \leq \delta.$$

Possiamo riscrivere la somma  $\sum_{m(i)=\delta} \text{LT}(h_i) g_i$  nel modo seguente:

$$\sum_{m(i)=\delta} \text{LT}(h_i) g_i = \sum_{m(i)=\delta} c_i x^{\alpha(i)} g_i$$

dove  $\text{LT}(h_i) = c_i x^{\alpha(i)}$ . Mostriamo ora che  $\sum_{m(i)=\delta} \text{LT}(h_i) g_i$  è combinazione lineare a coefficienti in  $\mathbb{K}$  degli  $\mathcal{S}(f_j, f_k)$ . Se la somma è zero allora è ovvio. Se non è zero segue dal Lemma 2.40. Infatti, ponendo  $f_i = x^{\alpha(i)} g_i$ , dal Lemma 2.40 e dalla Prop. 2.39 esistono costanti non nulle  $c_{jk}$  in  $\mathbb{K}$  tali che

$$\sum_{m(i)=\delta} \text{LT}(h_i) g_i = \sum_{j,k} c_{jk} x^{\delta - \gamma_{jk}} \mathcal{S}(g_j, g_k) \quad (2.12)$$

dove  $x^\delta = \text{mcm}(x^{\alpha(j)} \text{LM}(g_j), x^{\alpha(k)} \text{LM}(g_k))$  e  $x^{\gamma_{jk}} = \text{mcm}(\text{LM}(g_j), \text{LM}(g_k))$ . Inoltre sempre dal Lemma 2.40 si ha  $\text{multideg}(x^{\delta - \gamma_{jk}} \mathcal{S}(g_j, g_k)) \leq \delta$ .

Per ipotesi sappiamo che il resto di  $\mathcal{S}(g_i, g_k)$  diviso  $(g_1, \dots, g_t)$  è zero. In altre parole, utilizzando il Teor. 2.14,

$$\mathcal{S}(g_j, g_k) = \sum_{i=1}^t q_{ijk} g_i \quad (2.13)$$

dove  $q_{ijk} \in R_n$  e, per ogni  $i, j, k$ , vale

$$\text{multideg}(q_{ijk} g_i) \leq \text{multideg}(\mathcal{S}(g_j, g_k)). \quad (2.14)$$

In altre parole abbiamo trovato un'espressione per  $\mathcal{S}(g_j, g_k)$  come combinazione lineare di  $G$  dove non tutti i termini direttivi si elidono. Per sfruttare questo, moltiplichiamo l'equazione (2.13) per  $x^{\delta - \gamma_{jk}}$  e otteniamo

$$x^{\delta - \gamma_{jk}} \mathcal{S}(g_j, g_k) = \sum_{i=1}^t b_{ijk} g_i, \quad (2.15)$$

dove  $b_{ijk} = q_{ijk}x^{\delta-\gamma_{jk}}$ . Allora dal Lemma 2.40 e dalla (2.14) segue che

$$\text{multideg}(b_{ijk}g_i) \leq \text{multideg}(x^{\delta-\gamma_{jk}}\mathcal{S}(g_j, g_k)) \leq \delta. \quad (2.16)$$

Sostituiamo ora la (2.15) nella (2.12) e otteniamo

$$\begin{aligned} \sum_{m(i)=\delta} \text{LT}(h_i)g_i &= \sum_{j,k} c_{jk}x^{\delta-\gamma_{jk}}\mathcal{S}(g_j, g_k) = \\ &= \sum_{j,k} c_{jk} \left( \sum_{i=1}^t b_{ijk}g_i \right) = \sum_{i=1}^t \left( \sum_{j,k} c_{jk}b_{ijk} \right) g_i = \sum_{i=1}^t p_i g_i, \end{aligned}$$

dove  $p_i = \sum_{j,k} c_{jk}b_{ijk}$ . Osserviamo che per il Lemma 2.13 si ha

$$\text{multideg}(p_i) \leq \max_{ijk}(\text{multideg}(c_{jk}b_{ijk}))$$

dunque

$$\text{multideg}(p_i g_i) \leq \max_{ijk}(\text{multideg}(c_{jk}b_{ijk}g_i)).$$

Allora dalla (2.16) segue che per ogni  $1 \leq i \leq t$

$$\text{multideg}(p_i g_i) \leq \delta$$

poichè i  $c_{jk}$  sono costanti non nulle.

Infine, sostituiamo  $\sum_{m(i)=\delta} \text{LT}(h_i)g_i = \sum_{i=1}^t p_i g_i$  nell'equazione (2.11) e otteniamo

$$f = \sum_{i=1}^t p_i g_i + \sum_{m(i)=\delta} (h_i - \text{LT}(h_i))g_i + \sum_{m(i)<\delta} h_i g_i.$$

Allora  $f$  è una combinazione polinomiale dei  $g_i$  dove tutti i termini hanno multigrado strettamente minore di  $\delta$ , contraddizione perché avevamo assunto  $\delta$  minimo. In particolare

$$\text{multideg}(f) = \text{multideg}(h_i g_i)$$

per certi  $1 \leq i \leq t$  e per quanto detto prima,  $\text{LT}(f)$  è divisibile per  $\text{LT}(g_i)$ .

In conclusione,  $\text{LT}(f) \in \langle \text{LT}(g_1), \dots, \text{LT}(g_t) \rangle$ , dunque  $G$  è una base di Groebner.  $\square$

**Esempio 2.42** Dimostriamo che  $G = \{x^2 + y^2, xy, y^3\}$  è una base di Groebner per l'ideale  $I = \langle x^2 + y^2, xy \rangle$ .

Si ha  $\mathcal{S}(x^2 + y^2, xy) = y^3 \in G$  dunque  $\overline{\mathcal{S}(x^2 + y^2, xy)}^G = 0$ ;

inoltre,  $\mathcal{S}(x^2 + y^2, y^3) = y^5$  e  $y^5 = y^2 y^3$  quindi, per l'Oss. 2.36,  $\overline{\mathcal{S}(x^2 + y^2, y^3)}^G = 0$ ;

infine,  $\mathcal{S}(xy, y^3) = 0$  ed è ovvio che  $\overline{\mathcal{S}(xy, y^3)}^G = 0$ .

Per il criterio di Buchberger (vedasi Teor. 2.41)  $G = \{x^2 + y^2, xy, y^3\}$  è una base di Groebner per  $I$ .

Dato un ideale  $I \subset R_n$ , fino ad ora abbiamo capito come *dimostrare* che una data base è di Groebner per  $I$ . Mediante il seguente algoritmo possiamo *costruirla*.

**Teorema 2.43 (algoritmo di Buchberger)**

Sia  $I = \langle f_1, \dots, f_s \rangle$  un ideale non nullo. Allora una base di Groebner  $G = \{g_1, \dots, g_t\}$  per  $I$  può essere costruita in un numero finito di passi tramite il seguente algoritmo.

Input :  $F = (f_1, \dots, f_s)$

Output : una base di Groebner  $G = (g_1, \dots, g_t)$  per  $I$ , con  $F \subseteq G$ .

$G := F$

DO

$G' := G$

FOR ogni coppia  $\{p, q\}, p \neq q$  in  $G'$  DO

$S := \overline{\mathcal{S}(p, q)}^{G'}$

IF  $S \neq 0$  THEN  $G := G' \cup \{S\}$

WHILE  $G \neq G'$

*Dimostrazione.* Innanzitutto dobbiamo dimostrare che  $G \subseteq I$  per ogni passo dell'algoritmo. Inizialmente  $G' := G := F \subseteq I$ . Quando espandiamo  $G'$  lo facciamo aggiungendo il resto  $S := \overline{\mathcal{S}(p, q)}^{G'}$  per ogni  $p, q \in G'$ , ma  $S \in I$ . Infatti, per l'algoritmo della divisione, esistono  $q_1, q_2, S \in R_n$  tali che

$$\mathcal{S}(p, q) = q_1 p + q_2 q + S.$$

Poichè  $\mathcal{S}(p, q)$  è combinazione lineare di  $p, q \in G' \subseteq I$ , per ogni  $p, q \in G'$  si ha

$$S = \mathcal{S}(p, q) - q_1 p - q_2 q \in I.$$

Pertanto  $G := G' \cup \{S\} \subseteq I$ . Iterando l'algoritmo, otteniamo sempre degli elementi di  $I$ . Osserviamo anche che  $F \subseteq G$  quindi effettivamente  $G$  è una base per  $I$ . L'algoritmo termina quando  $G = G'$ , ovvero quando  $S = \overline{\mathcal{S}(p, q)}^G = 0$ , per ogni  $p, q \in G$ . Per il criterio di Buchberger (vedasi Teor. 2.41) allora tale  $G$  è una base di Groebner per  $I = \langle G \rangle$ .

Ora, manca da dimostrare che l'algoritmo termina in un numero finito di passi. Notiamo che  $G$  consiste di  $G'$  (la precedente  $G$ ) unita ai resti non nulli degli  $\mathcal{S}$ -polinomi di  $G'$ , allora  $G' \subseteq G$  e  $\langle \text{LT}(G') \rangle \subseteq \langle \text{LT}(G) \rangle$ . Inoltre se  $G' \neq G$  allora  $\langle \text{LT}(G') \rangle \subsetneq \langle \text{LT}(G) \rangle$ . Infatti supponiamo di aggiungere un resto non nullo  $r'$  di un  $\mathcal{S}$ -polinomio a  $G'$ . Siccome è ottenuto dalla divisione per  $G'$ , per il Teor. 2.14,

$LT(r)$  non è divisibile per i termini direttivi degli elementi di  $G'$ , dunque per il Lemma 2.20,  $LT(r') \notin \langle LT(G') \rangle$ . Tuttavia  $r' = S \in G$ , quindi  $LT(r') \in \langle LT(G) \rangle$  e, di conseguenza,  $LT(r') \in \langle LT(G) \rangle$ . Inoltre, gli ideali  $\langle LT(G') \rangle$  formano una catena ascendente in  $R_n$  che, per il Teor.2.32, dopo un numero finito di passi si stabilizza. Di conseguenza  $\langle LT(G') \rangle = \langle LT(G) \rangle$ . In particolare, affinché la catena si stabilizzi non può essere sempre  $S \neq 0$ . In un numero finito di passi troveremo  $S = 0$ , dunque  $G' = G$  e l'algoritmo termina. □

**Esempio 2.44** Sia  $I = \langle x^2 + 3x - 4, x^3 - 5x + 4 \rangle$  un ideale in  $\mathbb{R}[x]$ . Determinare una base di Groebner per  $I$ .

Siano  $f_1 = x^2 + 3x - 4$  e  $f_2 = x^3 - 5x + 4$ . Allora  $\mathcal{S}(f_1, f_2) = 3x^2 + x - 4$ . Per il Teor. 2.14,

$$3x^2 + x - 4 = 3(x^2 + 3x - 4) + 0 \cdot (x^3 - 5x + 4) + (-8x + 8),$$

dunque il resto di  $\mathcal{S}(f_1, f_2)$  per  $\{f_1, f_2\}$  è non nullo e vale  $r = -8x + 8 =: f_3$ . Ora, per l'algoritmo di Buchberger (vedasi Teor. 2.43),

$$G = \{x^2 + 3x - 4, x^3 - 5x + 4, -8x + 8\}.$$

Vediamo se tale base è di Groebner.

Ovviamente  $\overline{\mathcal{S}(f_1, f_2)}^G = 0$ . Inoltre  $\mathcal{S}(f_1, f_3) = 4x - 4$  e

$$4x - 4 = 0 \cdot (x^2 + 3x - 4) + 0 \cdot (x^3 - 5x + 4) - \frac{1}{2}(-8x + 8) + 0.$$

Dunque  $\overline{\mathcal{S}(f_1, f_3)}^G = 0$ . Infine calcoliamo  $\overline{\mathcal{S}(f_2, f_3)}^G$ . Si ha  $\mathcal{S}(f_2, f_3) = x^2 - 5x + 4$  e

$$x^2 - 5x + 4 = 1 \cdot (x^2 + 3x - 4) + 0 \cdot (x^3 - 5x + 4) + 1 \cdot (-8x + 8) + 0.$$

Di conseguenza anche  $\overline{\mathcal{S}(f_2, f_3)}^G = 0$ .

In conclusione, per il criterio di Buchberger (vedasi Teor. 2.41), una base di Groebner per  $I$  è

$$G = \{x^2 + 3x - 4, x^3 - 5x + 4, -8x + 8\}.$$

Osserviamo che

$$LT(x^2 + 3x - 4) = x^2 = -\frac{1}{8}x \cdot LT(-8x + 8)$$

e

$$LT(x^3 - 5x + 4) = x^3 = -\frac{1}{8}x^2 \cdot LT(-8x + 8).$$

In particolare, abbiamo trovato una base  $G$  di Groebner per  $I$  con la proprietà che per alcuni polinomi  $p \in G$  il termine direttivo  $\text{LT}(p) \in \langle \text{LT}(G \setminus \{p\}) \rangle$ . Siccome per la Def. 2.29 sono i termini direttivi che definiscono una base di Groebner, allora abbiamo trovato più elementi del necessario. Come vedremo nel seguente lemma, tali elementi sono ridondanti, nel senso che, rimuovendoli, si ottiene ancora una base di Groebner.

**Lemma 2.45** *Sia  $G$  una base di Groebner per un ideale  $I$  non nullo. Sia  $p \in G$  un polinomio tale che  $\text{LT}(p) \in \langle \text{LT}(G \setminus \{p\}) \rangle$ . Allora  $G \setminus \{p\}$  è ancora una base di Groebner per  $I$ .*

*Dimostrazione.* Supponiamo che  $G$  sia una base di Groebner, allora segue direttamente dalla Def. 2.29 che  $\langle \text{LT}(G) \rangle = \langle \text{LT}(I) \rangle$ . Se  $\text{LT}(p) \in \langle \text{LT}(G \setminus \{p\}) \rangle$ , allora  $\langle \text{LT}(I) \rangle \subseteq \langle \text{LT}(G \setminus \{p\}) \rangle$ . Viceversa,  $\text{LT}(G \setminus \{p\}) \subset \text{LT}(I)$  quindi la stessa inclusione vale per gli ideali e dunque  $\langle \text{LT}(G \setminus \{p\}) \rangle = \langle \text{LT}(I) \rangle$ . Per definizione segue che  $G \setminus \{p\}$  è anch'essa una base di Groebner per  $I$ . □

**Esempio 2.46** Sia  $I = \langle x^2 + 3x - 4, x^3 - 5x + 4 \rangle$ . Per quanto detto nell'esempio 2.44,  $G = \{x^2 + 3x - 4, x^3 - 5x + 4, -8x + 8\}$  è una base di Groebner per  $I$ . Applicando due volte il Lemma 2.45 alla base  $G$ , otteniamo una nuova base  $G = \{-8x + 8\}$  di Groebner per  $I$ .

## 2.6 Basi di Groebner minimali e ridotte

Utilizzando il Lemma 2.45 su una qualsiasi base di Groebner e rendendo monici i suoi elementi, otteniamo una base di Groebner particolare.

### Definizione 2.47 (base di Groebner minimale)

Una base di Groebner minimale  $G_m$  per un ideale  $I$  non nullo di  $R_n$  è una base di Groebner  $G$  per  $I$  tale che:

- i.  $LC(p) = 1$ , per ogni  $p \in G$ .
- ii. Per ogni  $p \in G$ ,  $LT(p) \notin \langle LT(G \setminus \{p\}) \rangle$ .

**Esempio 2.48** Sia  $I = \langle y^2 - 1, xy - 1 \rangle \subset \mathbb{R}[x, y]$ . Determiniamo una base di Groebner minimale.

Osserviamo che la base  $F = \{y^2 - 1, xy - 1\}$  non è di Groebner per  $I$  perché  $\mathcal{S}(y^2 - 1, xy - 1) = -x + y$  e il resto della divisione di  $-x + y$  per  $F$  non è zero anzi, è proprio  $-x + y$ . Per l'algoritmo di Buchberger (Teor. 2.43) estendiamo la base  $F$  con la base  $G = \{y^2 - 1, xy - 1, -x + y\}$ .

Ora,  $\mathcal{S}(y^2 - 1, -x + y) = x - y^3$ ,  $\mathcal{S}(xy - 1, -x + y) = y^2 - 1$  e i resti della divisione di  $\mathcal{S}(g_i, g_j)$  per  $G$  sono nulli, per ogni  $g_i, g_j \in G$ . Allora  $G = \{y^2 - 1, xy - 1, -x + y\}$  è una base di Groebner per il criterio di Buchberger (vedasi Teor. 2.41).

Affinchè sia una base minimale, dobbiamo rendere monico l'unico polinomio che non lo è, ossia  $-x + y$ . Dunque  $G = \{y^2 - 1, xy - 1, x - y\}$ . Ora, ogni polinomio  $p \in G$  deve avere termine direttore che non appartiene a  $\langle LT(G \setminus \{p\}) \rangle$ . Allora dato che  $LT(xy - 1) = xy = y \cdot LT(x - y)$ , per il Lemma 2.45 possiamo eliminare dalla base  $xy - 1$ . Una base minimale per  $I$  è  $G_m = \{y^2 - 1, x - y\}$ .

**Osservazione 2.49** Fissato un ordinamento monomiale, sia  $I$  un ideale non nullo di  $R_n$ . Siano  $G$  e  $G'$  due basi di Groebner minimali per  $I$ . Allora  $LT(G) = LT(G')$ .

*Dimostrazione.* Siano  $G$  e  $G'$  due basi di Groebner minimali per  $I$ . Ogni polinomio  $p$  in ciascuna delle due basi è monico, di conseguenza  $LT(p) = LM(p)$ . Sia  $g \in G$ .  $G'$  è una base di Groebner per  $I$  quindi  $\langle LT(G') \rangle = \langle LT(I) \rangle$  e ne segue che  $LT(g) \in \langle LT(G') \rangle$ . Essendo  $\langle LT(I) \rangle$  un ideale monomiale, per il Lemma 2.20 si ha che  $LT(g)$  è divisibile per  $LT(p)$  per almeno un  $p \in G'$  e sia  $g'$  tale polinomio. Ne segue che  $\langle LT(g) \rangle \subset \langle LT(g') \rangle$ . Scambiando il ruolo di  $G$  e  $G'$ , esiste un  $g_1 \in G$  tale che  $LT(g_1)$  divide  $LT(g')$ . Dal momento che  $G$  è una base di Groebner minimale,  $LT(g) \notin \langle LT(G \setminus \{g\}) \rangle$  e, poichè  $LT(g')$  divide  $LT(g)$ , anche  $LT(g') \notin \langle LT(G \setminus \{g\}) \rangle$ . Ne segue che nessun elemento di  $\langle LT(G \setminus \{g\}) \rangle$  divide  $LT(g')$ . Questo implica che

$g_1 = g$ . Ma allora  $\text{LT}(g')$  divide  $\text{LT}(g)$  e  $\text{LT}(g)$  divide  $\text{LT}(g')$ , quindi sono uguali. Pertanto  $\text{LT}(g) \in \text{LT}(G')$ . Dato che questo è vero per ogni  $g \in G$  ne segue che  $\text{LT}(G) \subseteq \text{LT}(G')$ . Scambiando il ruolo si ottiene l'altra inclusione, quindi l'uguaglianza. □

Esistono alcune basi di Groebner minimali che hanno una proprietà particolare. Non solo il termine direttore di un qualsiasi polinomio  $p$  della base non appartiene a  $\langle \text{LT}(G \setminus \{p\}) \rangle$  ma, qualunque monomio di ciascun polinomio in  $G$  soddisfa questo requisito.

**Definizione 2.50 (base di Groebner ridotta)**

Una base di Groebner ridotta  $G_{rid}$  per un ideale  $I$  di  $R_n$ , non nullo, è una base di Groebner  $G$  per  $I$  tale che:

i.  $\text{LC}(p) = 1$ , per ogni  $p \in G$ .

ii. Per ogni  $p \in G$ , nessun monomio di  $p$  è in  $\langle \text{LT}(G \setminus \{p\}) \rangle$ .

**Esempio 2.51** Dato l'ideale  $I = \langle x^2 + 3x - 4, x^3 - 5x + 4 \rangle \subset \mathbb{R}[x]$  troviamo la base di Groebner ridotta per  $I$ .

Per l'esempio 2.46, la base di Groebner ridotta (e quindi minimale) per  $I$  è

$$G_{rid} = \{x - 1\}.$$

Infatti, basta rendere monico il suo unico elemento mediante la relazione

$$-8x + 8 = -\frac{1}{8}(x - 1).$$

**Esempio 2.52** Sia  $I = \langle x^3 - 2xy, x^2y - 2y^2 + x \rangle \subset \mathbb{R}[x, y]$ , con i monomi ordinati secondo l'ordine *grlex*. Verifichiamo che  $f = 4x^3y^2 + 3xy^3 + x^2 \in I$ .

Per l'algoritmo di Buchberger (vedasi Teor. 2.41), una base di Groebner per  $I$  è  $G = \{x^3 - 2xy, x^2y - 2y^2 + x, -x^2, -2xy, -2y^2 + x\}$ . Rendendo monici tutti i suoi polinomi e osservando che

$$\text{LT}(x^3 - 2xy) = x^3 = x \cdot \text{LT}(x^2)$$

$$\text{LT}(x^2y - 2y^2 + x) = x^2y = y \cdot \text{LT}(x^2)$$

la base  $G_{rid} = \{x^2, xy, y^2 - \frac{1}{2}x\}$  è una base minimale e ridotta per  $I$  per il Lemma 2.45. Infine,

$$4x^3y^2 + 3xy^3 + x^2 = (4xy^2 + 1)x^2 + (3y^2)xy,$$

quindi  $f$  è combinazione lineare di due elementi di  $G_{rid}$ . Dato che  $G_{rid} \subset I$  allora  $f \in I$ .

**Proposizione 2.53** *Sia  $I$  un ideale di  $R_n$ , non nullo. Allora per un dato ordinamento monomiale,  $I$  ammette una e una sola base di Groebner ridotta.*

*Dimostrazione.* Sia  $G$  una base di Groebner minimale per  $I$ . Diremo che  $g \in G$  è ridotto se nessun monomio di  $g$  è in  $\langle \text{LT}(G \setminus \{g\}) \rangle$ . Osserviamo che se  $g$  è ridotto per  $G$ , allora è ridotto per ogni altra base minimale di  $I$  che contenga  $g$  e abbia lo stesso insieme di termini direttivi. Infatti, una volta che un elemento è ridotto, rimane tale in quanto non vengono modificati i termini direttivi. Ora, dato  $g \in G$ , sia  $g' = \bar{g}^{G \setminus \{g\}}$  e poniamo  $G' = (G \setminus \{g\}) \cup \{g'\}$ .

Vogliamo dimostrare che  $G'$  è una base minimale per  $I$ . Innanzitutto  $g' \neq 0$  altrimenti  $\text{LT}(g) \in \langle \text{LT}(G \setminus \{g\}) \rangle$ . Inoltre, osserviamo che  $\text{LT}(g') = \text{LT}(g)$ . Infatti, sia  $G = \{g_1, \dots, g_s, g\}$ . Nella divisione di  $g$  per  $G \setminus \{g\}$ , per l'algoritmo di divisione (Teor. 2.14), al primo passo, si aggiunge  $\text{LT}(g)$  a  $r = 0$  dato che  $\text{LT}(g)$  non è divisibile per nessun elemento di  $\text{LT}(G \setminus \{g\})$ . Poi ci sono passi di divisione o di resto. In un passo di divisione si ha  $p' = p - \frac{\text{LT}(p)}{\text{LT}(g_i)}g_i$  e  $\text{LT}(p') < \text{LT}(p)$ . In un passo di resto si ha  $p' = p - \text{LT}(p)$  quindi di nuovo  $\text{LT}(p') < \text{LT}(p)$ . Ma allora al primo resto  $\text{LT}(g)$  si potranno aggiungere, in passi di resto, degli  $\text{LT}(p')$  che sono sempre più piccoli di  $\text{LT}(g)$ . Per questo  $\text{LT}(g') = \text{LT}(g)$ . Per come abbiamo definito  $G$  e  $G'$  e dal fatto che  $\text{LT}(g') = \text{LT}(g)$  segue che  $\text{LT}(G') = \text{LT}(G)$ . Inoltre  $G'$  è contenuto in  $I$  perché  $g' \in I$  essendo il resto della divisione di  $g$  per  $(g_1, \dots, g_s)$  e  $\langle \text{LT}(G') \rangle = \langle \text{LT}(G) \rangle = \langle \text{LT}(I) \rangle$ , quindi  $G'$  è una base di Groebner per  $I$ .

Inoltre,  $\text{LC}(g') = \text{LC}(g) = 1$ , e per ogni  $g_i$  in  $G \setminus \{g\}$ , per ogni  $i$  tale che  $1 \leq i \leq s$ , si ha che  $\text{LC}(g_i) = 1$ , quindi  $G'$  soddisfa la proprietà  $i$ . della Def. 2.47. Da  $G = \{g_1, \dots, g_s, g\}$  segue che  $G' = \{g_1, \dots, g_s, g'\}$ . Sappiamo già che  $\text{LT}(g_i)$  non divide  $\text{LT}(g)$  (dato che  $G$  è minimale) e dunque  $\text{LT}(g')$ , quindi  $\text{LT}(g')$  non appartiene a  $\langle \text{LT}(G' \setminus \{g'\}) \rangle$ . Invece se per assurdo  $\text{LT}(g_i)$  appartenesse a  $\langle \text{LT}(G' \setminus \{g_i\}) \rangle$  allora, non essendo diviso da  $\text{LT}(g_j)$  per  $j \neq i$  (dato che  $G$  è minimale) si avrebbe che  $\text{LT}(g')$  divide  $\text{LT}(g_i)$ . Quindi  $\text{LT}(g)$  divide  $\text{LT}(g_i)$  assurdo essendo  $G$  minimale. Quindi  $G'$  è minimale.

Inoltre,  $g'$  è ridotto per  $G'$  perché per l'algoritmo di divisione nessun monomio in  $g'$  è diviso da  $\text{LT}(g_i)$ , quindi nessun monomio in  $g'$  appartiene a  $\langle \text{LT}(g_1), \dots, \text{LT}(g_s) \rangle = \langle \text{LT}(G \setminus \{g\}) \rangle = \langle \text{LT}(G' \setminus \{g'\}) \rangle$ .

Ora, ripetiamo il procedimento appena descritto per ogni elemento di  $G \setminus g$  fino a quando tutti gli elementi della base di Groebner sono stati ridotti. Per quanto detto prima, una volta che un elemento è ridotto, rimane tale in quanto non vengono modificati i termini direttivi, di conseguenza si ottiene una base di Groebner ridotta.

Per dimostrare l'unicità, supponiamo che  $G_1$  e  $G_2$  siano entrambe basi ridotte per  $I$ . In particolare,  $G_1$  e  $G_2$  sono basi minimali per  $I$  e per l'Oss. 2.49 hanno gli stessi termini direttivi, o equivalentemente,  $\text{LT}(G_1) = \text{LT}(G_2)$ . Quindi dato  $g_1 \in G_1$  esiste  $g_2 \in G_2$  tale che  $\text{LT}(g_1) = \text{LT}(g_2)$ . Se dimostriamo che  $g_1 = g_2$  allora, di conseguenza  $G_1$  è contenuto in  $G_2$  e scambiando il ruolo,  $G_2$  è contenuto in  $G_1$ , quindi l'unicità sarà così dimostrata. Per poter fare ciò, consideriamo la differenza  $g_1 - g_2$ . Ovviamente  $g_1 - g_2 \in I$  e poichè  $G_1$  è una base di Groebner, per il Cor. 2.34  $\overline{g_1 - g_2}^{G_1} = 0$ . Sappiamo che  $\text{LT}(g_1) = \text{LT}(g_2)$  quindi i termini direttivi si elidono in  $g_1 - g_2$ . Invece, i termini rimanenti non sono divisibili per alcun elemento di  $\text{LT}(G_1) = \text{LT}(G_2)$  essendo  $G_1$  e  $G_2$  basi di Groebner ridotte per  $I$ . Questo implica che  $g_1 - g_2 = \overline{g_1 - g_2}^{G_1} = 0$  e dunque  $g_1 = g_2$ .  $\square$

Allora se ne deduce il seguente fatto.

**Corollario 2.54** *Due ideali sono uguali se e solo se coincidono le loro basi di Groebner ridotte.*

In particolare, per verificare che  $\langle f_1, \dots, f_s \rangle$  e  $\langle g_1, \dots, g_t \rangle$  coincidono basta calcolare le basi di Groebner ridotte.

Avevamo accennato che la base utilizzata nella dimostrazione del lemma di Dickson (Teor. 2.23) fosse una base di Groebner ridotta. Questo fatto è garantito dal seguente risultato.

**Lemma 2.55** *Sia  $I = \langle x^\alpha, \alpha \in A \rangle \subset R_n$  un ideale monomiale e siano  $\alpha(1), \dots, \alpha(s)$  in  $A$  tali che  $I = \langle x^{\alpha(1)}, \dots, x^{\alpha(s)} \rangle$ . Allora  $G = \{x^{\alpha(1)}, \dots, x^{\alpha(s)}\}$  è una base di Groebner di  $I$ . Inoltre, se scegliamo  $s$  come il minimo numero di monomi in  $A$  che generano  $I$ , allora la base  $G$  è una base di Groebner ridotta.*

*Dimostrazione.* Affinchè  $G = \{x^{\alpha(1)}, \dots, x^{\alpha(s)}\}$  sia una base di Groebner, per il criterio di Buchberger (Teor. 2.41), occorre verificare che  $\overline{\mathcal{S}(x^{\alpha(i)}, x^{\alpha(j)})}^G = 0$  per ogni  $i, j$  tali che  $1 \leq i \neq j \leq s$ .

Sia  $x^\gamma = \text{mcm}(\text{LM}(x^{\alpha(i)}), \text{LM}(x^{\alpha(j)}))$ , allora

$$\begin{aligned} \mathcal{S}(x^{\alpha(i)}, x^{\alpha(j)}) &= \frac{x^\gamma}{\text{LT}(x^{\alpha(i)})} x^{\alpha(i)} - \frac{x^\gamma}{\text{LT}(x^{\alpha(j)})} x^{\alpha(j)} = \frac{x^\gamma}{x^{\alpha(i)}} x^{\alpha(i)} - \frac{x^\gamma}{x^{\alpha(j)}} x^{\alpha(j)} = \\ &= x^{\gamma - \alpha(i)} x^{\alpha(i)} - x^{\gamma - \alpha(j)} x^{\alpha(j)} = x^\gamma - x^\gamma = 0. \end{aligned}$$

Dunque possiamo concludere che  $G = \{x^{\alpha(1)}, \dots, x^{\alpha(s)}\}$  è una base di Groebner per  $I$ .

Ora, vogliamo dimostrare che  $G$  è una base di Groebner ridotta nel caso in cui  $s$  è il minimo numero di monomi di  $A$  che generano  $I$ . Innanzitutto, i monomi sono monici per definizione e  $LT(x^{\alpha(i)}) = x^{\alpha(i)}$ . Di conseguenza,  $LT(G \setminus \{x^{\alpha(i)}\}) = G \setminus \{x^{\alpha(i)}\}$ . Inoltre, le Def. 2.47 e 2.50 sono equivalenti perché ogni polinomio di  $G$  coincide con il suo monomio direttivo. Quindi, affinché  $G$  sia una base ridotta richiediamo le stesse condizioni di una base minimale. Quindi, richiediamo che per ogni  $x^{\alpha(i)} \in G$  deve valere che  $LT(x^{\alpha(i)}) \notin \langle LT(G \setminus \{x^{\alpha(i)}\}) \rangle$  o, equivalentemente,  $x^{\alpha(i)} \notin \langle G \setminus \{x^{\alpha(i)}\} \rangle$ . Ma, questo fatto è vero dal momento che abbiamo scelto  $s$  minimo. Se fosse  $x^{\alpha(i)} \in \langle G \setminus \{x^{\alpha(i)}\} \rangle$  allora, per il Lemma 2.20, esisterebbe  $j$  tale che  $1 \leq j \neq i \leq s$  e tale che  $x^{\alpha(j)}$  divide  $x^{\alpha(i)}$ . Ma allora  $I$  sarebbe generato anche da  $\{x^{\alpha(1)}, \dots, x^{\alpha(i-1)}, x^{\alpha(i+1)}, \dots, x^{\alpha(s)}\}$  contraddicendo il fatto che  $s$  è minimo. Di conseguenza, la base  $G$  è minimale e anche ridotta.

□

## 2.7 Zeri di un insieme di polinomi

Nel primo capitolo ci siamo posti alcune domande interessanti. Supponiamo di avere  $f_1, \dots, f_n \in R_n$ .

- Possiamo determinare se  $V(f_1, \dots, f_n)$  è non vuota? Ovvero, esiste una soluzione comune a tutte le equazioni  $f_i=0$ ?
- Possiamo determinare se  $V(f_1, \dots, f_n)$  è finita?

Adesso abbiamo le conoscenze necessarie per poter rispondere a tali domande.

### Risoluzione di equazioni polinomiali

Supponiamo di voler risolvere il sistema polinomiale

$$\begin{cases} f_1(x_1, \dots, x_n) = 0 \\ \vdots \\ f_s(x_1, \dots, x_n) = 0 \end{cases}$$

Equivalentemente, richiediamo di trovare tutti i punti di  $V(I)$ , dove  $I$  è l'ideale generato da  $\langle f_1, \dots, f_s \rangle$ .

Innanzitutto, per l'Oss. 1.9,

$$V(I) = V(\langle f_1, \dots, f_s \rangle) = V(f_1, \dots, f_s).$$

Inoltre, per la Prop. 1.10, se un ideale polinomiale ammette due basi  $F$  e  $G$ , allora

$$V(F) = V(G).$$

Dato un ideale  $I = \langle f_1, \dots, f_s \rangle$ , dagli esempi si osserva che per determinare tutti i punti di  $V(I)$  è conveniente trovare la base di Groebner ridotta  $G_{rid}$  per  $I$  e calcolare  $V(G_{rid})$ . In realtà questo fatto vale in generale (si veda [CLO, cap.3]).

**Esempio 2.56** Supponiamo di voler risolvere il seguente sistema di equazioni polinomiali.

$$\begin{cases} x^2 + 3x - 4 = 0 \\ x^3 - 5x + 4 = 0 \end{cases}$$

Queste equazioni determinano l'ideale  $I = \langle x^2 + 3x - 4, x^3 - 5x + 4 \rangle \subset \mathbb{R}[x]$ . Vogliamo trovare allora tutti i punti in  $V(I)$ . In particolare utilizzando la base di Groebner ridotta  $G_{rid} = \{x - 1\}$ , trovata nell'esempio 2.51, otteniamo

$$V(I) = \{1\}.$$

**Esempio 2.57** Risolviamo il seguente sistema polinomiale in  $\mathbb{R}[x, y]$  con i monomi ordinati secondo l'ordine *grlex*.

$$\begin{cases} x^3 - 2xy = 0 \\ x^2y - 2y^2 + x = 0 \end{cases}$$

Il problema è equivalente a trovare tutti i punti in  $V(I)$  con

$$I = \langle x^3 - 2xy, x^2y - 2y^2 + x \rangle.$$

Per quanto visto nell'esempio 2.52 la base  $G_m = \{x^2, xy, y^2 - \frac{1}{2}x\}$  è una base minimale per  $I$ . Di conseguenza,

$$V(I) = V(G_{rid}) = \{(0, 0)\}.$$

## Forma implicita

Siano  $f_1, \dots, f_n$  polinomi nelle variabili  $t_1, \dots, t_m$  e sia  $V$  il sottoinsieme di  $\mathbb{K}^n$  dato tramite la seguente parametrizzazione:

$$\begin{cases} x_1 - f_1(t_1, \dots, t_m) = 0 \\ \vdots \\ x_n - f_n(t_1, \dots, t_m) = 0 \end{cases} \quad (2.17)$$

Vogliamo trovare un sistema di equazioni polinomiali nelle variabili  $x_i$ , per  $i$  tale che  $1 \leq i \leq n$ , che definiscano  $V$ .

L'idea è quindi quella di eliminare le variabili  $t_1, \dots, t_m$  dalle equazioni del sistema (2.17).

Consideriamo l'ordine *lex* in  $\mathbb{K}[t_1, \dots, t_m, x_1, \dots, x_n]$

$$t_1 >_{lex} \dots >_{lex} t_m >_{lex} x_1 >_{lex} \dots >_{lex} x_n.$$

Supponiamo di avere una base di Groebner per l'ideale  $I = \langle x_1 - f_1, \dots, x_n - f_n \rangle$ . Dai risultati del [CLO, Cap.3] si osserva che trovare una base di Groebner per un ideale rispetto all'ordine *lex* semplifica la forma delle equazioni che definiscono l'ideale. In particolare, otteniamo equazioni in cui le variabili sono eliminate in successione. Notiamo inoltre che l'ordine di eliminazione sembra corrispondere all'ordinamento delle variabili. Quindi ci aspettiamo che le variabili  $t_1, \dots, t_m$  siano le prime ad essere eliminate. In questo modo i polinomi della base di Groebner per  $I$  dovrebbero contenere polinomi esclusivamente nelle variabili  $x_1, \dots, x_n$ . Allora tali polinomi sono i candidati come equazioni per  $V$ .

**Esempio 2.58** Consideriamo il sistema

$$\begin{cases} x = t_1 + t_2 + 1 \\ y = t_1 - t_2 + 3 \\ z = 2t_1 - 2 \\ w = t_1 + 2t_2 - 3, \end{cases} \quad (2.18)$$

il quale rappresenta parametricamente i punti del piano affine di  $\mathbb{K}^4$  passante per  $(1, 3, -2, -3)$  e di giacitura  $W = \langle (1, 1, 2, 1), (1, -1, 0, 2) \rangle$ .

Calcoliamo la base di Groebner per l'ideale

$$I = \langle t_1 + t_2 - x + 1, t_1 - t_2 - y + 3, 2t_1 - z - 2, t_1 + 2t_2 - w - 3 \rangle$$

i cui monomi sono ordinati secondo l'ordine *lex*. La base di Groebner  $G$  ha 4 elementi in totale,

$$G = \{4x - z - 2w - 12, 4y - 3z + 2w - 12, 4t_2 + z - 2w - 4, 2t_1 - z - 2\}.$$

Gli ultimi due polinomi dipendono dalle variabili  $t_1, t_2$  quindi le equazioni che definiscono la varietà affine sono  $4x - z - 2w - 12 = 0$  e  $4y - 3z + 2w - 12 = 0$ .

Ovviamente si poteva arrivare allo stesso risultato esplicitando le variabili  $t_1$  e  $t_2$  dalla terza e quarta equazione, rispettivamente, del sistema (2.18) e sostituendo tali valori nelle prime due equazioni.

## Varietà finite

Dato un ideale  $I$  non nullo di  $R_n$ , per determinare il metodo che ci permette di capire se  $V(I)$  è finita, dobbiamo poter introdurre alcuni concetti.

**Definizione 2.59** Sia  $I$  un ideale non nullo di  $R_n$ . Fissato un ordine monomiale  $>$ , diremo che un monomio  $x^\alpha$  con  $\alpha \in A$  è standard rispetto ad  $I$  se  $x^\alpha \notin \langle \text{LT}(I) \rangle$ .

**Esempio 2.60** Sia  $I = \langle y^2 - 1, xy - 1 \rangle$  in  $\mathbb{R}[x, y]$ . Abbiamo dimostrato nell'esempio 2.48 che una base di Groebner per  $I$  è  $G = \{y^2 - 1, xy - 1, -x + y\}$ . Quindi  $\langle \text{LT}(I) \rangle = \langle y^2, xy, -x \rangle$ . In particolare osserviamo che  $y$  è un monomio standard per  $I$ . Infatti  $y^2$ ,  $xy$  e  $x$  non dividono  $y$  quindi  $y \notin \langle \text{LT}(I) \rangle$ .

**Definizione 2.61** Sia  $I$  un ideale non nullo di  $R_n$  e siano  $\alpha(1), \dots, \alpha(m) \in \mathbb{Z}_{\geq 0}^n$ . Diremo che l'insieme  $S = \{x^{\alpha(1)}, \dots, x^{\alpha(m)}\}$  è standard per  $I$  se ogni monomio  $x^{\alpha(j)} \in S$  è standard per  $I$ .

**Osservazione 2.62** Sia  $I$  un ideale non nullo di  $R_n$ . Allora il numero di monomi standard è finito se e solo se qualche potenza di  $x_i$ , per  $i$  tale che  $1 \leq i \leq n$ , appare tra i generatori di  $\langle \text{LT}(I) \rangle$ .

*Dimostrazione.* Il numero di monomi standard è finito se per un numero finito di  $\alpha \in \mathbb{Z}_{\geq 0}^n$  si ha  $x^\alpha \notin \langle \text{LT}(I) \rangle$ . Tale condizione è equivalente a richiedere che qualche potenza di  $x_i$ , al variare di  $i$ , appare tra i generatori di  $\langle \text{LT}(I) \rangle$ . Infatti supponiamo che per ogni  $i$  tale che  $1 \leq i \leq n$ , esista un  $s_i \in \mathbb{Z}_{\geq 1}$  tale che  $x_i^{s_i}$  appare tra i generatori di  $\langle \text{LT}(I) \rangle$ . Allora  $x^\alpha \notin \langle \text{LT}(I) \rangle$  se e solo se  $\alpha(i) < s_i$  per ogni  $i$ , quindi i possibili  $\alpha$  sono un numero finito. Per dimostrare il viceversa, faremo vedere che se nessuna potenza di  $x_i$ , per qualche  $i$  tale che  $1 \leq i \leq n$ , appare tra i generatori di  $\langle \text{LT}(I) \rangle$ , allora l'insieme di monomi standard per  $I$  è infinito. Fissato  $i$  tale che  $1 \leq i \leq n$ , se nessuna di  $x_i$  appare fra i generatori di  $\langle \text{LT}(I) \rangle$  allora  $x_i^s \notin \langle \text{LT}(I) \rangle$ , per ogni intero  $s \geq 1$ . Allora il numero di monomi standard per  $I$  è infinito. □

**Esempio 2.63** Sia  $I = \langle t_1 + t_2 - x + 1, t_1 - t_2 - y + 3, 2t_1 - z - 2, t_1 + 2t_2 - w - 3 \rangle$  in  $\mathbb{K}[t_1, t_2, x, y, z]$  i cui monomi sono ordinati secondo l'ordine *lex*. Il numero di monomi standard è infinito perché nessuna potenza di  $z$  compare fra i generatori di  $\langle \text{LT}(I) \rangle$ . Infatti, la base di Groebner per  $I$  è

$$G = \{4x - z - 2w - 12, 4y - 3z + 2w - 12, 4t_2 + z - 2w - 4, 2t_1 - z - 2\}.$$

Dunque,  $\langle \text{LT}(I) \rangle = \langle x, y, t_2, t_1 \rangle$ .

Osserviamo anche che  $V(I)$  è infinito se  $\mathbb{K}$  è infinito.

**Esempio 2.64** Sia  $I = \langle x^4 + 7y, y^3 - 2y + 4, z^2 - 3 \rangle \subset \mathbb{R}[x, y, z]$ , i cui monomi sono ordinati secondo l'ordine *lex*. Il numero di monomi standard è finito perché le variabili  $x, y, z$  compaiono in  $\langle \text{LT}(I) \rangle$  con qualche potenza. Infatti,  $\{x^4 + 7y, y^3 - 2y + 4, z^2 - 3\}$  è una base di Groebner per  $I$  e  $x^4, y^3$  e  $z^2$  sono i generatori di  $\langle \text{LT}(I) \rangle$ .

**Osservazione 2.65** Dato un polinomio  $f$  di grado  $d$  in  $\mathbb{K}[x]$ , il numero dei monomi standard per  $I = \langle f \rangle$  è finito e coincide con il grado del polinomio.

*Dimostrazione.* Osserviamo che  $\langle \text{LT}(I) \rangle = \langle \text{LT}(f) \rangle = \langle x^d \rangle$ , dunque l'insieme dei monomi standard  $S$  è finito per l'Oss. 2.62. In particolare,  $S = \{1, x, x^2, \dots, x^{d-1}\}$ .  $\square$

Ora, possiamo definire un metodo che ci permette di capire se  $V(I)$  è finita.

**Teorema 2.66** *Sia  $\mathbb{K} = \mathbb{Q}$ . Dato un ideale  $I$  non nullo in  $R_n$ . La varietà  $V(I)$  è finita se e solo se l'insieme di monomi standard per  $I$  è finito. In questo caso il numero dei monomi standard equivale alla cardinalità di  $V(I)$ , dove gli zeri sono contati con molteplicità.*<sup>4</sup>

---

<sup>4</sup>Vedasi [PS, Cap. 3, Teor. 3.5].

# Capitolo 3

## Applicazione ai modelli epidemiologici

Le equazioni matematiche modellano i processi dell'epidemia: in particolare, tutti i modelli epidemiologici sono descritti da equazioni differenziali

$$\begin{cases} \dot{x}_1 = f_1(x_1, \dots, x_n) \\ \vdots \\ \dot{x}_n = f_n(x_1, \dots, x_n) \end{cases}$$

dove gli  $f_i(x_1, \dots, x_n)$ ,  $\forall i$  con  $1 \leq i \leq n$ , sono polinomi.

Come descritto nell'introduzione, attraverso l'analisi di queste equazioni, per ogni modello epidemiologico studieremo:

- i. le *varietà dei punti di equilibrio*: i punti di equilibrio sono tutti i punti che soddisfano il sistema

$$\begin{cases} f_1(x_1, \dots, x_n) = 0 \\ \vdots \\ f_n(x_1, \dots, x_n) = 0. \end{cases}$$

Per quanto affermato nella sezione 2.7, per determinare tali punti è conveniente determinare la base di Groebner ridotta  $G_{rid}$  dell'ideale associato  $\langle f_1(x_1, \dots, x_n), \dots, f_n(x_1, \dots, x_n) \rangle \subset R_n$  e calcolare  $V(G_{rid})$ .

- ii. la *stabilità dei punti di equilibrio* in funzione del *tasso di riproduzione* della malattia  $R_0$ , ovvero il numero medio di persone che ogni individuo infetto contagia. Una volta trovato il valore  $R_0$  è possibile capire l'andamento della diffusione del virus, quindi se l'epidemia tenderà a diffondersi in tutta la popolazione o se tenderà a scomparire.

Come indicato nell'introduzione, il metodo di studio di ciascun modello epidemio-  
logico prevede i seguenti passi.

- i. Presentiamo il modello matematico mediante la descrizione del corrispon-  
dente sistema di equazioni differenziali ordinarie.
- ii. Dal sistema di equazioni differenziali determiniamo l'ideale  $J$  associato a  
ciascun sistema.
- iii. Calcoliamo la base di Groebner, in particolare la base di Groebner ridotta  
 $G_{rid}$ , associata all'ideale  $J$ .
- iv. Calcolando  $V(G_{rid})$  otteniamo i punti di equilibrio di ciascun sistema.
- v. Calcoliamo la matrice Jacobiana associata al sistema in ciascun punto di  
equilibrio e i suoi rispettivi autovalori.
- vi. Definito  $R_0$ , applicando i teoremi di stabilità è possibile determinare se cia-  
scun punto di equilibrio è stabile o instabile in funzione del valore  $R_0$ . In  
questo modo sapremo se l'epidemia si espanderà o meno.

**Osservazione 3.1** In tutti i modelli che studieremo ((3.11), (3.39), (3.58), (3.66),  
(3.68)) è facile vedere che partendo da dati iniziali non negativi si rimane sempre  
nella regione in cui le variabili sono non negative. Questo è in accordo con il fatto  
che, nei modelli epidemiologici, le variabili sono non negative. Pertanto studieremo  
solo i punti di equilibrio con coordinate non negative.

Una volta trovato il valore  $R_0$  è possibile capire l'andamento della diffusione del  
virus, quindi se l'epidemia tenderà a diffondersi in tutta la popolazione o se tenderà  
a scomparire:

- se  $R_0 < 1$ , la malattia non supera il numero di infetti dell'istante iniziale  
e quindi decade a zero senza diffondersi in maniera drastica attraverso la  
popolazione.
- se  $R_0 > 1$  la malattia è sufficientemente forte da diffondersi attraverso la  
popolazione. In tal caso la malattia è tanto forte più è grande il valore di  
 $R_0$ .
- se  $R_0 = 1$  non possiamo concludere a prescindere se l'epidemia andrà a  
estinguersi o meno. Questo perchè sotto la condizione  $R_0 = 1$  la matrice  
Jacobiana del sistema che descrive i vari modelli epidemiologici ha almeno

un autovalore nullo, quindi non è possibile applicare i teoremi di stabilità. Si veda come esempio il modello SEIR.

Come visto nell'introduzione, i modelli epidemiologici che studiamo possono avere due tipi di struttura differente: epidemica o endemica. Ci occuperemo nello specifico di cinque modelli epidemiologici principali, due modelli epidemici e tre modelli endemici.

### 3.1 Cenni di analisi

I modelli epidemiologici sono espressi matematicamente da sistemi di equazioni differenziali. In particolare, tutti i modelli epidemiologici sono descritti da equazioni differenziali ordinarie (ODE). Dunque, per poterli studiare nel dettaglio, abbiamo bisogno di richiamare alcuni concetti di analisi.

I teoremi e gli enunciati che seguono sono stati riportati da [GE].

#### Equazioni differenziali ordinarie

##### Definizione 3.2 (sistema dinamico)

Dati uno spazio vettoriale reale  $V$  e un insieme aperto  $W \subseteq V$ , definiremo sistema dinamico una coppia  $(W, \varphi)$ , dove

$$\varphi : J \times W \longrightarrow W,$$

con  $J$  intervallo di  $\mathbb{R}$ , è tale che  $\varphi \in C^1$  e verifica le seguenti proprietà:

- i.  $\exists t_0 \in J$  tale che  $\varphi(t_0, x) = x, \forall x \in W$ ;
- ii.  $\varphi(t + s, x) = \varphi(t, \varphi(s, x)), \forall x \in W, \forall t, s \in J : t + s \in J$ .

Data un'applicazione  $(t, x) \longrightarrow \varphi(t, x)$ , per l'ipotesi di differenziabilità, è possibile definire un campo vettoriale

$$f(x) = \left. \frac{d}{dt} \varphi(t, x) \right|_{t=t_0}, \quad (3.1)$$

per ogni  $x \in W$  e per qualche  $t_0 \in J$  fissato.

Se poniamo  $x(t) = \varphi(t, x)$ , sottolineando la sola dipendenza dal tempo della  $\varphi$ , la (3.1) diventa

$$\dot{x} = \frac{dx}{dt} = f(x). \quad (3.2)$$

Se introduciamo in  $V$  un sistema di coordinate, allora  $x = (x_1, \dots, x_n)$  e la (3.2) è equivalente al sistema di  $n$  equazioni

$$\begin{cases} \dot{x}_1 = f_1(x_1, \dots, x_n) \\ \vdots \\ \dot{x}_n = f_n(x_1, \dots, x_n) \end{cases} \quad (3.3)$$

o, equivalentemente,

$$\dot{x} = f(x) \quad (3.4)$$

con  $x \in W \subseteq \mathbb{R}^n$  e  $f : W \rightarrow \mathbb{R}^n$  è un campo vettoriale continuo.

**Definizione 3.3 (equazioni differenziali ordinarie del primo ordine)**

*Dati uno spazio vettoriale reale  $V$  e un insieme aperto  $W \subseteq V$ , chiameremo sistema di equazioni differenziali ordinarie delle relazioni che legano  $x : J \rightarrow W$ , con  $J$  intervallo di  $\mathbb{R}$ , ad alcune delle sue derivate:*

$$F(t, x(t), x^{(1)}(t), \dots, x^{(p)}(t)) = 0, \quad p \in \mathbb{N}, \quad (3.5)$$

con  $F : J \times W \times V^p \rightarrow \mathbb{R}^n$  continua nei suoi argomenti. Indichiamo con

$$x^{(j)} = \frac{d^j x}{dt^j}.$$

*Diremo che il sistema (3.5) è in forma normale se la derivata di ordine più alto è isolata al primo membro, ossia:*

$$x^{(p)}(t) = F(t, x(t), x^{(1)}(t), \dots, x^{(p-1)}(t)). \quad (3.6)$$

*Diremo inoltre che il sistema (3.5) è del primo ordine se  $p = 1$ . Infine, diremo che il sistema (3.5) è autonomo se  $F$  non dipende esplicitamente dal tempo, cioè  $\frac{\partial F}{\partial t} = 0$ .*

Ci concentreremo nell'analisi di equazioni differenziali ordinarie in forma normale del primo ordine. Studieremo sistemi autonomi e scriveremo il sistema già in forma normale.

**Definizione 3.4 (problema di Cauchy)**

*Il problema della determinazione delle soluzioni del sistema di equazioni differenziali del primo ordine con condizioni iniziali*

$$\begin{cases} \dot{x} = f(x) \\ x(t_0) = x_0 \end{cases} \quad (3.7)$$

con  $x_0 \in W \subseteq V$  e  $f : W \rightarrow V$ , prende il nome di problema di Cauchy.

Come vedremo nel Teor. 3.6, se la funzione  $f$  è di classe  $C^1$  la soluzione del problema di Cauchy esiste ed è unica.

**Lemma 3.5** Una funzione continua  $u : J \rightarrow W$  è soluzione del problema integrale

$$u(t) = x_0 + \int_{t_0}^t f(u(s)) ds \quad (3.8)$$

con  $f$  continua, se e solo se  $u$  è di classe  $C^1$  e risolve il problema di Cauchy (3.7).

**Teorema 3.6** Siano  $W \subseteq V$  un sottoinsieme aperto di uno spazio vettoriale normato  $V$  e sia  $x_0 \in W$ . Sia  $f : W \rightarrow V$  un'applicazione di classe  $C^1$ . Allora esiste  $a > 0$  e un'unica soluzione  $u : J \rightarrow W$ , con  $J = [t_0 - a, t_0 + a]$ , del problema di Cauchy (3.7).

## Stabilità e punti di equilibrio

### Definizione 3.7 (autovalori)

Sia  $A$  una matrice quadrata di ordine  $n$  a coefficienti in un campo  $\mathbb{K}$ , gli autovalori  $\lambda_1, \dots, \lambda_n$  di  $A$  sono le radici in  $\mathbb{K}$  del polinomio

$$p_n(\lambda) = \det(A - \lambda \mathbb{I}_n)$$

dove  $\mathbb{I}_n$  è la matrice identità di ordine  $n$ .

### Definizione 3.8 (punto di equilibrio)

Chiameremo punto di equilibrio (o punto critico) del sistema (3.4) un punto  $x_e \in \mathbb{R}^n$  tale che  $f(x_e) = 0$ .

### Definizione 3.9 (stabilità secondo Ljapunov)

Dato un punto di equilibrio  $x_e$  per il sistema (3.4) e il dato iniziale  $\varphi(0, \bar{x}) = \bar{x}$ , con  $\bar{x} \in W \subseteq \mathbb{R}^n$ , diremo che:

- i.  $x_e$  è un punto di equilibrio stabile secondo Ljapunov se per ogni  $\epsilon > 0$  esiste un  $\delta > 0$  tale che  $|\bar{x} - x_e| < \delta$  allora  $|\varphi(t, \bar{x}) - x_e| < \epsilon$  per ogni  $t \geq 0$ ;
- ii.  $x_e$  è un punto di equilibrio instabile se non è stabile;
- iii.  $x_e$  è attrattivo se esiste un intorno  $B(x_e)$  di  $x_e$  in  $W$  tale che, per ogni  $\bar{x} \in B(x_e)$  si ha

$$\lim_{t \rightarrow \infty} |\varphi(t, \bar{x}) - x_e| = 0;$$

- iv.  $x_e$  è asintoticamente stabile se è stabile e attrattivo.

**Definizione 3.10** Sia  $x_e$  un punto di equilibrio per il sistema dinamico (3.4). Poichè  $f$  è di classe  $C^1$ , e  $f(x_e) = 0$ , possiamo scrivere

$$\dot{x} = A(x - x_e) + Q(x), \quad \lim_{x \rightarrow x_0} \frac{|Q(x)|}{|x - x_e|} = 0,$$

dove  $A$  è la matrice  $n \times n$  di elementi  $A_{ij} = [\partial f_i / \partial x_j](x_e)$  e  $Q(x)$  è un infinitesimo di ordine superiore al primo rispetto a  $|x - x_e|$ .

Trascurando la correzione  $Q(x)$ , si ottiene il sistema

$$\dot{x} = A(x - x_e) \tag{3.9}$$

che chiamiamo sistema linearizzato del sistema (3.4) nell'intorno del punto d'equilibrio  $x_e$ .

**Teorema 3.11** Sia (3.9) il sistema linearizzato del sistema dinamico (3.4) nell'intorno del punto di equilibrio  $x_e$ . Se  $\text{Re}(\lambda) \leq -c$ , con  $c > 0$ , per ogni autovalore  $\lambda$  della matrice  $A$ , allora esiste un intorno  $B(x_e)$  di  $x_e$  in  $W$  tale che

- i. per ogni  $\bar{x} \in B(x_e)$ ,  $\varphi(t, \bar{x})$  è definito e in  $B(x_e)$  per ogni  $t \geq 0$ ;
- ii. esiste una costante  $C_0$  tale che

$$|\varphi(t, \bar{x}) - x_e| \leq C_0 e^{-ct/2} |\bar{x} - x_e|, \quad \forall t \geq 0, \forall \bar{x} \in B(x_e),$$

dove  $|\cdot|$  denota la norma euclidea. Quindi in particolare  $x_e$  è asintoticamente stabile.

**Teorema 3.12** Sia (3.9) il sistema linearizzato del sistema dinamico (3.4) nell'intorno del punto di equilibrio  $x_e$ . Se  $\text{Re}(\lambda) > 0$  per qualche autovalore  $\lambda$  della matrice  $A$ , allora  $x_e$  è un punto di equilibrio instabile.

**Definizione 3.13** Siano  $\delta > 0$ ,  $\delta \in \mathbb{R}$ , e  $x_e \in \mathbb{R}^n$ . Allora, per ogni  $n \geq 1$ , l' $n$ -disco aperto di raggio  $\delta$  e centro  $x_e$  è l'insieme

$$D_\delta^n(x_e) = \{x \in \mathbb{R}^n : d(x, x_e) < \delta\},$$

dove  $d$  è la distanza Euclidea.

## 3.2 Modelli epidemiologici

Storicamente, il primo modello matematico in ambito epidemiologico fu formulato, nel 1760, da Daniel Bernoulli nel tentativo di supportare la vaccinazione contro il vaiolo. L'uso di modelli deterministici e probabilistici nell'analisi dello sviluppo di malattie infettive si è grandemente diffuso a partire dal ventesimo secolo. Dopo i primi modelli a tempi discreti di Hamer (1906) e Ross (1911, lavori sulla prevenzione della malaria), negli anni venti, Kermack e McKendrick proposero un modello di tipo differenziale (di tipo SIR) per spiegare la rapida crescita e successiva decrescita del numero di persone infette osservate in alcune epidemie, come la peste (Londra 1665-1666, Bombay 1906) e il colera (Londra 1865).

Nei modelli epidemiologici la popolazione viene suddivisa in sottogruppi, ognuno dei quali rappresenta uno stadio specifico dell'epidemia. Le categorie principali comuni a tutti i modelli sono due:

- I *suscettibili*, indicati con **S**, sono le persone non infette, ma che possono contrarre la malattia.
- Gli *infetti*, indicati con **I**, sono le persone infette, quindi quelle che possono diffondere il contagio per contatto.

Ovviamente il numero di persone suscettibili e infette non è fisso ma cambia con il tempo. Nelle primissime fasi di una epidemia, con la maggior parte delle persone che non hanno ancora contratto la patologia, abbiamo infatti un numero molto elevato di suscettibili, indicato con  $S(t)$ , e un numero di infetti  $I(t)$  molto piccolo. Man mano che l'epidemia si diffonde, decresce il numero di suscettibili mentre il numero di infetti cresce.

Possiamo distinguere diversi modelli epidemiologici a seconda delle differenti categorie in cui vengono suddivisi gli individui della popolazione. Indichiamo con  $N$  la popolazione totale.

Un approccio particolarmente semplice per descrivere la diffusione delle malattie infettive è il modello SI.

Le informazioni che seguono riguardo il modello SI sono state tratte da Wikipedia. Tale modello presuppone che la comunità sia divisa in due categorie, i suscettibili  $S$  e gli infetti  $I$ . In questo modello non è prevista né la guarigione né la perdita di infettività e tutti i suscettibili diventano infetti.

Il tasso di diffusione della malattia dipende statisticamente dal numero di individui

infettivi e in secondo luogo dal numero di individui che possono ancora essere infettati. Secondo questo modello la malattia si diffonderà in tutta la popolazione, dunque

$$N = S(t) + I(t), \forall t \geq 0.$$

Le estensioni al modello SI sono il modello SIS, in cui gli individui possono guarire e riammalarsi, e il modello SIR, in cui gli individui possono diventare immuni alla malattia.

Le informazioni che seguono riguardo il modello SIS sono state tratte da Wikipedia. Il modello SIS distingue due gruppi di persone però, a differenza del modello SI, dopo che la malattia è guarita ogni individuo ritorna immediatamente al gruppo sano e può essere nuovamente infettato. Per una popolazione costante  $N$  si ha

$$N = S(t) + I(t), \forall t \geq 0.$$

Il modello SIS può quindi essere utilizzato per le malattie che hanno le seguenti proprietà:

- i. Dopo che la malattia è guarita, ogni individuo ritorna immediatamente al gruppo sano e può essere nuovamente infettato.
- ii. Gli infetti sono immediatamente contagiosi.
- iii. Ogni gruppo interagisce con la stessa probabilità.

## Modelli SIR

E' stato possibile descrivere tali modelli a partire dalle informazioni riportate su Wikipedia e [MBK] alle quali è stato aggiunto un contributo personale.

Il modello SIR è uno dei modelli epidemiologici più semplici e molti modelli sono derivati da questa forma base. Tale modello prevede di suddividere la popolazione in 3 categorie: i Suscettibili (S), gli Infetti (I) e i Recuperati o Rimossi (**R**). Quest'ultimi sono i pazienti già infettati che non possono più diffondere la malattia perché sono guariti o deceduti.

Nel tempo una parte crescente della popolazione suscettibile diventa infetta e gli infetti diventano rimossi. Per il modello SIR vale quindi la seguente legge:

$$N = S(t) + I(t) + R(t), \quad \forall t \geq 0. \quad (3.10)$$

Naturalmente comprendiamo subito che questi numeri non sono fissi. Nelle primissime fasi di una epidemia, con la maggior parte delle persone che non hanno ancora contratto la patologia, abbiamo infatti un numero molto elevato di suscettibili, un numero piccolo di infettati e un numero R ancora di valore zero. Man mano che l'epidemia procede, la I cresce, la S diminuisce mentre comincia a crescere anche la R.

Il modello SIR può essere utilizzato per le malattie che hanno le seguenti proprietà:

- i. Ogni individuo può essere infettato da un agente patogeno solo una volta e poi diventa immune o muore.
- ii. La popolazione è costante. Le morti dovute alla malattia sono incluse nel gruppo R.
- iii. Gli infetti sono immediatamente contagiosi.
- iv. Ogni gruppo interagisce con la stessa probabilità.

I modelli epidemiologici che studiamo possono avere due tipi di struttura differente: epidemica o endemica. La prima fa riferimento ad un modello in cui la malattia infettiva è localizzata nel tempo, ed è tanto rapida da rendere trascurabili la nascita e le morti sulla popolazione (causate da motivi quali non la malattia infettiva in questione). La seconda invece descrive un modello in cui la malattia permane su tempi lunghi e quindi bisogna considerare le nascite e le morti sulla popolazione.

Possiamo allora distinguere il modello SIR epidemico da quello endemico.

### Modello SIR epidemico

Lo studio fatto relativo al modello SIR epidemico è stato possibile a partire dai dati riportati da Wikipedia ai quali è stato aggiunto un contributo personale.

Per scrivere una formulazione matematica del fenomeno dobbiamo introdurre una serie di equazioni differenziali in modo da indicare i tassi di trasferimento di individui da un compartimento all'altro. Il modo più semplice per descrivere questo tipo di processo infettivo è il seguente:

$$\begin{cases} \frac{dS}{dt} = -\beta SI \\ \frac{dI}{dt} = \beta SI - \gamma I \\ \frac{dR}{dt} = \gamma I \end{cases} \quad (3.11)$$

dove i parametri  $\beta > 0$  e  $\gamma > 0$ , rappresentano rispettivamente il *tasso di infezione* della malattia e il *tasso di recupero* dalla categoria degli infetti, con condizioni iniziali  $S(0) = S_0 > 0$ ,  $I(0) = I_0 > 0$  e  $R(0) = 0$ .

Dalla (3.10) segue che il numero di rimossi è pari a  $R(t) = N - S(t) - I(t)$ .

Il problema fondamentale è quello di cercare di capire se l'infezione sia nelle condizioni tali da diffondersi o se invece resterà un fenomeno ristretto, e a quale valore inizierà il suo declino. Per capirlo abbiamo bisogno di studiare i punti di equilibrio del sistema (3.11) e determinare se sono punti di equilibrio stabili o instabili.

**Teorema 3.14** *Sia  $J_{SIR_{ep}} = \langle -\beta SI, \beta SI - \gamma I, \gamma I \rangle$  l'ideale associato al modello SIR epidemico (3.11) nell'anello  $\mathbb{R}[S, I]$  con l'ordine monomiale  $\text{lex } S > I$ . Allora:*

*i. la base di Groebner ridotta di  $J_{SIR_{ep}}$  è*

$$G_{rid} = \{I\};$$

*ii. l'insieme dei punti di equilibrio è*

$$V_{SIR_{ep}} = \{(S_e, 0, R_e), \forall S_e, R_e \in \mathbb{R}\}.$$

*Dimostrazione.* Per la Def. 3.8 i punti di equilibrio per (3.11) sono tutti i punti che soddisfano il sistema:

$$\begin{cases} -\beta SI = 0 \\ \beta SI - \gamma I = 0 \\ \gamma I = 0. \end{cases} \quad (3.12)$$

Per quanto detto nella sezione 2.7, per risolvere il sistema (3.12) è conveniente determinare la base di Groebner ridotta  $G_{rid}$  dell'ideale  $\langle -\beta SI, \beta SI - \gamma I, \gamma I \rangle$  in  $\mathbb{R}[S, I]$  e, infine, determinare i punti che appartengono a  $V(G_{rid})$ .

Per il criterio di Buchberger (vedasi Teor. 2.41) la base di Groebner per l'ideale  $\langle -\beta SI, \beta SI - \gamma I, \gamma I \rangle$  è  $G = \{-\beta SI, \beta SI - \gamma I, \gamma I\}$ . Infatti:

$$\mathcal{S}(-\beta SI, \beta SI - \gamma I) = \frac{\gamma}{\beta} I = \frac{1}{\beta}(\gamma I), \text{ dunque } \overline{\mathcal{S}(-\beta SI, \beta SI - \gamma I)}^G = 0;$$

$$\mathcal{S}(-\beta SI, \gamma I) = 0;$$

$$\mathcal{S}(\beta SI - \gamma I, \gamma I) = -\frac{\gamma}{\beta} I = -\frac{1}{\beta}(\gamma I) \text{ e dunque } \overline{\mathcal{S}(\beta SI - \gamma I, \gamma I)}^G = 0.$$

In particolare, osserviamo che

$$\text{LT}(-\beta SI) = \frac{-\beta S}{\gamma} \text{LT}(\gamma I)$$

e

$$\text{LT}(\beta SI - \gamma I) = \frac{\beta S}{\gamma} \text{LT}(\gamma I),$$

dunque per il Lemma 2.45 possiamo eliminare dalla base  $G$  i polinomi  $-\beta SI$  e  $\beta SI - \gamma I$ . Inoltre, rendendo monico  $\gamma I$ , per la Def. 2.50 si ha che  $G_{rid} = \{I\}$  è la base di Groebner ridotta per  $\langle -\beta SI, \beta SI - \gamma I, \gamma I \rangle$ .

E' facile vedere che i punti in  $V(G_{rid})$ , ovvero i punti di equilibrio per il sistema (3.11) appartengono al piano  $I = 0$  e dunque sono infiniti punti del tipo  $(S_e, 0, R_e)$ .

□

Ci occupiamo ora della stabilità dei punti di equilibrio trovati nel teorema precedente.

**Teorema 3.15** *Sia  $V_{SIR_{ep}} = \{(S_e, 0, R_e), \forall S_e, R_e \in \mathbb{R}\}$  la varietà associata al modello SIR (vedasi Teor. 3.14). Allora il punto di equilibrio  $(S_e, 0, R_e)$ , con  $S_e \geq 0, R_e \geq 0$ , è:*

*i. instabile se  $S_e > \frac{\gamma}{\beta}$ ;*

*ii. stabile (ma non asintoticamente stabile) se  $S_e \leq \frac{\gamma}{\beta}$ .*

*Dimostrazione.* Per poter applicare i teoremi di stabilità e quindi determinare se i punti di equilibrio sono stabili o instabili dobbiamo trovare la matrice Jacobiana relativa al sistema (3.11), calcolarla nei punti di equilibrio e determinarne gli autovalori.

Osserviamo che la matrice Jacobiana associata al sistema (3.11) è

$$J(S, I, R) = \begin{pmatrix} -\beta I & -\beta S & 0 \\ \beta I & \beta S - \gamma & 0 \\ 0 & \gamma & 0 \end{pmatrix}. \quad (3.13)$$

Calcoliamo allora la matrice Jacobiana (3.13) nei punti di equilibrio e otteniamo

$$J(S_e, 0, R_e) = \begin{pmatrix} 0 & -\beta S_e & 0 \\ 0 & \beta S_e - \gamma & 0 \\ 0 & \gamma & 0 \end{pmatrix}. \quad (3.14)$$

Gli autovalori associati alla matrice (3.14) sono gli zeri del polinomio

$$p(\lambda) = \begin{vmatrix} -\lambda & -\beta S_e & 0 \\ 0 & \beta S_e - \gamma - \lambda & 0 \\ 0 & \gamma & -\lambda \end{vmatrix},$$

dunque  $p(\lambda) = \lambda^2(\beta S_e - \gamma - \lambda)$ .

Se  $S_e \neq \frac{\gamma}{\beta}$ , gli autovalori sono  $\lambda_1 = 0$  e  $\lambda_2 = \beta S_e - \gamma$ . Altrimenti, la matrice (3.14) ammette come unico autovalore  $\lambda_1 = 0$ .

Se  $S_e > \frac{\gamma}{\beta}$  allora  $\lambda_2 > 0$  e per il Teor. 3.12 i punti di equilibrio sono instabili.

Se  $S_e < \frac{\gamma}{\beta}$  allora  $\lambda_2 < 0$ , ma ciò non è sufficiente per concludere che i punti di equilibrio sono stabili in quanto la matrice (3.14) ammette autovalore nullo con molteplicità algebrica 2. Quindi non è possibile applicare i teoremi di stabilità e non è possibile concludere in questo modo. Per questo motivo studiamo la stabilità dei punti di equilibrio del sistema (3.11) facendo altri calcoli.

Sia  $x_e = (S_e, 0, R_e)$  con  $0 \leq S_e < \frac{\gamma}{\beta}$  e  $R_e \geq 0$ .

$\forall \epsilon > 0$ , sia  $\delta > 0$  tale che:

- i.  $\gamma - \beta S_e - \beta \delta > 0$ ;
- ii.  $\delta < \frac{\epsilon}{6}$ ;
- iii.  $S_e - S_e e^{\frac{-\beta \delta}{\gamma - \beta S_e - \beta \delta}} + \delta e^{\frac{-\beta \delta}{\gamma - \beta S_e - \beta \delta}} < \frac{\epsilon}{6}$ .

Osserviamo che la condizione iii. è possibile perchè

$$\lim_{\delta \rightarrow 0^+} \left( S_e - S_e e^{\frac{-\beta \delta}{\gamma - \beta S_e - \beta \delta}} + \delta e^{\frac{-\beta \delta}{\gamma - \beta S_e - \beta \delta}} \right) = 0.$$

Sia  $\bar{x} = (\bar{S}, \bar{I}, \bar{R}) \in D_\delta^3(x_e) \cap \{S \geq 0, I \geq 0, R \geq 0\}$  il dato iniziale.

Dal sistema (3.11) segue che  $\frac{dS}{dt} < 0$  e dunque  $S$  decresce sempre. Di conseguenza, si ha che  $S(t) \leq \bar{S}$  per ogni  $t \geq 0$ .

Dall'equazione  $\frac{dI}{dt} = (\beta S - \gamma)I$  del sistema (3.11) segue che

$$\frac{dI}{dt} \leq 0 \iff S \leq \frac{\gamma}{\beta}.$$

Ora, per la condizione i. sappiamo che  $\delta < \frac{\gamma}{\beta} - S_e$  quindi

$$S(t) \leq \bar{S} < S_e + \delta < \frac{\gamma}{\beta}.$$

Per quanto detto,  $I(t)$  decresce quindi dalla condizione ii. segue che

$$0 \leq I(t) \leq \bar{I} < \delta < \frac{\epsilon}{6}. \quad (3.15)$$

Ricordiamo che  $S(t)$  è decrescente quindi  $S(t) \leq \bar{S}$ .

Sia  $a := \gamma - \beta S_e - \beta \delta > 0$  per la condizione i. Allora

$$\beta S - \gamma \leq \beta \bar{S} - \gamma < \beta S_e + \beta \delta - \gamma = -a.$$

e, di conseguenza,

$$\frac{dI}{dt} = (\beta S - \gamma)I \leq -aI. \quad (3.16)$$

Integrando la (3.16) otteniamo

$$I(t) \leq \bar{I}e^{-at}. \quad (3.17)$$

L'equazione  $\frac{dS}{dt} = -\beta SI$  del sistema (3.11) si può riscrivere nel modo seguente:  $\frac{dS}{dt} \frac{1}{S} = -\beta I$ , che integrata dà

$$\log(S(t)) = \log(\bar{S}) - H(t)$$

o, equivalentemente,

$$S(t) = \bar{S}e^{-H(t)} \quad (3.18)$$

dove

$$H(t) := \beta \int_0^t I(s) ds,$$

( $I(t)$  è integrabile, decadendo almeno esponenzialmente). Poiché  $I$  e  $S$  decrescono e sono sempre positivi, la traiettoria rimane sempre in un compatto, quindi (per

il teorema del prolungamento <sup>1)</sup>  $S(t)$  è definita per ogni  $t$ .  
Ora, utilizzando la (3.17) otteniamo

$$H(t) := \beta \int_0^t I(s) ds \leq \beta \int_0^t \bar{I} e^{-as} ds = \frac{\beta \bar{I}}{a} (1 - e^{-at}) < \frac{\beta \bar{I}}{a} \leq \frac{\beta \delta}{a}. \quad (3.19)$$

Dall'equazione (3.18) segue che

$$S(t) = \bar{S} e^{-H(t)} \geq \bar{S} e^{-\frac{\beta \delta}{a}} \geq (S_e - \delta) e^{-\frac{\beta \delta}{a}}. \quad (3.20)$$

Ne segue che

$$|S(t) - S_e| < \frac{\epsilon}{6}. \quad (3.21)$$

Infatti, se  $S(t) \geq S_e$ , allora per la condizione ii. si ha che

$$|S(t) - S_e| = S(t) - S_e \leq \bar{S} - S_e < \delta < \frac{\epsilon}{6}.$$

Se  $S(t) \leq S_e$ , allora per la (3.20) e per la condizione iii.

$$|S(t) - S_e| = S_e - S(t) \leq S_e - (S_e - \delta) e^{-\frac{\beta \delta}{a}} = S_e - S_e e^{\frac{-\beta \delta}{\gamma - \beta S_e - \beta \delta}} + \delta e^{\frac{-\beta \delta}{\gamma - \beta S_e - \beta \delta}} < \frac{\epsilon}{6}.$$

Osserviamo che

$$|R(t) - R_e| < \frac{5}{6} \epsilon.$$

Infatti,

$$|R(t) - R_e| \leq |R(t) - \bar{R}| + |\bar{R} - R_e|. \quad (3.22)$$

Consideriamo  $|R(t) - \bar{R}|$ . Dall'equazione (3.10) segue che  $R(t) = N - S(t) - I(t)$ .  
Quindi

$$\begin{aligned} R(t) - \bar{R} &= N - S(t) - I(t) - (N - \bar{S} - \bar{I}) = \bar{S} - S(t) + \bar{I} - I(t) = \\ &= \bar{S} - S_e + S_e - S(t) + \bar{I} - I(t). \end{aligned}$$

Di conseguenza, per le (3.15), (3.21) e per la condizione ii.

$$|R(t) - \bar{R}| \leq |\bar{S} - S_e| + |S_e - S(t)| + |\bar{I}| + |I(t)| \leq \frac{2}{3} \epsilon.$$

Inoltre,

$$|\bar{R} - R_e| < \delta < \frac{\epsilon}{6}.$$

---

<sup>1)</sup>Vedasi [G, Cap.3, Cor. 12.23]

Quindi la (3.22) diventa

$$|R(t) - R_e| \leq |R(t) - \bar{R}| + |\bar{R} - R_e| < \frac{5\epsilon}{6}. \quad (3.23)$$

Ma allora dalle (3.15), (3.21) e (3.23) segue che

$$\begin{aligned} \|x(t) - x_e\| &= \sqrt{(S(t) - S_e)^2 + I(t)^2 + (R(t) - R_e)^2} \\ &\leq \sqrt{\left(\frac{\epsilon}{6}\right)^2 + \left(\frac{\epsilon}{6}\right)^2 + \left(\frac{5\epsilon}{6}\right)^2} = \frac{\sqrt{3}}{2}\epsilon < \epsilon. \end{aligned}$$

In conclusione, i punti di equilibrio  $x_e = (S_e, 0, R_e)$ , con  $S_e \geq 0$  e  $R_e \geq 0$ , sono stabili se  $S_e < \frac{\gamma}{\beta}$ .

Il punto di equilibrio  $x_e$  non può essere asintoticamente stabile se  $S_e < \frac{\gamma}{\beta}$ . Infatti se prendiamo un  $\bar{x} \in D_\delta^3(x_e) \cap \{S \geq 0, I \geq 0, R \geq 0\}$  tale che  $\bar{S} < S_e$  allora  $S(t) < S_e$  per ogni  $t$  quindi non è possibile che

$$\lim_{t \rightarrow \infty} x(t) = x_e.$$

Consideriamo ora il caso in cui  $S_e = \frac{\gamma}{\beta}$  e sia quindi  $x_e = (\frac{\gamma}{\beta}, 0, R_e)$  con  $R_e \geq 0$ . Sia  $P_e = (\frac{\gamma}{\beta}, 0)$  la sua proiezione.

Sia  $\epsilon > 0$  e sia  $\delta > 0$  tale che:

- i.  $\delta < \frac{\gamma}{\beta}$ ;
- ii.  $\delta < 1$ ;
- iii.  $\delta < \frac{\epsilon^2}{180}$ ;
- iv.  $\delta < \frac{\epsilon}{12}$ ;
- v.  $[\frac{\gamma}{\beta} - (\frac{\gamma}{\beta} - \sqrt{\delta})e^{-\sqrt{\delta}}]^2 + \delta^2 < (\frac{\epsilon}{6})^2$ .

La condizione v. è possibile perchè

$$\lim_{\delta \rightarrow 0^+} ([\frac{\gamma}{\beta} - (\frac{\gamma}{\beta} - \sqrt{\delta})e^{-\sqrt{\delta}}]^2 + \delta^2) = 0.$$

Sia  $\bar{x} = (\bar{S}, \bar{I}, \bar{R}) \in D_\delta^3(x_e) \cap \{S \geq 0, I \geq 0, R \geq 0\}$ . Dimostriamo che

$$\|x(t) - x_e\| < \epsilon, \quad \forall t \geq 0. \quad (3.24)$$

Sia  $\bar{P} = (\bar{S}, \bar{I})$ . Allora

$$d_2(\bar{P}, P_e) = \sqrt{(\bar{S} - \frac{\gamma}{\beta})^2 + \bar{I}^2} \leq \sqrt{(\bar{S} - \frac{\gamma}{\beta})^2 + \bar{I}^2 + (\bar{R} - R_e)^2} < \delta.$$

Dunque

$$\bar{I} \leq \sqrt{(\bar{S} - \frac{\gamma}{\beta})^2 + \bar{I}^2} < \delta$$

e  $\bar{P} \in D_\delta^2(P_e) \cap \{S \geq 0, I \geq 0\}$ .

Sia  $x(t) = (S(t), I(t), R(t))$  la soluzione di (3.11) con dato iniziale  $\bar{x}$  e sia  $y(t) = (S(t), I(t))$ .

Sia  $\epsilon' = \frac{\epsilon}{6}$ , quindi  $\delta < \epsilon'$ , per la condizione iv.

Asseriamo che

$$y(t) \in D_{\epsilon'}^2(P_e) \cap \{S \geq 0, I \geq 0\}, \quad \forall t \geq 0.$$

Supponiamo per assurdo che

$$\exists t \geq 0 : y(t) \notin D_{\epsilon'}^2(P_e) \cap \{S \geq 0, I \geq 0\}. \quad (3.25)$$

Dato che  $y(0) = (S(0), I(0)) = (\bar{S}, \bar{I}) = \bar{P} \in D_\delta^2(P_e) \subseteq D_{\epsilon'}^2(P_e)$ , allora  $\exists t > 0$  tale che  $y(t) \notin D_{\epsilon'}^2(P_e) \cap \{S \geq 0, I \geq 0\}$ . Sia  $\epsilon'' = \frac{\epsilon'}{\sqrt{2}}$ . Allora  $\epsilon'' < \epsilon'$ . Dunque  $\exists t > 0$  tale che  $y(t) \notin D_{\epsilon''}^2(P_e) \cap \{S \geq 0, I \geq 0\}$ .

Dal sistema (3.11) segue che  $\frac{dS}{dt} = -\beta SI \leq 0$ , quindi  $S(t)$  è decrescente  $\forall t \geq 0$ . Inoltre, da  $\frac{dI}{dt} = (\beta S - \gamma)I$  segue che  $I(t)$  è decrescente per ogni  $t$  tale che  $S(t) \leq \frac{\gamma}{\beta}$ .

Sia

$$W(t) = (S(t) - \frac{\gamma}{\beta})^2 + I(t)^2.$$

Dal sistema (3.11) segue che

$$W'(t) = -2(\beta S(t) - \gamma)I(t)(S(t) - I(t)).$$

Allora  $W'(t) \leq 0$  (e quindi  $W(t)$  è decrescente) se  $S(t) \geq I(t)$  e  $S(t) \geq \frac{\gamma}{\beta}$ .

Ora mostriamo che

$$\exists t \geq 0 : S(t) < \frac{\gamma}{\beta}. \quad (3.26)$$

Se per assurdo fosse  $S(t) \geq \frac{\gamma}{\beta}, \forall t \geq 0$ , allora dobbiamo distinguere due casi:

a.  $I(t) < \delta \forall t \geq 0$ ;

b.  $\exists t \geq 0 : I(t) \geq \delta$ .

Nel caso a. si ha che  $S(t) - I(t) > \frac{\gamma}{\beta} - \delta > 0 \forall t \geq 0$ . Quindi  $W(t)$  è decrescente  $\forall t \geq 0$ . Allora

$$d_2(y(t), P_e) = \sqrt{W(t)} \leq \sqrt{W(0)} = d_2(\bar{P}, P_e) < \delta < \epsilon'.$$

Questo implica che  $y(t) \in D_e^2(P_e) \cap \{S \geq 0, I \geq 0\}$ ,  $\forall t \geq 0$ , contraddicendo (3.25).

Nel caso b. essendo  $I(0) = \bar{I} < \delta$ , si ha che  $\exists t_1 > 0$  tale che  $I(t_1) = \delta, I(t) < \delta \forall t \in [0, t_1)$ . Ma allora  $\forall t \in [0, t_1]$  si ha:  $S(t) \geq \frac{\gamma}{\beta}$  e  $S(t) - I(t) \geq \frac{\gamma}{\beta} - \delta > 0$ . Perciò  $W(t)$  è decrescente in  $[0, t_1]$ . Quindi

$$\delta \leq \sqrt{(S(t_1) - \frac{\gamma}{\beta})^2 + I(t_1)^2} = \sqrt{W(t_1)} \leq \sqrt{W(0)} = d_2(\bar{P}, P_e) < \delta,$$

contraddizione.

Allora la (3.26) è dimostrata.

Per poter dimostrare la (3.24) dobbiamo distinguere i seguenti casi:  $\bar{S} > \frac{\gamma}{\beta}$  e  $\bar{S} \leq \frac{\gamma}{\beta}$ .

Sia  $\bar{S} > \frac{\gamma}{\beta}$ .

Dato che  $S(0) = \bar{S} > \frac{\gamma}{\beta}$ , per la (3.26),  $\exists t_0 > 0$  tale che  $S(t_0) = \frac{\gamma}{\beta}, S(t) > \frac{\gamma}{\beta} \forall t \in [0, t_0)$ .

Ne segue che

$$I(t) < \delta, \forall t \geq 0. \quad (3.27)$$

Infatti se per assurdo  $\exists t \in [0, t_0]$  tale che  $I(t) \geq \delta$ , essendo  $I(0) = \bar{I} < \delta$  allora  $\exists t_2 \in (0, t_0]$  tale che  $I(t_2) = \delta, I(t) < \delta \forall t \in [0, t_2)$ . Ma se  $t \in [0, t_2]$  si ha:  $S(t) \geq \frac{\gamma}{\beta}$  dato che  $[0, t_2] \subseteq [0, t_0]$  e  $I(t) \leq \delta < \frac{\gamma}{\beta} \leq S(t)$ . Pertanto  $W(t)$  è decrescente in  $[0, t_2]$  e

$$\delta \leq \sqrt{(S(t_2) - \frac{\gamma}{\beta})^2 + I(t_2)^2} = \sqrt{W(t_2)} \leq \sqrt{W(0)} = d_2(\bar{P}, P_e) < \delta,$$

contraddizione.

Quindi  $I(t) < \delta \forall t \in [0, t_0]$ .

Se  $t \geq t_0$  allora, essendo  $S(t)$  decrescente,  $S(t) \leq S(t_0) = \frac{\gamma}{\beta}$ . Di conseguenza,  $I(t)$  è decrescente per  $t \geq t_0$ . Dunque

$$I(t) \leq I(t_0) < \delta,$$

e la (3.27) è dimostrata.

Osserviamo che

$$t \in [0, t_0] \implies y(t) \in D_{\epsilon'}^2(P_e) \subseteq D_{\epsilon'}^2(P_e). \quad (3.28)$$

Infatti, sia  $t \in [0, t_0]$ . Allora si ha che  $\frac{\gamma}{\beta} \leq S(t) \leq \bar{S}$  e  $0 \leq I(t) < \delta$  per (3.27). Dunque

$$d_2(y(t), P_e) = \sqrt{(S(t) - \frac{\gamma}{\beta})^2 + I(t)^2} \leq \sqrt{(\bar{S} - \frac{\gamma}{\beta})^2 + I(t)^2} < \sqrt{2}\delta < \epsilon''.$$

Ma, allora,  $y(t) \in D_{\epsilon'}^2(P_e) \cap \{S \geq 0, I \geq 0\}$  per  $t \in [0, t_0]$  e la (3.28) è dimostrata.

Ora, da (3.25) segue che  $\exists t \geq 0 : y(t) \notin D_{\epsilon'}^2(P_e) \cap \{S \geq 0, I \geq 0\}$ . Quindi per la (3.28) si ha che

$$\exists t \geq t_0 : y(t) \notin D_{\epsilon'}^2(P_e) \cap \{S \geq 0, I \geq 0\}. \quad (3.29)$$

Asseriamo inoltre che

$$\exists \bar{t} \geq t_0 : S(\bar{t}) \leq \frac{\gamma}{\beta} - \epsilon''. \quad (3.30)$$

Infatti se per assurdo fosse  $S(t) > \frac{\gamma}{\beta} - \epsilon'' \forall t \geq t_0$ , allora  $S(t) \leq S(t_0) = \frac{\gamma}{\beta}$  e quindi  $0 \leq \frac{\gamma}{\beta} - S(t) < \epsilon''$  e  $I(t) < \delta < \epsilon''$  per (3.27). Quindi

$$d_2(y(t), P_e) = \sqrt{(\frac{\gamma}{\beta} - S(t))^2 + I(t)^2} \leq \sqrt{2}\epsilon'' = \epsilon'.$$

Allora  $y(t) \in D_{\epsilon'}^2(P_e) \cap \{S \geq 0, I \geq 0\}$  per  $t \geq t_0$ . Da (3.28) segue che  $y(t) \in D_{\epsilon'}^2(P_e) \cap \{S \geq 0, I \geq 0\} \forall t \geq 0$ , contraddicendo (3.25). Allora la (3.30) è dimostrata.

Sia ora  $d = \sqrt{\delta}$ . Mostriamo ora che

$$\exists t_3 > t_0 : S(t_3) = \frac{\gamma}{\beta} - d, I(t_3) \leq \delta \text{ e } S(t) \geq \frac{\gamma}{\beta} - d, \text{ per } t_0 \leq t \leq t_3. \quad (3.31)$$

Infatti, dato che  $S(t_0) = \frac{\gamma}{\beta}$ , se fosse  $S(t) > \frac{\gamma}{\beta} - d, \forall t \geq t_0$ , preso  $\bar{t} \geq t_0$  tale che  $S(\bar{t}) \leq \frac{\gamma}{\beta} - \epsilon''$  (notiamo che  $\bar{t}$  esiste per (3.30)), allora  $\frac{\gamma}{\beta} - d < S(\bar{t}) \leq \frac{\gamma}{\beta} - \epsilon''$ . Di conseguenza,  $\frac{\epsilon}{6\sqrt{2}} = \epsilon'' < d = \sqrt{\delta}$ . Allora  $\delta > \frac{\epsilon^2}{72}$ . Ma, per la condizione iii. si ha che  $\delta < \frac{\epsilon^2}{180} < \frac{\epsilon^2}{72} < \delta$ , contraddizione. Inoltre,  $I(t_3) \leq \delta$  segue da (3.27). La (3.31) è così dimostrata.

Supponiamo ora  $t \geq t_3$ . Per la (3.31),  $S(t) \leq S(t_3) = \frac{\gamma}{\beta} - d$ . Dal sistema (3.11) ricaviamo

$$\frac{dI}{dt} = (\beta S - \gamma)I \leq (\beta S(t_3) - \gamma)I = -\beta d I.$$

Ponendo  $K(t) = I(t_3)e^{-\beta d(t-t_3)}$  si ha che

$$\frac{dI}{dt} \frac{1}{I} \leq -\beta d$$

e integrando otteniamo

$$\int_{t_3}^t \frac{dI}{I} \leq \int_{t_3}^t -\beta d ds$$

o, equivalentemente,

$$I(t) \leq I(t_3)e^{-\beta d(t-t_3)} = K(t). \quad (3.32)$$

Ora, dal sistema (3.11)  $\frac{dS}{dt} = -\beta SI$  e quindi, per la (3.32),  $\frac{dS}{dt} \frac{1}{S} \geq -\beta K(t)$  per  $t \geq t_3$ . Allora,

$$\int_{t_3}^t \frac{dS}{S} \geq -\beta \int_{t_3}^t K(s) ds.$$

Posto  $J(t) = \int_{t_3}^t K(s) ds$ , si ottiene che

$$S(t) \geq S(t_3)e^{-\beta J(t)}. \quad (3.33)$$

Inoltre, per (3.27) si ha

$$J(t) = \int_{t_3}^t K(s) ds = \int_{t_3}^t I(t_3)e^{-\beta d(s-t_3)} ds = \frac{I(t_3)}{\beta d} (1 - e^{-\beta d(t-t_3)}) \leq \frac{\delta}{\beta d} = \frac{d}{\beta}.$$

Di conseguenza, per  $t \geq t_3$  la (3.33) diventa

$$S(t) \geq S(t_3)e^{-\beta J(t)} \geq S(t_3)e^{-d} = \left(\frac{\gamma}{\beta} - d\right)e^{-d}.$$

Quindi, per  $t \geq t_3$

$$\left(\frac{\gamma}{\beta} - d\right)e^{-d} \leq S(t) \leq S(t_3) = \frac{\gamma}{\beta} - d.$$

Allora,

$$0 < d \leq \frac{\gamma}{\beta} - S(t) \leq \frac{\gamma}{\beta} - \left(\frac{\gamma}{\beta} - d\right)e^{-d}.$$

Segue che, da (3.27) e dalla condizione v. si ha per  $t \geq t_3$

$$d_2(y(t), P_e) = \sqrt{(S(t) - \frac{\gamma}{\beta})^2 + I(t)^2} \leq \sqrt{\left[\frac{\gamma}{\beta} - \left(\frac{\gamma}{\beta} - d\right)e^{-d}\right]^2 + \delta^2} < \epsilon'.$$

In conclusione  $y(t) \in D_e^2(P_e) \cap \{S \geq 0, I \geq 0\}$ , per  $t \geq t_3$ .

Per (3.25) e (3.28),  $\exists t \in (t_0, t_3) : y(t) \notin D_e^2(P_e) \cap \{S \geq 0, I \geq 0\}$ . Ma, per  $t \in (t_0, t_3)$

$$\frac{\gamma}{\beta} - d \leq S(t) \leq S(t_0) = \frac{\gamma}{\beta}.$$

Quindi  $0 \leq \frac{\gamma}{\beta} - S(t) \leq d$ . Dunque, usando (3.27) si ha che

$$d_2(y(t), P_e) = \sqrt{(S(t) - \frac{\gamma}{\beta})^2 + I(t)^2} \leq \sqrt{d^2 + \delta^2} = \sqrt{\delta + \delta^2} < \sqrt{2\delta} < \epsilon'.$$

Allora  $y(t) \in D_e^2(P_e) \cap \{S \geq 0, I \geq 0\}$ , contraddizione.

In conclusione,  $y(t) \in D_e^2(P_e) \cap \{S \geq 0, I \geq 0\}$ ,  $\forall t \geq 0$ , per dati iniziali con  $\bar{S} > \frac{\gamma}{\beta}$ .

Ora, sia  $\bar{S} \leq \frac{\gamma}{\beta}$ .

Allora  $S(t) \leq \bar{S} \leq \frac{\gamma}{\beta}$ ,  $\forall t \geq 0$ .

Dunque  $\beta S - \gamma \leq 0$ ,  $\forall t \geq 0$ . Quindi  $I(t)$  è decrescente e  $I(t) \leq \bar{I} < \delta$ ,  $\forall t \geq 0$ .

Osserviamo che  $\bar{S} > \frac{\gamma}{\beta} - d$ . Infatti se fosse  $\bar{S} \leq \frac{\gamma}{\beta} - d$  avremmo  $\frac{\gamma}{\beta} - \delta < \bar{S} \leq \frac{\gamma}{\beta} - d$ . Quindi  $d < \delta$  ovvero  $\delta > 1$ , contraddicendo la condizione ii.

Allora

$$\exists t_4 > 0 : S(t_4) = \frac{\gamma}{\beta} - d, \quad I(t_4) \leq \delta \text{ e } S(t) > \frac{\gamma}{\beta} - d, \text{ per } 0 \leq t < t_4. \quad (3.34)$$

Infatti, intanto vale la seguente

$$\exists \bar{t} \geq 0 : S(\bar{t}) \leq \frac{\gamma}{\beta} - \epsilon''. \quad (3.35)$$

Se per assurdo fosse  $S(t) > \frac{\gamma}{\beta} - \epsilon'' \forall t \geq 0$ , allora  $S(t) \leq S(0) = \bar{S} \leq \frac{\gamma}{\beta}$ . Quindi  $0 \leq \frac{\gamma}{\beta} - S(t) < \epsilon''$  e  $I(t) < \delta < \epsilon''$ . Dunque,  $\forall t \geq 0$ ,

$$d_2(y(t), P_e) = \sqrt{(\frac{\gamma}{\beta} - S(t))^2 + I(t)^2} \leq \sqrt{2\epsilon''} = \epsilon'.$$

Allora  $y(t) \in D_e^2(P_e) \cap \{S \geq 0, I \geq 0\}$  per  $\forall t \geq 0$ , contraddicendo (3.25). Quindi la (3.35) è dimostrata.

Ora, se fosse  $S(t) > \frac{\gamma}{\beta} - d$ ,  $\forall t \geq 0$ , preso  $\bar{t} \geq 0$  tale che  $S(\bar{t}) \leq \frac{\gamma}{\beta} - \epsilon''$  ( $\bar{t}$  esiste per (3.35)), allora  $\frac{\gamma}{\beta} - d < S(\bar{t}) \leq \frac{\gamma}{\beta} - \epsilon''$ . Di conseguenza,  $\frac{\epsilon}{6\sqrt{2}} = \epsilon'' < d = \sqrt{\delta}$ . Allora

$\delta > \frac{\epsilon^2}{72}$ . Ma, per la condizione iii. si ha che  $\delta < \frac{\epsilon^2}{180} < \frac{\epsilon^2}{72} < \delta$ , contraddizione.

La (3.34) è così dimostrata.

Supponiamo ora  $t \geq t_4$ . Per la (3.34),  $S(t) \leq S(t_4) = \frac{\gamma}{\beta} - d$ . Dal sistema (3.11) ricaviamo

$$\frac{dI}{dt} = (\beta S - \gamma)I \leq (\beta S(t_4) - \gamma)I = -\beta d I.$$

Ponendo  $K(t) = I(t_4)e^{-\beta d(t-t_4)}$  si ha che

$$I(t) \leq I(t_4)e^{-\beta d(t-t_4)} = K(t). \quad (3.36)$$

Ora, per la (3.36), posto  $J(t) = \int_{t_4}^t K(s)ds$ ,

$$S(t) \geq S(t_4)e^{-\beta J(t)}. \quad (3.37)$$

Inoltre,

$$J(t) = \int_{t_4}^t K(s)ds = \int_{t_4}^t I(t_4)e^{-\beta d(s-t_4)}ds = \frac{I(t_4)}{\beta d}(1 - e^{-\beta d(t-t_4)}) \leq \frac{\delta}{\beta d} = \frac{d}{\beta}.$$

Di conseguenza, per  $t \geq t_4$  la (3.37) diventa

$$S(t) \geq S(t_4)e^{-\beta J(t)} \geq S(t_4)e^{-d} = \left(\frac{\gamma}{\beta} - d\right)e^{-d}.$$

Quindi, per  $t \geq t_4$

$$\left(\frac{\gamma}{\beta} - d\right)e^{-d} \leq S(t) \leq S(t_4) = \frac{\gamma}{\beta} - d.$$

Allora,

$$0 < d \leq \frac{\gamma}{\beta} - S(t) \leq \frac{\gamma}{\beta} - \left(\frac{\gamma}{\beta} - d\right)e^{-d}.$$

Segue che, dalla condizione v. si ha

$$d_2(y(t), P_e) = \sqrt{\left(S(t) - \frac{\gamma}{\beta}\right)^2 + I(t)^2} \leq \sqrt{\left[\frac{\gamma}{\beta} - \left(\frac{\gamma}{\beta} - d\right)e^{-d}\right]^2 + \delta^2} < \epsilon'.$$

In conclusione  $y(t) \in D_{\epsilon'}^2(P_e) \cap \{S \geq 0, I \geq 0\}$ , per  $t \geq t_4$ .

Per (3.25),  $\exists t \in (0, t_4) : y(t) \notin D_{\epsilon'}^2(P_e) \cap \{S \geq 0, I \geq 0\}$ . Ma, per  $t \in (0, t_4)$

$$\frac{\gamma}{\beta} - d < S(t) \leq S(0) = \bar{S} \leq \frac{\gamma}{\beta}.$$

Quindi  $0 \leq \frac{\gamma}{\beta} - S(t) < d$ . Dunque

$$d_2(y(t), P_e) = \sqrt{\left(S(t) - \frac{\gamma}{\beta}\right)^2 + I(t)^2} \leq \sqrt{d^2 + \delta^2} = \sqrt{\delta + \delta^2} < \sqrt{2\delta} < \epsilon'.$$

Allora  $y(t) \in D_e^2(P_e) \cap \{S \geq 0, I \geq 0\}$ , contraddizione.

In conclusione,  $y(t) \in D_e^2(P_e) \cap \{S \geq 0, I \geq 0\}$ ,  $\forall t \geq 0$ , per dati iniziali con  $\bar{S} \leq \frac{\gamma}{\beta}$ .

Quindi  $\forall t \geq 0$ ,

$$|S(t) - \frac{\gamma}{\beta}| \leq \sqrt{(S(t) - \frac{\gamma}{\beta})^2 + I(t)^2} < \epsilon'$$

e

$$|I(t)| \leq \sqrt{(S(t) - \frac{\gamma}{\beta})^2 + I(t)^2} < \epsilon'$$

Ora,

$$\begin{aligned} |R(t) - R_e| &\leq |R(t) - \bar{R}| + |\bar{R} - R_e| < |(N - S(t) - I(t)) - (N - \bar{S} - \bar{I})| + \delta \leq \\ &\leq |S(t) - \bar{S}| + |I(t) - \bar{I}| + \delta \leq |S(t) - \frac{\gamma}{\beta}| + |\frac{\gamma}{\beta} - \bar{S}| + |I(t)| + |\bar{I}| + \delta < 5\epsilon' \end{aligned}$$

Allora, per ogni punto  $x(t)$  si ha che

$$\begin{aligned} \|x(t) - x_e\| &= \sqrt{(S(t) - \frac{\gamma}{\beta})^2 + I(t)^2 + (R(t) - R_e)^2} < \\ &< \sqrt{(\epsilon')^2 + (\epsilon')^2 + 25(\epsilon')^2} = \sqrt{27}\epsilon' = \frac{\sqrt{27}}{6}\epsilon < \epsilon. \end{aligned}$$

In conclusione,  $x(t) \in D_e^3(P_e) \cap \{S \geq 0, I \geq 0, R_e\}$ ,  $\forall t \geq 0$ . Pertanto il punto  $(\frac{\gamma}{\beta}, 0, R_e)$  con  $R_e \geq 0$  è un punto di equilibrio stabile.

Come nel caso  $S_e > \frac{\gamma}{\beta}$ , si ha che  $x_e$  non può essere asintoticamente stabile.

□

**Osservazione 3.16** Sia  $\bar{x} = (\bar{S}, \bar{I}, \bar{R})$  con  $\bar{S} \geq 0$ ,  $\bar{I} \geq 0$  e  $\bar{R} \geq 0$ . Definiamo

$$R_0 = \frac{\beta\bar{S}}{\gamma}.$$

Se  $R_0 \leq 1$  allora  $S \leq \bar{S} \leq \frac{\gamma}{\beta}$  dunque  $I$  è decrescente.

Se  $R_0 > 1$  abbiamo che  $\bar{S} > \frac{\gamma}{\beta}$ , quindi  $I$  è crescente finchè  $S > \frac{\gamma}{\beta}$ .

Ma  $S$  è decrescente quindi si arriverà ad un tempo  $T$  tale che  $S(T) = \frac{\gamma}{\beta}$ ,  $\frac{dS}{dt} < 0$  e  $\frac{dI}{dt} = 0$ . Quindi, quale che sia il dato iniziale  $(\bar{S}, \bar{I}, \bar{R})$  la traiettoria entra nella regione  $S < \frac{\gamma}{\beta}$ .

Quindi o  $\bar{S} < \frac{\gamma}{\beta}$  o comunque, se  $\bar{S} > \frac{\gamma}{\beta}$ , esiste un tempo  $T_1$  tale che  $S(T_1) < \frac{\gamma}{\beta}$ . Dunque,  $I(t)$  diminuisce in ogni caso e ci si può limitare ai dati iniziali tali che  $\bar{S} < \frac{\gamma}{\beta}$ . A livello pratico però, è differente se  $\bar{S} < \frac{\gamma}{\beta}$  o  $\bar{S} > \frac{\gamma}{\beta}$ . Infatti nel primo

caso,  $I(t)$  diminuisce, mentre nel caso  $\bar{S} > \frac{\gamma}{\beta}$ ,  $I(t)$  prima aumenta e poi diminuisce. In particolare, se  $\bar{S} \gg \frac{\gamma}{\beta}$  oppure  $\bar{I} \gg 0$ , allora  $I(t)$  aumenta moltissimo prima di diminuire. Dunque il numero dei malati diminuisce comunque, però in un caso diminuisce perché si ammalano sempre meno persone e nell'altro perché si ammalano tutti e a un certo punto o guariscono o muoiono, così che alla fine in ogni caso non ci sono più malati.

Ora, vogliamo determinare il punto limite.

Dall'equazione (3.17) e ricordando che  $I(t)$  è positivo, se ne deduce che

$$\lim_{t \rightarrow \infty} I(t) = 0,$$

(almeno esponenzialmente).

Da (3.18) segue che

$$\lim_{t \rightarrow \infty} S(t) = A\bar{S},$$

con  $A = e^{-H(\infty)} > 0$  e finito (poiché  $I(t)$  è integrabile, decadendo almeno esponenzialmente).

Ora, integrando  $\frac{dR}{dt} = \gamma I$  del sistema (3.11) otteniamo

$$R(t) = \bar{R} + H_1(t), \quad (3.38)$$

dove

$$H_1(t) := \gamma \int_0^t I(s) ds = \frac{\gamma}{\beta} H(t).$$

Infine, da (3.38) segue che

$$\lim_{t \rightarrow \infty} R(t) = \bar{R} + B,$$

con  $B = \frac{\gamma}{\beta} H(\infty) > 0$  e finito.

Dunque preso un dato iniziale  $\bar{x} = (\bar{S}, \bar{I}, \bar{R})$ , con  $\bar{S} \geq 0$ ,  $\bar{I} \geq 0$  e  $\bar{R} \geq 0$ , il punto limite della traiettoria sarà del tipo  $(S_{\bar{x}}, 0, R_{\bar{x}})$  dove  $S_{\bar{x}}$  e  $R_{\bar{x}}$  dipendono da  $\bar{x}$  e  $S_{\bar{x}} < \frac{\gamma}{\beta}$ .

Osserviamo che gli  $S_{\bar{x}}$  tendono a 0 al variare di  $\bar{x}$ . Infatti supponiamo che non sia così. Allora esiste un  $b > 0$  tale che  $S_{\bar{x}} \geq b$  per ogni  $\bar{x}$ . Ma, preso il dato iniziale  $\bar{x}_1 = (b, \bar{I}_1, \bar{R}_1)$  sappiamo che  $S$  decresce e quindi per qualche  $t$  avremmo che  $S(t) < b$ , e quindi anche  $S_{\bar{x}_1} < b$ , contraddizione.

Restringendoci al piano  $(S, I)$ , possiamo osservare le traiettorie, ovvero le curve  $(S(t), I(t))$  in figura.

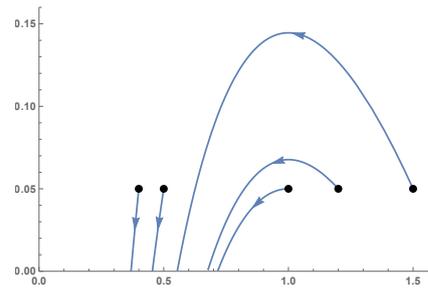


Figura 3.1: Traiettorie nel piano  $(S, I)$  con  $\gamma = \beta = 1$  e  $\bar{x}$  (pallino nero) dato iniziale

Notiamo che due traiettorie non si intersecano mai perchè altrimenti preso un punto di di intersezione come dato iniziale ci sarebbero due soluzioni, violando l'unicità.

Diremo allora che  $R_0 = \frac{\beta \bar{S}}{\gamma}$  assume un *valore di soglia*. Per evitare quindi che la malattia si diffonda ulteriormente bisogna ridurre il numero di suscettibili  $\bar{S}$ . La vaccinazione, ad esempio, è uno dei metodi più efficaci per evitare che la malattia si diffonda ulteriormente.

### Modello SIRD

Lo studio relativo al modello SIRD ha come riferimento lo studio del modello in [CAC] al quale è stato aggiunto un contributo personale.

Gli studi dell'andamento della diffusione del COVID-19 hanno fatto riferimento a questo modello epidemiologico.

Può succedere che l'infezione che si sta diffondendo potrebbe causare la morte di alcuni individui della popolazione. In tal caso è conveniente utilizzare un modello matematico che tenga in considerazione tale fatto. Indichiamo con  $D(t)$  il numero di individui morti a causa dell'infezione al tempo  $t$  con un *tasso di morte*  $\phi > 0$  e  $R(t)$  il numero di guariti. Possiamo allora modificare il sistema (3.11), che definiva il modello SIR epidemico senza tener conto della morte causata dalla pandemia, nel sistema seguente:

$$\begin{cases} \frac{dS}{dt} = -\beta SI \\ \frac{dI}{dt} = \beta SI - \phi I - \gamma I \\ \frac{dR}{dt} = \gamma I \\ \frac{dD}{dt} = \phi I, \end{cases} \quad (3.39)$$

dove i parametri  $\beta > 0$  e  $\gamma > 0$  rappresentano rispettivamente il *tasso di infezione* della malattia e il *tasso di recupero* dalla categoria degli infetti, con condizioni iniziali  $S(0) = S_0 > 0$ ,  $I(0) = I_0 > 0$ ,  $R(0) = 0$  e  $D(0) = 0$ .

Per il modello SIRD vale la seguente legge

$$N = S(t) + I(t) + R(t) + D(t) \quad \forall t \quad (3.40)$$

e, di conseguenza,  $D(t) = N - S(t) - I(t) - R(t)$ .

**Teorema 3.17** *Sia  $J_{SIRD} = \langle -\beta SI, \beta SI - \phi I - \gamma I, \gamma I \rangle$  l'ideale associato al modello SIRD (3.39) nell'anello  $\mathbb{R}[S, I]$  con l'ordine monomiale *lex*  $S > I$ . Allora:*

*i. la base di Groebner ridotta di  $J_{SIRD}$  è*

$$G_{rid} = \{I\};$$

*ii. l'insieme dei punti di equilibrio è*

$$V_{SIRD} = \{(S_e, 0, R_e, D_e), \forall S_e, R_e, D_e \in \mathbb{R}\}.$$

*Dimostrazione.* Per la Def. 3.8 i punti di equilibrio per il sistema (3.39) sono tutti i punti che soddisfano il sistema

$$\begin{cases} -\beta SI = 0 \\ \beta SI - \phi I - \gamma I = 0 \\ \gamma I = 0 \\ \phi I = 0. \end{cases} \quad (3.41)$$

Osserviamo che le ultime due equazioni del sistema (3.41) sono equivalenti, quindi per trovare i punti di equilibrio di (3.39) è sufficiente risolvere il sistema

$$\begin{cases} -\beta SI = 0 \\ \beta SI - \phi I - \gamma I = 0 \\ \gamma I = 0 \end{cases} \quad (3.42)$$

o, equivalentemente, trovare i punti che appartengono a  $V(G_{rid})$ , dove  $G_{rid}$  è la base di Groebner ridotta per l'ideale  $\langle -\beta SI, \beta SI - \phi I - \gamma I, \gamma I \rangle \in \mathbb{R}[S, I]$ .

Una base di Groebner per l'ideale  $\langle -\beta SI, \beta SI - \phi I - \gamma I, \gamma I \rangle$  è

$$G = \{-\beta SI, \beta SI - \phi I - \gamma I, \gamma I\}.$$

Infatti,  $\mathcal{S}(-\beta SI, \beta SI - \phi I - \gamma I) = \frac{\phi + \gamma}{\beta} I = \frac{\phi + \gamma}{\beta \gamma} (\gamma I)$ ,  $\mathcal{S}(-\beta SI, \gamma I) = 0$  e infine  $\mathcal{S}(\beta SI - \phi I - \gamma I, \gamma I) = -\frac{\phi + \gamma}{\beta} I = -\frac{\phi + \gamma}{\beta \gamma} (\gamma I)$ .

Dunque  $\overline{\mathcal{S}(-\beta SI, \beta SI - \phi I - \gamma I)}^G = 0 = \overline{\mathcal{S}(\beta SI - \phi I - \gamma I, \gamma I)}^G$ .

In conclusione, per il criterio di Buchberger (vedasi Teor. 2.41)  $G$  è una base di Groebner per  $\langle -\beta SI, \beta SI - \phi I - \gamma I, \gamma I \rangle$ .

In particolare, osserviamo che

$$\text{LT}(-\beta SI) = \frac{-\beta S}{\gamma} \text{LT}(\gamma I)$$

e

$$\text{LT}(\beta SI - \phi I - \gamma I) = \frac{\beta S}{\gamma} \text{LT}(\gamma I),$$

dunque per il Lemma 2.45 possiamo eliminare dalla base  $G$  i polinomi  $-\beta SI$  e  $\beta SI - \phi I - \gamma I$ . Inoltre, rendendo monico  $\gamma I$ , per la Def. 2.50 si ha che  $G_{rid} = \{I\}$  è la base di Groebner ridotta per  $\langle -\beta SI, \beta SI - \phi I - \gamma I, \gamma I \rangle$ .

E' facile vedere che i punti in  $V(G_{rid})$ , ovvero i punti di equilibrio per il sistema (3.39) appartengono al piano  $I = 0$  e dunque sono infiniti punti del tipo  $(S_e, 0, R_e, D_e)$ .

□

Ci occupiamo ora della stabilità dei punti di equilibrio trovati nel teorema precedente.

**Teorema 3.18** *Sia  $V_{SIRD} = \{(S_e, 0, R_e, D_e), \forall S_e, R_e, D_e \in \mathbb{R}\}$  la varietà associata al modello SIRD (vedasi Teor. 3.17). Allora il punto di equilibrio  $(S_e, 0, R_e, D_e)$ , con  $S_e \geq 0, R_e \geq 0, D_e \geq 0$ , è:*

- i. instabile se  $S_e > \frac{\gamma+\phi}{\beta}$ ;*
- ii. stabile (ma non asintoticamente stabile) se  $S_e < \frac{\gamma+\phi}{\beta}$ .*

*Dimostrazione.* Per poter applicare i teoremi di stabilità e quindi determinare se i punti di equilibrio sono stabili o instabili dobbiamo trovare la matrice Jacobiana relativa al sistema (3.39), calcolarla nei punti di equilibrio e determinarne gli autovalori.

La matrice Jacobiana associata al sistema (3.39) è:

$$J(S, I, R, D) = \begin{pmatrix} -\beta I & -\beta S & 0 & 0 \\ \beta I & \beta S - \phi - \gamma & 0 & 0 \\ 0 & \gamma & 0 & 0 \\ 0 & \phi & 0 & 0 \end{pmatrix}. \quad (3.43)$$

La matrice Jacobiana calcolata nei punti di equilibrio  $(S_e, 0, R_e, D_e)$  è

$$J(S_e, 0, R_e, D_e) = \begin{pmatrix} 0 & -\beta S_e & 0 & 0 \\ 0 & \beta S_e - \phi - \gamma & 0 & 0 \\ 0 & \gamma & 0 & 0 \\ 0 & \phi & 0 & 0 \end{pmatrix}, \quad (3.44)$$

e i suoi rispettivi autovalori sono gli zeri del polinomio

$$p(\lambda) = \begin{vmatrix} -\lambda & -\beta S_e & 0 & 0 \\ 0 & \beta S_e - \phi - \gamma - \lambda & 0 & 0 \\ 0 & \gamma & -\lambda & 0 \\ 0 & \phi & 0 & -\lambda \end{vmatrix} = -\lambda^3(\beta S_e - \phi - \gamma - \lambda).$$

Se  $S_e \neq \frac{\phi+\gamma}{\beta}$ , gli autovalori sono  $\lambda_1 = 0$  e  $\lambda_2 = \beta S_e - \phi - \gamma$ . Altrimenti, la matrice (3.44) ammette come unico autovalore  $\lambda_1 = 0$ .

Se  $S_e > \frac{\phi+\gamma}{\beta}$  la matrice Jacobiana ammette un autovalore positivo e per il Teor. 3.12, i punti di equilibrio sono instabili.

Se  $S_e < \frac{\phi + \gamma}{\beta}$  allora  $\lambda_2 < 0$  ma questo non è sufficiente per applicare i teoremi di stabilità dato che la matrice Jacobiana (3.44) ammette un autovalore nullo. Dobbiamo allora eseguire altri calcoli.

Sia  $x_e = (S_e, 0, R_e, D_e)$  con  $0 \leq S_e < \frac{\gamma + \phi}{\beta}$ ,  $R_e \geq 0$  e  $D_e \geq 0$ .  
 $\forall \epsilon > 0$ , sia  $\delta > 0$  tale che:

- i.  $\gamma + \phi - \beta S_e - \beta \delta > 0$ ;
- ii.  $\delta < \frac{\epsilon}{8}$ ;
- iii.  $S_e - S_e e^{\frac{-\beta \delta}{\gamma + \phi - \beta S_e - \beta \delta}} + \delta e^{\frac{-\beta \delta}{\gamma + \phi - \beta S_e - \beta \delta}} < \frac{\epsilon}{8}$ ;
- iv.  $\delta + \frac{\gamma \delta}{\gamma + \phi - \beta S_e - \beta \delta} < \frac{\epsilon}{8}$ .

Osserviamo che per le condizioni iii. e iv. sono possibili perchè

$$\lim_{\delta \rightarrow 0^+} \left( S_e - S_e e^{\frac{-\beta \delta}{\gamma + \phi - \beta S_e - \beta \delta}} + \delta e^{\frac{-\beta \delta}{\gamma + \phi - \beta S_e - \beta \delta}} \right) = 0$$

e

$$\lim_{\delta \rightarrow 0} \left( \delta + \frac{\gamma \delta}{\gamma + \phi - \beta S_e - \beta \delta} \right) = 0$$

Sia  $\bar{x} = (\bar{S}, \bar{I}, \bar{R}, \bar{D}) \in D_\delta^4(x_e) \cap \{S \geq 0, I \geq 0, R \geq 0, D \geq 0\}$  il dato iniziale.

Dal sistema (3.39) segue che  $\frac{dS}{dt} < 0$  e dunque  $S$  decresce sempre. Di conseguenza, si ha che  $S(t) \leq \bar{S}$  per ogni  $t \geq 0$ .

Dall'equazione  $\frac{dI}{dt} = (\beta S - \gamma - \phi)I$  del sistema (3.39) segue che

$$\frac{dI}{dt} \leq 0 \iff S \leq \frac{\gamma + \phi}{\beta}.$$

Ora, per la condizione i. sappiamo che  $\delta < \frac{\gamma + \phi}{\beta} - S_e$  quindi

$$S(t) \leq \bar{S} < S_e + \delta < \frac{\gamma + \phi}{\beta}.$$

Per quanto detto,  $I(t)$  decresce quindi dalla condizione ii. segue che

$$0 \leq I(t) \leq \bar{I} < \delta < \frac{\epsilon}{8}. \quad (3.45)$$

Ricordiamo che  $S(t)$  è decrescente quindi  $S(t) \leq \bar{S}$ .

Sia  $a := \gamma + \phi - \beta S_e - \beta \delta > 0$  per la condizione i. Allora

$$\beta S - (\gamma + \phi) \leq \beta \bar{S} - (\gamma + \phi) < \beta S_e + \beta \delta - (\gamma + \phi) = -a.$$

e, di conseguenza,

$$\frac{dI}{dt} = (\beta S - \gamma - \phi)I \leq -aI. \quad (3.46)$$

Integrando la (3.46) otteniamo

$$I(t) \leq \bar{I}e^{-at}. \quad (3.47)$$

L'equazione  $\frac{dS}{dt} = -\beta SI$  del sistema (3.39) si può riscrivere nel modo seguente:  $\frac{dS}{dt} \frac{1}{S} = -\beta I$ , che integrata dà

$$\log(S(t)) = \log(\bar{S}) - H(t)$$

o, equivalentemente,

$$S(t) = \bar{S}e^{-H(t)} \quad (3.48)$$

dove

$$H(t) := \beta \int_0^t I(s) ds,$$

( $I(t)$  è integrabile, decadendo almeno esponenzialmente). Poiché  $I$  e  $S$  decrescono e sono sempre positivi, la traiettoria rimane sempre in un compatto, quindi (per il teorema del prolungamento <sup>2</sup>)  $S(t)$  è definita per ogni  $t$ .

Ora, utilizzando la (3.47) otteniamo

$$H(t) := \beta \int_0^t I(s) ds \leq \beta \int_0^t \bar{I}e^{-as} ds = \frac{\beta \bar{I}}{a}(1 - e^{-at}) < \frac{\beta \bar{I}}{a} \leq \frac{\beta \delta}{a}. \quad (3.49)$$

Di conseguenza, utilizzando la (3.48), segue che

$$S(t) = \bar{S}e^{-H(t)} \geq \bar{S}e^{-\frac{\beta \delta}{a}} \geq (S_e - \delta)e^{-\frac{\beta \delta}{a}}. \quad (3.50)$$

Ora,

$$|S(t) - S_e| < \frac{\epsilon}{8}. \quad (3.51)$$

Infatti, se  $S(t) \geq S_e$ , allora per la condizione ii.

$$|S(t) - S_e| = S(t) - S_e \leq \bar{S} - S_e < \delta < \frac{\epsilon}{8}.$$

Se  $S(t) \leq S_e$ , allora per la (3.50) e per la condizione iii.

$$|S(t) - S_e| = S_e - S(t) \leq S_e - (S_e - \delta)e^{-\frac{\beta \delta}{a}} = S_e - S_e e^{\frac{-\beta \delta}{\gamma + \phi - \beta S_e - \beta \delta}} + \delta e^{\frac{-\beta \delta}{\gamma + \phi - \beta S_e - \beta \delta}} < \frac{\epsilon}{8}.$$

---

<sup>2</sup>Vedasi [G, Cap.3, Cor. 12.23]

Ora,  $\frac{dR}{dt} = \gamma I > 0$  quindi  $R(t)$  è crescente e dunque  $R(t) \geq \bar{R}$ . Inoltre integrando l'equazione  $\frac{dR}{dt} = \gamma I$  otteniamo

$$R(t) = \bar{R} + H_1(t), \quad (3.52)$$

dove

$$H_1(t) := \gamma \int_0^t I(s) ds = \frac{\gamma}{\beta} H(t).$$

Allora per la (3.49)

$$H_1(t) \leq \frac{\gamma}{\beta} \frac{\beta \delta}{a} = \frac{\gamma \delta}{a}. \quad (3.53)$$

Osserviamo che

$$|R(t) - R_e| < \frac{\epsilon}{8}. \quad (3.54)$$

Infatti, se  $R(t) \geq R_e$  allora dalla (3.52) e dalla (3.53) segue che

$$|R(t) - R_e| = R(t) - R_e = \bar{R} + H_1(t) - R_e < \delta + \frac{\gamma \delta}{a}.$$

Allora per la condizione iv.

$$|R(t) - R_e| < \delta + \frac{\gamma \delta}{\gamma + \phi - \beta S_e - \beta \delta} < \frac{\epsilon}{8}.$$

Se,  $R(t) \leq R_e$  allora

$$|R(t) - R_e| = R_e - R(t) \leq R_e - \bar{R} < \delta < \frac{\epsilon}{8}.$$

Osserviamo che

$$|D(t) - D_e| < \frac{7}{8}\epsilon.$$

Infatti,

$$|D(t) - D_e| \leq |D(t) - \bar{D}| + |\bar{D} - D_e|. \quad (3.55)$$

Consideriamo  $|D(t) - \bar{D}|$ .

Dall'equazione (3.40) segue che  $D(t) = N - S(t) - I(t) - R(t)$ . Quindi

$$\begin{aligned} D(t) - \bar{D} &= N - S(t) - I(t) - R(t) - (N - \bar{S} - \bar{I} - \bar{R}) = \bar{S} - S(t) + \bar{I} - I(t) + \bar{R} - R(t) = \\ &= \bar{S} - S_e + S_e - S(t) + \bar{I} - I(t) + \bar{R} - R_e + R_e - R(t). \end{aligned}$$

Di conseguenza, per le (3.45), (3.51) e (3.54) e per la condizione ii.

$$|D(t) - \bar{D}| \leq |\bar{S} - S_e| + |S_e - S(t)| + |I| + |I(t)| + |\bar{R} - R_e| + |R_e - R(t)| \leq \frac{3}{4}\epsilon.$$

Inoltre,

$$|\bar{D}(t) - D_e| < \delta < \frac{\epsilon}{8}.$$

Quindi la (3.55) diventa

$$|D(t) - D_e| \leq |D(t) - \bar{D}| + |\bar{D} - D_e| < \frac{7\epsilon}{8}. \quad (3.56)$$

Ma allora

$$\begin{aligned} \|x(t) - x_e\| &= \sqrt{(S(t) - S_e)^2 + I(t)^2 + (R(t) - R_e)^2 + (D(t) - D_e)^2} \\ &\leq \sqrt{\left(\frac{\epsilon}{8}\right)^2 + \left(\frac{\epsilon}{8}\right)^2 + \left(\frac{\epsilon}{8}\right)^2 + \left(\frac{7\epsilon}{8}\right)^2} = \frac{\sqrt{13}}{4}\epsilon < \epsilon. \end{aligned}$$

In conclusione, i punti di equilibrio  $x_e = (S_e, 0, R_e, D_e)$ , con  $S_e \geq 0, R_e \geq 0$  e  $D_e \geq 0$ , sono stabili se  $S_e < \frac{\gamma+\phi}{\beta}$ .

Il punto di equilibrio  $x_e$  non può essere asintoticamente stabile dato che se prendiamo un  $\bar{x} \in D_\delta^4(x_e) \cap \{S \geq 0, I \geq 0, R \geq 0\}$  tale che  $\bar{S} < S_e$  allora  $S(t) < S_e$  per ogni  $t$  quindi non è possibile che

$$\lim_{t \rightarrow \infty} x(t) = x_e.$$

□

**Osservazione 3.19** Sia  $S_e = \frac{\gamma+\phi}{\beta}$ . In questo caso sarebbe necessario uno studio a parte simile a quanto fatto nel modello SIR per determinare se il punto di equilibrio  $(S_e, 0, R_e, D_e)$ , con  $R_e \geq 0$  e  $D_e \geq 0$ , è stabile o instabile.

In realtà, non è particolarmente rilevante dal punto di vista pratico studiare il caso  $S_e = \frac{\gamma+\phi}{\beta}$  perché difficilmente i parametri fisici assumono valori per cui  $S_e$  sia uguale esattamente a quel valore. In realtà ciò che è importante è che ci sia una transizione tra due regimi completamente diversi.

**Osservazione 3.20** Sia  $\bar{x} = (\bar{S}, \bar{I}, \bar{R}, \bar{D})$  il dato iniziale con  $\bar{S} \geq 0, \bar{I} \geq 0, \bar{R} \geq 0$  e  $\bar{D} \geq 0$ . Definiamo

$$R_0 = \frac{\beta\bar{S}}{\gamma + \phi}.$$

Se  $R_0 \leq 1$  allora  $S \leq \bar{S} \leq \frac{\gamma+\phi}{\beta}$  dunque  $I$  è decrescente.

Se  $R_0 > 1$  abbiamo che  $\bar{S} > \frac{\gamma+\phi}{\beta}$ , quindi  $I$  è crescente finché  $S > \frac{\gamma+\phi}{\beta}$ . Ma  $S$  è decrescente quindi si arriverà ad un tempo  $T$  tale che  $S(T) = \frac{\gamma+\phi}{\beta}$ ,  $\frac{dS}{dt} < 0$  e  $\frac{dI}{dt} = 0$ . Quindi, quale che sia il dato iniziale  $\bar{x}$  la traiettoria entra nella regione

$$S < \frac{\gamma + \phi}{\beta}.$$

Quindi o  $\bar{S} < \frac{\gamma + \phi}{\beta}$  o comunque, se  $\bar{S} > \frac{\gamma + \phi}{\beta}$ , esiste un tempo  $T_1$  tale che  $S(T_1) < \frac{\gamma + \phi}{\beta}$ . Dunque,  $I(t)$  diminuisce in ogni caso e ci si può limitare ai dati iniziali tali che  $\bar{S} < \frac{\gamma + \phi}{\beta}$ . A livello pratico però, è differente se  $\bar{S} < \frac{\gamma + \phi}{\beta}$  o  $\bar{S} > \frac{\gamma + \phi}{\beta}$ . Infatti nel primo caso,  $I(t)$  diminuisce, mentre nel caso  $\bar{S} > \frac{\gamma + \phi}{\beta}$ ,  $I(t)$  prima aumenta e poi diminuisce. In particolare, se  $\bar{S} \gg \frac{\gamma + \phi}{\beta}$  oppure  $\bar{I} \gg 0$ , allora  $I(t)$  aumenta moltissimo prima di diminuire. Dunque il numero dei malati diminuisce comunque, però in un caso diminuisce perché si ammalano sempre meno persone e nell'altro perché si ammalano tutti e a un certo punto o guariscono o muoiono, così che alla fine in ogni caso non ci sono più malati.

Ora, vogliamo determinare il punto limite.

Dall'equazione (3.47) e ricordando che  $I(t)$  è positivo, se ne deduce che

$$\lim_{t \rightarrow \infty} I(t) = 0,$$

(almeno esponenzialmente).

Da (3.48) segue che

$$\lim_{t \rightarrow \infty} S(t) = A\bar{S},$$

con  $A = e^{-H(\infty)} > 0$  (poiché  $I(t)$  è integrabile, decadendo almeno esponenzialmente).

Da (3.52) segue che

$$\lim_{t \rightarrow \infty} R(t) = \bar{R} + B,$$

con  $B = \frac{\gamma}{\beta}H(\infty) > 0$  e finito.

Infine, integrando l'equazione  $\frac{dD}{dt} = \phi I$  del sistema (3.39) otteniamo

$$D(t) = \bar{D} + H_2(t), \quad (3.57)$$

dove

$$H_2(t) := \phi \int_0^t I(s) ds = \frac{\phi}{\beta} H(t).$$

Di conseguenza,

$$\lim_{t \rightarrow \infty} D(t) = \bar{D} + C,$$

dove  $C = \frac{\phi}{\beta}H(t)(\infty) > 0$  e finito.

Dunque preso un dato iniziale  $\bar{x} = (\bar{S}, \bar{I}, \bar{R}, \bar{D})$ , con  $\bar{S} \geq 0$ ,  $\bar{I} \geq 0$ ,  $\bar{R} \geq 0$  e  $\bar{D} \geq 0$  il punto limite della traiettoria sarà del tipo  $(S_{\bar{x}}, 0, R_{\bar{x}}, D_{\bar{x}})$  dove  $S_{\bar{x}}$ ,  $R_{\bar{x}}$  e  $D_{\bar{x}}$  dipendono da  $\bar{x}$  e  $S_{\bar{x}} < \frac{\gamma + \phi}{\beta}$ .

Osserviamo che gli  $S_{\bar{x}}$  tendono a 0 al variare di  $\bar{x}$ . Infatti supponiamo che non sia così. Allora esiste un  $b > 0$  tale che  $S_{\bar{x}} \geq b$  per ogni  $\bar{x}$ . Ma, preso il dato iniziale  $\bar{x}_1 = (b, \bar{I}_1, \bar{R}_1, \bar{D}_1)$  sappiamo che  $S$  decresce e quindi per qualche  $t$  avremmo che  $S(t) < b$ , e quindi anche  $S_{\bar{x}_1} < b$ , contraddizione.

Notiamo che due traiettorie non si intersecano mai perchè altrimenti preso un punto di intersezione come dato iniziale ci sarebbero due soluzioni, violando l'unicità.

Restringendoci al piano  $(S, I)$ , possiamo osservare le traiettorie, ovvero le curve  $(S(t), I(t))$  in figura.

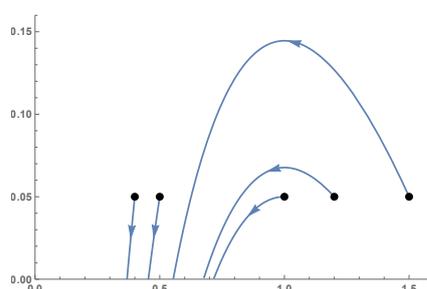


Figura 3.2: Traiettorie nel piano  $(S, I)$  con  $\gamma + \phi = \beta = 1$  e  $\bar{x}$  (pallino nero) dato iniziale

Diremo allora che  $R_0 = \frac{\beta \bar{S}}{\gamma + \phi}$  assume un *valore di soglia*. Per evitare quindi che la malattia si diffonda ulteriormente bisogna ridurre il numero di suscettibili  $\bar{S}$ . La vaccinazione, ad esempio, è uno dei metodi più efficaci per evitare che la malattia si diffonda ulteriormente.

### Modelli SIR endemici

Si parla di modello endemico quando la durata della malattia è tale da influenzare l'andamento della vita della popolazione colpita. Supponiamo che il *tasso di natalità e mortalità* sia lo stesso valore e lo indichiamo con  $\mu > 0$ .

Studieremo due modelli SIR endemici: nel primo, il modello SIR con vaccinazione, una parte dei suscettibili passa direttamente nella categoria dei rimossi con un *tasso di vaccinazione*  $\rho \geq 0$ , dunque il numero di persone suscettibili è  $(1 - \rho)N$ ; nel secondo invece, tutti i suscettibili sono predisposti ad essere infettati, dunque il tasso di vaccinazione è  $\rho = 0$ .

Studieremo quindi il caso più generale in cui  $\rho \geq 0$  e applicheremo i risultati ottenuti al caso  $\rho = 0$ .

### Modello SIR e caso della vaccinazione

Lo studio del modello SIR e caso della vaccinazione è stato fatto a partire dai dati riportati in [MBK, p.671] ai quali è stato aggiunto un contributo personale.

E' possibile studiare un particolare modello SIR endemico in cui una parte dei suscettibili passa direttamente nella categoria dei rimossi con un *tasso di vaccinazione*  $\rho \geq 0$ . Questo significa che il numero di suscettibili non è più  $N$  ma gli individui a rischio infezione sono  $(1 - \rho)N$ . Questo implica che  $\rho \leq 1$ .

Tale modello può essere espresso dal sistema:

$$\begin{cases} \frac{dS}{dt} = -\beta SI - \mu S + (1 - \rho)\mu N \\ \frac{dI}{dt} = \beta SI - \gamma I - \mu I \\ \frac{dR}{dt} = \gamma I - \mu R + \rho\mu N, \end{cases} \quad (3.58)$$

dove i parametri  $\beta > 0$ ,  $\gamma > 0$  e  $\mu > 0$ , rappresentano rispettivamente il *tasso di infezione* della malattia, il *tasso di recupero* dalla categoria degli infetti e il *tasso di natalità e mortalità*, con condizioni iniziali  $S(0) = S_0 > 0$ ,  $I(0) = I_0 > 0$  e  $R(0) = 0$ .

Dalla (3.10) segue che il numero di rimossi è pari a  $R(t) = N - S(t) - I(t)$ .

**Teorema 3.21** Sia  $J_{SIR_{vac}} = \langle -\beta SI - \mu S + (1 - \rho)\mu N, \beta SI - \gamma I - \mu I \rangle$  l'ideale associato al modello SIR con vaccinazione (3.58) nell'anello  $\mathbb{R}[S, I]$ , con l'ordine monomiale *lex*  $S > I$ . Sia

$$R_0 = \frac{(1 - \rho)\beta N}{\gamma + \mu}. \quad (3.59)$$

Allora:

i. la base di Groebner ridotta di  $J_{SIR_{vac}}$  è

$$G_{rid} = \left\{ S + \frac{(1-\rho)\beta N}{\mu R_0} I - (1-\rho)N, I^2 - \frac{\mu}{\beta}(R_0 - 1)I \right\};$$

ii. l'insieme dei punti di equilibrio è  $V_{SIR_{vac}} = \{P_1, P_2\}$  dove

$$P_1 = ((1-\rho)N, 0, \rho N)$$

e

$$P_2 = \left( \frac{(1-\rho)N}{R_0}, \frac{\mu}{\beta}(R_0 - 1), \frac{\gamma}{\beta}(R_0 - 1) + \rho N \right).$$

*Dimostrazione.* Per la Def. 3.8 i punti di equilibrio per (3.58) sono tutti i punti che soddisfano il sistema

$$\begin{cases} -\beta SI - \mu S + (1-\rho)\mu N = 0 \\ \beta SI - \gamma I - \mu I = 0 \\ \gamma I - \mu R + \rho\mu N = 0. \end{cases} \quad (3.60)$$

Osserviamo che dalla terza equazione del sistema (3.60) ricaviamo l'equazione

$$R_e = \frac{\gamma I_e + \rho\mu N}{\mu}. \quad (3.61)$$

Di conseguenza, per trovare i punti di equilibrio per (3.58) è sufficiente trovare tutti i punti che soddisfino il sistema

$$\begin{cases} -\beta SI - \mu S + (1-\rho)\mu N = 0 \\ \beta SI - \gamma I - \mu I = 0 \end{cases} \quad (3.62)$$

e sfruttare la (3.61).

In particolare, cerchiamo tutti i punti che appartengono a  $V(G_{rid})$ , dove  $G_{rid}$  è la base di Groebner ridotta per l'ideale  $\langle -\beta SI - \mu S + (1-\rho)\mu N, \beta SI - \gamma I - \mu I \rangle$  in  $\mathbb{R}[S, I]$ , tenendo presente la (3.61).

Una base di Groebner per l'ideale di riferimento è

$$G = \left\{ -\beta SI - \mu S + (1-\rho)\mu N, \beta SI - \gamma I - \mu I, \right. \\ \left. \frac{1}{\beta}(\mu S + \gamma I + \mu I - (1-\rho)\mu N), -\frac{\gamma + \mu}{\mu} I^2 + \frac{1}{\beta}((1-\rho)N\beta - \gamma - \mu)I \right\}.$$

Per dimostrare tale fatto basta applicare l'algoritmo di Buchberger (vedasi Teor. 2.43). Consideriamo la base  $G' = \{-\beta SI - \mu S + (1-\rho)\mu N, \beta SI - \gamma I - \mu I\}$ .

Osserviamo che

$$\mathcal{S}(-\beta SI - \mu S + (1-\rho)\mu N, \beta SI - \gamma I - \mu I) = \frac{1}{\beta}(\mu S + \gamma I + \mu I - (1-\rho)\mu N) =$$

$$= \overline{\mathcal{S}(-\beta SI - \mu S + (1 - \rho)\mu N, \beta SI - \gamma I - \mu I)}^{G'}.$$

Per il Teor. 2.41,  $G'$  non è una base di Groebner per l'ideale di riferimento. Sia  $G'' = \{-\beta SI - \mu S + (1 - \rho)\mu N, \beta SI - \gamma I - \mu I, \frac{1}{\beta}(\mu S + \gamma I + \mu I - (1 - \rho)\mu N)\}$ . Osserviamo che

$$\overline{\mathcal{S}(-\beta SI - \mu S + (1 - \rho)\mu N, \beta SI - \gamma I - \mu I)}^{G''} = 0;$$

$$\begin{aligned} \mathcal{S}(-\beta SI - \mu S + (1 - \rho)\mu N, \frac{1}{\beta}(\mu S + \gamma I + \mu I - (1 - \rho)\mu N)) &= \\ &= \frac{\mu}{\beta}S - \frac{\gamma + \mu}{\mu}I^2 + (1 - \rho)NI - \frac{(1 - \rho)\mu}{\beta}N, \\ \mathcal{S}(\beta SI - \gamma I - \mu I, \frac{1}{\beta}(\mu S + \gamma I + \mu I - (1 - \rho)\mu N)) &= \\ &= -\frac{\gamma + \mu}{\mu}I^2 + \frac{(1 - \rho)N\beta - \gamma - \mu}{\beta}I, \end{aligned}$$

e

$$\begin{aligned} \overline{\mathcal{S}(-\beta SI - \mu S + (1 - \rho)\mu N, \frac{1}{\beta}(\mu S + \gamma I + \mu I - (1 - \rho)\mu N))}^{G''} &= \\ &= -\frac{\gamma + \mu}{\mu}I^2 + \frac{1}{\beta}((1 - \rho)N\beta - \gamma - \mu)I = \\ &= \overline{\mathcal{S}(\beta SI - \gamma I - \mu I, \frac{1}{\beta}(\mu S + \gamma I + \mu I - (1 - \rho)\mu N))}^{G''}. \end{aligned}$$

Quindi per il Teor. 2.41,  $G''$  non è una base di Groebner per l'ideale di riferimento. Analizziamo allora la base  $G = \{-\beta SI - \mu S + (1 - \rho)\mu N, \beta SI - \gamma I - \mu I, \frac{1}{\beta}(\mu S + \gamma I + \mu I - (1 - \rho)\mu N), -\frac{\gamma + \mu}{\mu}I^2 + \frac{1}{\beta}((1 - \rho)N\beta - \gamma - \mu)I\}$ .

Calcolando  $\mathcal{S}(g_i, g_j)$  per ogni  $g_i, g_j \in G$  osserviamo che  $\overline{\mathcal{S}(g_i, g_j)}^G = 0$  per ogni  $g_i, g_j \in G$ . Di conseguenza per il criterio di Buchberger (vedasi Teor. 2.41),  $G$  è una base di Groebner per  $\langle -\beta SI - \mu S + (1 - \rho)\mu N, \beta SI - \gamma I - \mu I \rangle$ .

Osserviamo che

$$\text{LT}(-\beta SI - \mu S + (1 - \rho)\mu N) = \frac{-\beta^2 I}{\mu} \text{LT}\left(\frac{1}{\beta}(\mu S + \gamma I + \mu I - (1 - \rho)\mu N)\right)$$

e

$$\text{LT}(\beta SI - \gamma I - \mu I) = \frac{\beta^2 I}{\mu} \text{LT}\left(\frac{1}{\beta}(\mu S + \gamma I + \mu I - (1 - \rho)\mu N)\right),$$

dunque per il Lemma 2.45 possiamo eliminare dalla base  $G$  i polinomi  $-\beta SI - \mu S + (1 - \rho)\mu N$  e  $\beta SI - \gamma I - \mu I$ . Inoltre, rendendo monici i restanti polinomi

della base  $G$ , per la Def. 2.50 otteniamo la base ridotta per l'ideale  $\langle -\beta SI - \mu S + (1 - \rho)\mu N, \beta SI - \gamma I - \mu I \rangle$ :

$$G_{rid} = \left\{ S + \frac{\gamma + \mu}{\mu} I - (1 - \rho)N, I^2 - \frac{\mu}{\beta(\gamma + \mu)}((1 - \rho)N\beta - \gamma - \mu)I \right\}$$

o, equivalentemente per la (3.59),

$$G_{rid} = \left\{ S + \frac{(1 - \rho)\beta N}{\mu R_0} I - (1 - \rho)N, I^2 - \frac{\mu}{\beta}(R_0 - 1)I \right\}.$$

Dall'equazione (3.61) e calcolando  $V(G_{rid})$  otteniamo che l'insieme dei punti di equilibrio per il sistema (3.58) è  $V_{SIR_{vac}} = \{P_1, P_2\}$  dove

$$P_1 = ((1 - \rho)N, 0, \rho N)$$

e

$$P_2 = \left( \frac{(1 - \rho)N}{R_0}, \frac{\mu}{\beta}(R_0 - 1), \frac{\gamma}{\beta}(R_0 - 1) + \rho N \right).$$

□

Ci occupiamo ora della stabilità dei punti di equilibrio trovati nel teorema precedente.

**Osservazione 3.22** Come premesso all'inizio del capitolo, supporremo che le coordinate dei punti di equilibrio siano non negative. Infatti, con dati iniziali  $S(0) \geq 0$ ,  $I(0) \geq 0$ ,  $R(0) \geq 0$  si rimane nella regione  $S(t) \geq 0$ ,  $I(t) \geq 0$ ,  $R(t) \geq 0 \forall t \geq 0$ . In particolare per  $P_2$  questo implica che  $R_0 \geq 1$ . Del resto se  $R_0 = 1$ ,  $P_2$  coincide con  $P_1$ . Quindi considereremo  $P_2$  solo nel caso  $R_0 > 1$ .

**Teorema 3.23** *Sia  $V_{SIR_{vac}} = \{P_1, P_2\}$  la varietà associata al modello SIR con vaccinazione (vedasi Teor. 3.21). Allora nel caso di coordinate non negative (vedasi Oss. 3.22):*

- i. se  $R_0 < 1$ ,  $P_1$  è un punto di equilibrio asintoticamente stabile. Invece, se  $R_0 > 1$ ,  $P_1$  è un punto di equilibrio instabile;*
- ii.  $P_2$  è un punto di equilibrio asintoticamente stabile.*

*Dimostrazione.* Per poter applicare i teoremi di stabilità e quindi determinare se  $P_1$  e  $P_2$  sono stabili o instabili dobbiamo trovare la matrice Jacobiana relativa al sistema (3.58), calcolarla nei punti di equilibrio e determinarne gli autovalori.

La matrice Jacobiana del sistema (3.58) è:

$$J(S, I, R) = \begin{pmatrix} -\beta I - \mu & -\beta S & 0 \\ \beta I & \beta S - \gamma - \mu & 0 \\ 0 & \gamma & -\mu \end{pmatrix}.$$

Soffermiamoci ad analizzare il punto di equilibrio  $P_1 = ((1 - \rho)N, 0, \rho N)$ . La matrice Jacobiana calcolata in tale punto è

$$J((1 - \rho)N, 0, \rho N) = \begin{pmatrix} -\mu & -\beta(1 - \rho)N & 0 \\ 0 & \beta(1 - \rho)N - \gamma - \mu & 0 \\ 0 & \gamma & -\mu \end{pmatrix},$$

e i suoi rispettivi autovalori sono gli zeri del polinomio

$$\begin{aligned} p(\lambda) &= \begin{vmatrix} -\mu - \lambda & -\beta(1 - \rho)N & 0 \\ 0 & \beta(1 - \rho)N - \gamma - \mu - \lambda & 0 \\ 0 & \gamma & -\mu - \lambda \end{vmatrix} = \\ &= (-\mu - \lambda)^2(\beta(1 - \rho)N - \gamma - \mu - \lambda). \end{aligned}$$

Gli autovalori sono  $\lambda_1 = -\mu < 0$  e  $\lambda_2 = \beta(1 - \rho)N - \gamma - \mu$ .

Se  $\beta(1 - \rho)N - \gamma - \mu < 0$ , o equivalentemente per la (3.59),  $R_0 < 1$ , tutti gli autovalori sono negativi, allora per il Teor. 3.11,  $P_1 = ((1 - \rho)N, 0, \rho N)$  è un punto di equilibrio asintoticamente stabile. Invece, se  $\beta(1 - \rho)N - \gamma - \mu > 0$ , o equivalentemente,  $R_0 > 1$ , la matrice Jacobiana ammette qualche autovalore positivo e per il Teor. 3.12,  $P_1 = ((1 - \rho)N, 0, \rho N)$  è un punto di equilibrio instabile.

Consideriamo ora il punto di equilibrio

$$P_2 = \left( \frac{(1 - \rho)N}{R_0}, \frac{\mu}{\beta}(R_0 - 1), \frac{\gamma}{\beta}(R_0 - 1) + \rho N \right)$$

per il sistema (3.58).

Ricordiamo che per ipotesi vale l'Oss. 3.22, studieremo solo il caso  $R_0 > 1$ .

Ponendo  $S_e = \frac{(1 - \rho)N}{R_0}$ ,  $I_e = \frac{\mu}{\beta}(R_0 - 1)$  e  $R_e = \frac{\gamma}{\beta}(R_0 - 1) + \rho N$ , la matrice Jacobiana calcolata in  $P_2 = (S_e, I_e, R_e)$  è

$$J(S_e, I_e, R_e) = \begin{pmatrix} -\beta I_e - \mu & -\beta S_e & 0 \\ \beta I_e & 0 & 0 \\ 0 & \gamma & -\mu \end{pmatrix}. \quad (3.63)$$

Gli autovalori della matrice (3.63) sono gli zeri del polinomio

$$p(\lambda) = \begin{vmatrix} -\beta I_e - \mu - \lambda & -\beta S_e & 0 \\ \beta I_e & -\lambda & 0 \\ 0 & \gamma & -\mu - \lambda \end{vmatrix} = \quad (3.64)$$

$$= (-\mu - \lambda)(\lambda^2 + \lambda(\beta I_e + \mu) + \beta^2 S_e I_e).$$

Osserviamo che gli zeri del polinomio (3.64) sono tutti reali se e solo se

$$\mu R_0^2 \geq 4(R_0 - 1)(\gamma + \mu). \quad (3.65)$$

Gli autovalori sono  $\lambda_1 = -\mu < 0$  e, se vale la (3.65), allora

$$\lambda_2 = -\frac{\mu R_0 + \sqrt{\mu^2 R_0^2 - 4\mu(\gamma + \mu)(R_0 - 1)}}{2} < 0$$

poichè  $R_0 > 0$  e

$$\lambda_3 = \frac{-\mu R_0 + \sqrt{\mu^2 R_0^2 - 4\mu(\gamma + \mu)(R_0 - 1)}}{2} < 0$$

poichè  $R_0 > 1$  per ipotesi.

Dunque, se vale la (3.65), tutti gli autovalori sono negativi e, di conseguenza, per il Teor. 3.11,  $P_2$  è un punto di equilibrio asintoticamente stabile.

Se la (3.65) non è soddisfatta, gli autovalori  $\lambda_2$  e  $\lambda_3$  non sono reali. Consideriamo allora  $\text{Re}(\lambda_{2,3}) = -\frac{\mu R_0}{2}$ . Per il Teor. 3.11, essendo  $\text{Re}(\lambda_{2,3}) < 0$  poichè  $R_0 > 0$ , il punto di equilibrio  $P_2$  è un punto di equilibrio asintoticamente stabile.

In conclusione, nel caso di coordinate non negative (vedasi Oss. 3.22) il punto di equilibrio  $P_2$  è asintoticamente stabile.  $\square$

**Osservazione 3.24** Se  $R_0 < 1$  le traiettorie vanno verso  $P_1$  (i rimossi sono i vaccinati e i suscettibili sono i restanti), se  $R_0 > 1$  le traiettorie vanno verso  $P_2$  (i rimossi aumentano perché gli infetti o guariscono o muoiono, i suscettibili diventano pochi perché si sono in gran parte ammalati).

**Osservazione 3.25** Se  $R_0 = 1$ , la stabilità del punto di equilibrio  $P_1$  resta un problema aperto. A dire il vero, non è particolarmente rilevante dal punto di vista pratico studiare il caso  $R_0 = 1$ . Infatti, difficilmente i parametri fisici assumono valori per cui  $R_0$  sia uguale esattamente a quel valore. In realtà ciò che è importante è che ci sia una transizione tra due regimi completamente diversi.

### Modello SIR endemico

Lo studio relativo al modello SIR endemico ha come riferimento lo studio del modello in [MBK, p.669] al quale è stato aggiunto un contributo personale.

Supponiamo ora di applicare i risultati ottenuti nel modello SIR con vaccinazione al caso  $\rho = 0$ . Questo significa che il numero di suscettibili è  $N$ , ovvero tutte le persone possono diventare infette.

Sostituendo  $\rho = 0$  al sistema (3.58) otteniamo le equazioni che rappresentano il modello SIR endemico:

$$\begin{cases} \frac{dS}{dt} = -\beta SI - \mu S + \mu N \\ \frac{dI}{dt} = \beta SI - \gamma I - \mu I \\ \frac{dR}{dt} = \gamma I - \mu R, \end{cases} \quad (3.66)$$

con condizioni iniziali  $S(0) = S_0 > 0$ ,  $I(0) = I_0 > 0$  e  $R(0) = 0$ , dove  $\mu > 0$ ,  $\beta > 0$  e  $\gamma > 0$  siano rispettivamente il *tasso di natalità e mortalità*, il *tasso di infezione* della malattia e il *tasso di recupero* dalla categoria degli infetti.

**Teorema 3.26** Sia  $J_{SIR_{end}} = \langle -\beta SI - \mu S + \mu N, \beta SI - \gamma I - \mu I \rangle$  l'ideale associato al modello SIR endemico (3.66) nell'anello  $\mathbb{R}[S, I]$ , con l'ordine monomiale  $lex$   $S > I$ . Sia

$$R_0 = \frac{\beta N}{\gamma + \mu}. \quad (3.67)$$

Allora:

i. la base di Groebner ridotta di  $J_{SIR_{end}}$  è

$$G_{rid} = \left\{ S + \frac{\beta N}{\mu R_0} I - N, I^2 - \frac{\mu}{\beta} (R_0 - 1) I \right\};$$

ii. l'insieme dei punti di equilibrio è  $V_{SIR_{end}} = \{P_1, P_2\}$  dove

$$P_1 = (N, 0, 0)$$

e

$$P_2 = \left( \frac{N}{R_0}, \frac{\mu}{\beta} (R_0 - 1), \frac{\gamma}{\beta} (R_0 - 1) \right).$$

*Dimostrazione.* Basta applicare il Teor. 3.21 al caso  $\rho = 0$ .

□

Ci occupiamo ora della stabilità dei punti di equilibrio trovati nel teorema precedente.

**Teorema 3.27** *Sia  $V_{SIR_{end}} = \{P_1, P_2\}$  la varietà associata al modello SIR endemico (vedasi Teor. 3.26). Nel caso di coordinate non negative (vedasi Oss. 3.22):*

- i. se  $R_0 < 1$ ,  $P_1$  è un punto di equilibrio asintoticamente stabile. Invece, se  $R_0 > 1$ ,  $P_1$  è un punto di equilibrio instabile;*
- ii.  $P_2$  è un punto di equilibrio asintoticamente stabile.*

*Dimostrazione.* Basta applicare il Teor. 3.23 al caso  $\rho = 0$ . □

**Osservazione 3.28** Se  $R_0 \leq 1$  l'epidemia si estinguerà. Se  $R_0 > 1$  l'epidemia si propagherà fintanto che  $S$  è vicino a  $N$  e si estinguerà se  $S \ll N$ .

*Dimostrazione.* Dalla seconda equazione del sistema (3.66),  $\frac{dI}{dt} = \beta SI - \gamma I - \mu I$ , otteniamo che

$$\frac{dI}{dt} \leq 0 \iff S \leq \frac{\gamma + \mu}{\beta}.$$

Se  $R_0 \leq 1$  allora  $S \leq N \leq \frac{\gamma + \mu}{\beta}$  quindi  $I$  è decrescente e l'epidemia si estinguerà. Se  $R_0 > 1$ , o equivalentemente  $N > \frac{\gamma + \mu}{\beta}$ , se  $S$  è molto vicino a  $N$ , allora  $S > \frac{\gamma + \mu}{\beta}$  e dunque  $\frac{dI}{dt} > 0$ , l'epidemia si propagherà finchè  $S(t) > \frac{\gamma + \mu}{\beta}$  poichè la funzione  $I(t)$  è crescente. Invece, se  $R_0 > 1$  (ovvero  $N > \frac{\gamma + \mu}{\beta}$ ), se  $S \ll N$  allora  $S < \frac{\gamma + \mu}{\beta}$ , quindi  $I$  è decrescente e l'epidemia si estinguerà. □

Il modello SIR può essere espanso per ottenere simulazioni più realistiche suddividendo la popolazione in altre categorie come:

- **M**, gli individui immuni dalla nascita: neonati che ricevono dalla propria madre, se questa è stata infetta, anticorpi che li difendono dalla malattia.
- **E**, gli esposti: soggetti contagiati ma non ancor infettivi.

Per molte infezioni, i bambini non nascono nel compartimento suscettibile ma sono immuni alla malattia per i primi mesi di vita a causa della protezione degli anticorpi materni. Questo fenomeno può essere mostrato includendo una classe M (per l'immunità derivata dalla madre) all'inizio del modello. Nel momento in cui scompaiono gli anticorpi, l'individuo diventa suscettibile. Se entra in contatto con una persona malata, l'individuo passa dalla categoria S a I essendosi infettato. Nel momento in cui guarisce però, guadagna l'immunità dal virus. Per malattie che si sviluppano seguendo queste tappe si parla di modello MSIR.

Può succedere però che quando gli anticorpi materni scompaiono dal corpo, l'individuo diventa suscettibile, S. Quando l'individuo entra in contatto per un tempo sufficientemente lungo con un infetto, il suscettibile entra nella classe E degli esposti. Questo compartimento raccoglie gli individui che hanno contratto la malattia, ma che non sono ancora in grado di diffonderla. Al termine del periodo di latenza, l'esposto entra nella classe degli infetti I. Alla fine del periodo infettivo, l'individuo entra nella classe dei rimossi R quindi acquisisce l'immunità permanente dalla malattia. In tal caso invece si parla di modello MSEIR.

## Modello SEIR

Lo studio fatto relativo al modello SEIR è stato possibile a partire dalle informazioni riportate da Wikipedia alle quali è stato aggiunto un contributo personale.

Una variazione del modello SIR è il cosiddetto modello SEIR. L'unica differenza fra i due modelli è che l'individuo suscettibile nel momento in cui è stato infettato, non viene spostato direttamente nella classe degli infetti, perchè potrebbe ancora non essere infettivo. Durante questo periodo l'individuo si trova nella classe degli esposti (E). Solo nel momento in cui diventa un vero e proprio infetto allora entra a far parte della popolazione della classe I con un *tasso di contagio*  $\sigma > 0$ .

Le equazioni che rappresentano il modello SEIR endemico sono:

$$\begin{cases} \frac{dS}{dt} = -\beta SI - \mu S + \mu N \\ \frac{dE}{dt} = \beta SI - \sigma E - \mu E \\ \frac{dI}{dt} = \sigma E - \gamma I - \mu I \\ \frac{dR}{dt} = \gamma I - \mu R, \end{cases} \quad (3.68)$$

dove i parametri  $\beta > 0$ ,  $\gamma > 0$  e  $\mu > 0$ , rappresentano rispettivamente il *tasso di infezione* della malattia, il *tasso di recupero* dalla categoria degli infetti e il *tasso di natalità e mortalità*, con condizioni iniziali  $S(0) = S_0 > 0$ ,  $E(0) = E_0 > 0$ ,  $I(0) = I_0 > 0$  e  $R(0) = 0$ .

L'equazione che descrive il modello SEIR è

$$N = S(t) + E(t) + I(t) + R(t), \quad \forall t \geq 0$$

e, di conseguenza, il numero di rimossi è pari a  $R(t) = N - S(t) - E(t) - I(t)$ .

**Teorema 3.29** Sia  $J_{SEIR} = \langle -\frac{\beta\sigma}{\gamma+\mu}SE - \mu S + \mu N, \frac{\beta\sigma}{\gamma+\mu}SE - (\sigma + \mu)E \rangle$  l'ideale associato al modello SEIR (3.68) nell'anello  $\mathbb{R}[S, E]$ , con l'ordine monomiale *lex*  $S > E$ . Sia

$$R_0 = \frac{N\beta\sigma}{(\sigma + \mu)(\gamma + \mu)}. \quad (3.69)$$

Allora:

i. la base di Groebner ridotta di  $J_{SEIR}$  è

$$G_{rid} = \left\{ S + \frac{\sigma + \mu}{\mu}E - N, E^2 - \frac{\mu(\gamma + \mu)}{\beta\sigma}(R_0 - 1)E \right\};$$

ii. l'insieme dei punti di equilibrio è  $V_{SEIR} = \{P_1, P_2\}$  dove

$$P_1 = (N, 0, 0, 0) \quad \text{e} \quad P_2 = \left( \frac{N}{R_0}, \frac{\gamma + \mu}{\sigma} \frac{\mu}{\beta} (R_0 - 1), \frac{\mu}{\beta} (R_0 - 1), \frac{\gamma}{\beta} (R_0 - 1) \right).$$

*Dimostrazione.* Per la Def. 3.8 i punti di equilibrio per (3.68) sono tutti i punti che soddisfano il sistema seguente:

$$\begin{cases} -\beta SI - \mu S + \mu N = 0 \\ \beta SI - \sigma E - \mu E = 0 \\ \sigma E - \gamma I - \mu I = 0 \\ \gamma I - \mu R = 0. \end{cases} \quad (3.70)$$

Osserviamo che da (3.70) ricaviamo le equazioni

$$I_e = \frac{\sigma}{\gamma + \mu} E_e \quad (3.71)$$

e

$$R_e = \frac{\gamma}{\mu} I_e. \quad (3.72)$$

Sostituendo la (3.71) in (3.72) otteniamo

$$R_e = \frac{\gamma\sigma}{\mu(\gamma + \mu)} E_e. \quad (3.73)$$

Sfruttando (3.71) e (3.73), per determinare i punti di equilibrio di (3.68) cerchiamo tutti i punti che soddisfano il sistema

$$\begin{cases} -\frac{\beta\sigma}{\gamma + \mu} SE - \mu S + \mu N = 0 \\ \frac{\beta\sigma}{\gamma + \mu} SE - (\sigma + \mu)E = 0 \end{cases} \quad (3.74)$$

o, equivalentemente, tutti i punti che appartengono a  $V(G_{rid})$ , dove  $G_{rid}$  è la base di Groebner ridotta per l'ideale  $\langle -\frac{\beta\sigma}{\gamma + \mu} SE - \mu S + \mu N, \frac{\beta\sigma}{\gamma + \mu} SE - (\sigma + \mu)E \rangle \in \mathbb{R}[S, E]$ . Una base di Groebner per l'ideale di riferimento è

$$G = \left\{ -\frac{\beta\sigma}{\gamma + \mu} SE - \mu S + \mu N, \frac{\beta\sigma}{\gamma + \mu} SE - (\sigma + \mu)E, \frac{\mu(\gamma + \mu)}{\beta\sigma} S + \frac{(\sigma + \mu)(\gamma + \mu)}{\beta\sigma} E - \frac{\mu(\gamma + \mu)}{\beta\sigma} N, -\frac{\sigma + \mu}{\mu} E^2 + \frac{N\beta\sigma - (\sigma + \mu)(\gamma + \mu)}{\beta\sigma} E \right\}.$$

Per dimostrare tale fatto basta applicare l'algoritmo di Buchberger (vedasi Teor. 2.43) all'ideale  $\langle -\frac{\beta\sigma}{\gamma + \mu} SE - \mu S + \mu N, \frac{\beta\sigma}{\gamma + \mu} SE - (\sigma + \mu)E \rangle$ .

Consideriamo la base

$$G' = \left\{ -\frac{\beta\sigma}{\gamma + \mu} SE - \mu S + \mu N, \frac{\beta\sigma}{\gamma + \mu} SE - (\sigma + \mu)E \right\}.$$

Osserviamo che

$$\begin{aligned} & \mathcal{S}\left(-\frac{\beta\sigma}{\gamma+\mu}SE - \mu S + \mu N, \frac{\beta\sigma}{\gamma+\mu}SE - (\sigma + \mu)E\right) = \\ & = \frac{\mu(\gamma + \mu)}{\beta\sigma}S + \frac{(\sigma + \mu)(\gamma + \mu)}{\beta\sigma}E - \frac{\mu(\gamma + \mu)}{\beta\sigma}N = \\ & \overline{\frac{\mu(\gamma + \mu)}{\beta\sigma}S + \frac{(\sigma + \mu)(\gamma + \mu)}{\beta\sigma}E - \frac{\mu(\gamma + \mu)}{\beta\sigma}N}^{G'} \\ & = \mathcal{S}\left(-\frac{\beta\sigma}{\gamma+\mu}SE - \mu S + \mu N, \frac{\beta\sigma}{\gamma+\mu}SE - (\sigma + \mu)E\right) . \end{aligned}$$

Per il Teor. 2.41,  $G'$  non è una base di Groebner per l'ideale di riferimento.

Ora, sia

$$\begin{aligned} G'' = \{ & -\frac{\beta\sigma}{\gamma+\mu}SE - \mu S + \mu N, \frac{\beta\sigma}{\gamma+\mu}SE - (\sigma + \mu)E, \\ & \frac{\mu(\gamma + \mu)}{\beta\sigma}S + \frac{(\sigma + \mu)(\gamma + \mu)}{\beta\sigma}E - \frac{\mu(\gamma + \mu)}{\beta\sigma}N\}. \end{aligned}$$

E' ovvio che

$$\overline{\mathcal{S}\left(-\frac{\beta\sigma}{\gamma+\mu}SE - \mu S + \mu N, \frac{\beta\sigma}{\gamma+\mu}SE - (\sigma + \mu)E\right)}^{G''} = 0.$$

Inoltre,

$$\begin{aligned} & \mathcal{S}\left(-\frac{\beta\sigma}{\gamma+\mu}SE - \mu S + \mu N, \frac{\mu(\gamma + \mu)}{\beta\sigma}S + \frac{(\sigma + \mu)(\gamma + \mu)}{\beta\sigma}E - \frac{\mu(\gamma + \mu)}{\beta\sigma}N\right) = \\ & = \frac{\mu(\gamma + \mu)}{\beta\sigma}S - \frac{\sigma + \mu}{\mu}E^2 + NE - \frac{\mu(\gamma + \mu)}{\beta\sigma}N \end{aligned}$$

e

$$\begin{aligned} & \overline{\mathcal{S}\left(-\frac{\beta\sigma}{\gamma+\mu}SE - \mu S + \mu N, \frac{\mu(\gamma + \mu)}{\beta\sigma}S + \frac{(\sigma + \mu)(\gamma + \mu)}{\beta\sigma}E - \frac{\mu(\gamma + \mu)}{\beta\sigma}N\right)}^{G''} = \\ & = -\frac{\sigma + \mu}{\mu}E^2 + \frac{N\beta\sigma - (\sigma + \mu)(\gamma + \mu)}{\beta\sigma}E. \end{aligned}$$

Infine,

$$\begin{aligned} & \mathcal{S}\left(\frac{\beta\sigma}{\gamma+\mu}SE - (\sigma + \mu)E, \frac{\mu(\gamma + \mu)}{\beta\sigma}S + \frac{(\sigma + \mu)(\gamma + \mu)}{\beta\sigma}E - \frac{\mu(\gamma + \mu)}{\beta\sigma}N\right) = \\ & = -\frac{\sigma + \mu}{\mu}E^2 + \frac{N\beta\sigma - (\sigma + \mu)(\gamma + \mu)}{\beta\sigma}E = \end{aligned}$$

$$= \overline{\mathcal{S}\left(\frac{\beta\sigma}{\gamma+\mu}SE - (\sigma+\mu)E, \frac{\mu(\gamma+\mu)}{\beta\sigma}S + \frac{(\sigma+\mu)(\gamma+\mu)}{\beta\sigma}E - \frac{\mu(\gamma+\mu)}{\beta\sigma}N\right)}^{G''}.$$

Per il Teor. 2.41,  $G''$  non è una base di Groebner per l'ideale di riferimento.

Ora, sia

$$G = \left\{ -\frac{\beta\sigma}{\gamma+\mu}SE - \mu S + \mu N, \frac{\beta\sigma}{\gamma+\mu}SE - (\sigma+\mu)E, \right. \\ \left. \frac{\mu(\gamma+\mu)}{\beta\sigma}S + \frac{(\sigma+\mu)(\gamma+\mu)}{\beta\sigma}E - \frac{\mu(\gamma+\mu)}{\beta\sigma}N, -\frac{\sigma+\mu}{\mu}E^2 + \frac{N\beta\sigma - (\sigma+\mu)(\gamma+\mu)}{\beta\sigma}E \right\}.$$

Calcolando  $\mathcal{S}(g_i, g_j)$  per ogni  $g_i, g_j \in G$  osserviamo che  $\overline{\mathcal{S}(g_i, g_j)}^G = 0$  per ogni  $g_i, g_j \in G$ . Di conseguenza per il criterio di Buchberger (vedasi Teor. 2.41),  $G$  è una base di Groebner per  $\langle -\frac{\beta\sigma}{\gamma+\mu}SE - \mu S + \mu N, \frac{\beta\sigma}{\gamma+\mu}SE - (\sigma+\mu)E \rangle$ .

Osserviamo che

$$\text{LT}\left(-\frac{\beta\sigma}{\gamma+\mu}SE - \mu S + \mu N\right) = -\text{LT}\left(\frac{\beta\sigma}{\gamma+\mu}SE - (\sigma+\mu)E\right) = \\ = -\frac{(\beta\sigma)^2 E}{\mu(\gamma+\mu)^2} \text{LT}\left(\frac{\mu(\gamma+\mu)}{\beta\sigma}S + \frac{(\sigma+\mu)(\gamma+\mu)}{\beta\sigma}E - \frac{\mu(\gamma+\mu)}{\beta\sigma}N\right).$$

Di conseguenza per il Lemma 2.45 possiamo eliminare dalla base  $G$  i polinomi  $-\frac{\beta\sigma}{\gamma+\mu}SE - \mu S + \mu N$  e  $\frac{\beta\sigma}{\gamma+\mu}SE - (\sigma+\mu)E$ .

Rendendo monici i restanti polinomi della base  $G$ , per la Def. 2.50 la base di Groebner ridotta

$$G_{rid} = \left\{ S + \frac{\sigma+\mu}{\mu}E - N, E^2 - \frac{\mu(\gamma+\mu)}{\beta\sigma}(R_0 - 1)E \right\}.$$

Dalle equazioni (3.71) e (3.73) e calcolando  $V(G_{rid})$  otteniamo che i punti di equilibrio per il sistema (3.68) sono :

$$P_1 = (N, 0, 0, 0) \quad \text{e} \quad P_2 = \left( \frac{N}{R_0}, \frac{\gamma+\mu}{\sigma} \frac{\mu}{\beta}(R_0 - 1), \frac{\mu}{\beta}(R_0 - 1), \frac{\gamma}{\beta}(R_0 - 1) \right).$$

□

Ci occupiamo ora della stabilità dei punti di equilibrio trovati nel teorema precedente.

**Teorema 3.30** Sia  $V_{SEIR} = \{P_1, P_2\}$  la varietà associata al modello SEIR (vedasi Teor. 3.29). Allora, nel caso di coordinate non negative (vedasi Oss. 3.22), si ha che:

i. supponiamo che

$$(\gamma + 2\mu + \sigma)^2 \geq 4(\gamma + \mu)(\sigma + \mu)(1 - R_0). \quad (3.75)$$

Se  $R_0 < 1$  allora  $P_1$  è un punto di equilibrio asintoticamente stabile; se  $R_0 > 1$  allora  $P_1$  è un punto instabile.

ii. se la (3.75) non è soddisfatta, il punto di equilibrio  $P_1$  è asintoticamente stabile;

iii.  $P_2$  è un punto di equilibrio asintoticamente stabile.

*Dimostrazione.* Per poter applicare i teoremi di stabilità e quindi determinare se  $P_1$  e  $P_2$  sono stabili o instabili dobbiamo trovare la matrice Jacobiana relativa al sistema (3.68), calcolarla nei punti di equilibrio e determinarne gli autovalori.

La matrice Jacobiana del sistema (3.68) è:

$$J(S, E, I, R) = \begin{pmatrix} -\beta I - \mu & 0 & -\beta S & 0 \\ \beta I & -(\sigma + \mu) & \beta S & 0 \\ 0 & \sigma & -(\gamma + \mu) & 0 \\ 0 & 0 & \gamma & -\mu \end{pmatrix}.$$

Soffermiamoci ad analizzare il punto di equilibrio  $P_1 = (N, 0, 0, 0)$ . La matrice Jacobiana calcolata in tale punto è

$$J(N, 0, 0, 0) = \begin{pmatrix} -\mu & 0 & -\beta N & 0 \\ 0 & -(\sigma + \mu) & \beta N & 0 \\ 0 & \sigma & -(\gamma + \mu) & 0 \\ 0 & 0 & \gamma & -\mu \end{pmatrix}. \quad (3.76)$$

Gli autovalori della matrice (3.76) sono gli zeri del polinomio

$$p(\lambda) = \begin{vmatrix} -\mu - \lambda & 0 & -\beta N & 0 \\ 0 & -(\sigma + \mu + \lambda) & \beta N & 0 \\ 0 & \sigma & -(\gamma + \mu + \lambda) & 0 \\ 0 & 0 & \gamma & -(\mu + \lambda) \end{vmatrix} = \quad (3.77)$$

$$= (\mu + \lambda)^2 [\lambda^2 + (\gamma + 2\mu + \sigma)\lambda + (\sigma + \mu)(\gamma + \mu)(1 - R_0)],$$

dunque, gli autovalori sono  $\lambda_1 = -\mu < 0$  e

$$\lambda_{2,3} = \frac{1}{2} \left[ -(\gamma + 2\mu + \sigma) \mp \sqrt{(\gamma + 2\mu + \sigma)^2 - 4(\sigma + \mu)(\gamma + \mu)(1 - R_0)} \right].$$

Gli autovalori  $\lambda_{2,3}$  sono reali se e solo se vale (3.75).

Se vale la (3.75) allora

$$\lambda_2 = \frac{1}{2} \left[ -(\gamma + 2\mu + \sigma) - \sqrt{(\gamma + 2\mu + \sigma)^2 - 4(\sigma + \mu)(\gamma + \mu)(1 - R_0)} \right] < 0$$

indipendentemente da  $R_0$ , mentre

$$\lambda_3 = \frac{1}{2} \left[ -(\gamma + 2\mu + \sigma) + \sqrt{(\gamma + 2\mu + \sigma)^2 - 4(\sigma + \mu)(\gamma + \mu)(1 - R_0)} \right] > 0$$

se e solo se

$$R_0 > 1.$$

Dunque, se  $R_0 > 1$ , allora necessariamente un autovalore è positivo e gli altri negativi e, di conseguenza,  $P_1$  è un punto instabile per il Teor. 3.12.

Se  $R_0 < 1$  tutti gli autovalori sono negativi e, per il Teor. 3.11,  $P_1$  è un punto di equilibrio asintoticamente stabile.

Se la (3.75) non è soddisfatta, gli autovalori  $\lambda_2$  e  $\lambda_3$  della matrice Jacobiana (3.76) non sono reali. Dobbiamo allora considerare la parte reale degli autovalori. In particolare,  $\text{Re}(\lambda_{2,3}) = -\frac{\gamma+2\mu+\sigma}{2} < 0$  indipendentemente da  $R_0$ . Di conseguenza, se la (3.75) non è soddisfatta allora la parte reale di tutti gli autovalori della matrice (3.76) è negativa quindi per il Teor. 3.11 il punto di equilibrio  $P_1$  è asintoticamente stabile.

Consideriamo ora il secondo punto di equilibrio

$$P_2 = \left( \frac{N}{R_0}, \frac{\gamma + \mu}{\sigma} \frac{\mu}{\beta} (R_0 - 1), \frac{\mu}{\beta} (R_0 - 1), \frac{\gamma}{\beta} (R_0 - 1) \right),$$

per il sistema (3.68).

Ricordiamo che per ipotesi vale l'Oss. 3.22, studieremo solo il caso  $R_0 > 1$ .

La matrice Jacobiana calcolata in  $P_2$  è

$$J(P_2) = \begin{pmatrix} -\mu R_0 & 0 & -\beta \frac{N}{R_0} & 0 \\ \mu(R_0 - 1) & -(\sigma + \mu) & \beta \frac{N}{R_0} & 0 \\ 0 & \sigma & -(\gamma + \mu) & 0 \\ 0 & 0 & \gamma & -\mu \end{pmatrix}. \quad (3.78)$$

Gli autovalori della matrice Jacobiana (3.78) sono gli zeri del polinomio

$$p(\lambda) = \begin{vmatrix} -\mu R_0 - \lambda & 0 & -\beta \frac{N}{R_0} & 0 \\ \mu(R_0 - 1) & -(\sigma + \mu + \lambda) & \beta \frac{N}{R_0} & 0 \\ 0 & \sigma & -(\gamma + \mu + \lambda) & 0 \\ 0 & 0 & \gamma & -(\mu + \lambda) \end{vmatrix} = \quad (3.79)$$

$$= (-\mu - \lambda)(-\lambda^3 - b\lambda^2 - a\lambda - d),$$

dove  $a = \mu R_0(\gamma + 2\mu + \sigma)$ ,  $b = \gamma + \mu(2 + R_0) + \sigma$  e  $d = \mu(\gamma + \mu)(\sigma + \mu)(R_0 - 1)$ , ottenuti sfruttando la (3.69). Osserviamo che  $a > 0$ ,  $b > 0$  e  $d > 0$  dato che  $R_0 > 1$ .

E' facile osservare che un autovalore del polinomio (3.79) è  $\lambda_1 = -\mu < 0$ . Per poter capire se la parte reale degli altri tre autovalori della matrice (3.78) è positiva o negativa, possiamo applicare *la regola dei segni di Cartesio* al polinomio

$$g(\lambda) := -\lambda^3 - b\lambda^2 - a\lambda - d \quad (3.80)$$

dove  $a, b, d$  sono definiti in (3.79).

Ricordiamo che dato un polinomio  $g(x)$  di grado  $n$  a coefficienti reali e non tutti nulli, la regola di Cartesio stabilisce che il massimo numero di radici reali positive di un polinomio è dato dal numero di variazioni di segno fra coefficienti consecutivi, trascurando eventuali coefficienti nulli. L'informazione relativa al numero di radici negative si deduce applicando la stessa regola al polinomio trasformato a radici opposte, ossia al polinomio  $g(-x)$ . Esso ha infatti radici opposte a quelle di  $g(x)$ , perciò le variazioni relative ai coefficienti del polinomio  $g(-x)$  danno informazioni circa le sue radici positive e, di conseguenza, circa le radici negative di  $g(x)$ .

Consideriamo il polinomio  $g(\lambda)$  definito in (3.80). Tale polinomio non ha variazioni di segno quindi non ci sono zeri reali positivi, mentre  $g(-\lambda)$  ha 3 variazioni di segno, quindi ci sono 1 o 3 zeri reali negativi. Di conseguenza, ci sono 0 o 3 autovalori reali negativi o 1 reale negativo e 2 complessi coniugati.

Se gli autovalori sono reali e negativi allora, tenendo presente che  $\lambda_1 < 0$ , per il Teor. 3.11,  $P_2$  è un punto di equilibrio asintoticamente stabile per il sistema (3.68). Se (3.80) ammette 1 autovalore reale negativo e 2 complessi coniugati, supponiamo che gli autovalori complessi coniugati siano  $\lambda_2 = u + iv$  e  $\lambda_3 = u - iv$ , con  $u, v$  reali e  $v \neq 0$ . Sappiamo che  $\lambda_2$  è un autovalore, dunque  $g(\lambda_2) = 0$ . Ricaviamo allora che

$$-u^3 - bu^2 - (a - 3v^2)u - (d - bv^2) = 0 \quad (3.81)$$

e

$$-3u^2 + v^2 - 2bu - a = 0. \quad (3.82)$$

Dalla (3.82) otteniamo  $v^2 = 3u^2 + 2bu + a$  e sostituendola nella (3.81) ricaviamo

$$f(u) = 8u^3 + 8bu^2 + (2a + 2b^2)u + (ab - d) = 0, \quad (3.83)$$

dove  $ab - d = \mu R_0(\gamma + 2\mu + \sigma)(\gamma + \mu(2 + R_0) + \sigma) - \mu(\gamma + \mu)(\sigma + \mu)(R_0 - 1) > 0$ , o equivalentemente  $\mu R_0(\gamma + 2\mu + \sigma)(\gamma + \mu(2 + R_0) + \sigma) > \mu(\gamma + \mu)(\sigma + \mu)(R_0 - 1)$ , dato che  $R_0, \mu, \gamma, \sigma > 0$  e dunque  $R_0 > R_0 - 1$ ,  $\gamma + 2\mu + \sigma > \gamma + \mu$  e  $\gamma + \mu(2 + R_0) + \sigma > \mu + \sigma$ . Quindi il polinomio di terzo grado in  $u$ , non avendo variazioni di segno, non ha zeri  $u$  positivi; di conseguenza, gli eventuali autovalori di (3.78) complessi coniugati hanno parte reale negativa: sapendo che due autovalori sono reali e negativi e che la parte reale degli autovalori complessi e coniugati di (3.78) è negativa, per il Teor. 3.11 si ha che  $P_2$  è un punto di equilibrio asintoticamente stabile per il sistema (3.68).

□

**Osservazione 3.31** Se  $R_0 = 1$ , la stabilità del punto di equilibrio  $P_1$  resta un problema aperto. A dire il vero, non è particolarmente rilevante dal punto di vista pratico studiare il caso  $R_0 = 1$ . Infatti, difficilmente i parametri fisici assumono valori per cui  $R_0$  sia uguale esattamente a quel valore. In realtà ciò che è importante è che ci sia una transizione tra due regimi completamente diversi.

**Osservazione 3.32** Se  $R_0 < 1$  le traiettorie vanno verso  $P_1$  (la malattia non è abbastanza forte da diffondersi), se  $R_0 > 1$  le traiettorie vanno verso  $P_2$  (i rimossi aumentano perché gli infetti o guariscono o muoiono, i suscettibili diventano pochi perché si sono in gran parte ammalati).

Come è stato possibile osservare dallo studio dei modelli epidemiologici sopra riportati, i punti di equilibrio di ciascun sistema che rappresenta un qualsiasi modello epidemiologico endemico sono un numero finito. Questo risultato non è valido però per i punti di equilibrio dei modelli epidemici. Sorge dunque spontaneo chiedersi se vi è un motivo specifico come spiegazione di questa osservazione. Sarebbe interessante approfondire la questione.

# Bibliografia

- [C] L.Caporaso, *Note del corso di Geometria Algebrica*.
- [CA] G.Catania. Tesi di Laurea, *Basi di Groebner e le loro applicazioni*, Università degli studi di Milano, Facoltà di Scienze Matematiche, Fisiche e Naturali, Corso di laurea in Matematica, a.a. 2003/2004.
- [CAC] D.Caccavo. *Chinese and Italian COVID – 19, outbreaks can be correctly described by a modified SIRD model*. 2020.
- [CLO] D.Cox, J.Little and D.O’Shea. *Ideals, varieties and algorithms, an Introduction to Computational Algebraic Geometry and Commutative Algebra*, Springer, 2007.
- [F] R.Fassina. *Modelli matematici per lo studio delle epidemie*, Tesina di Laurea in Ingegneria dell’Informazione, a.a. 2012/2013.
- [G] A.Gathmann. *Commutative Algebra*, Class Notes TU Kaiserslautern,2013/2014.
- [GE] G.Gentile. *Equazioni differenziali ordinarie analisi qualitativa e alcune applicazioni*.
- [HOAG] D. Hincapié P., J. Ospina G., A. Uyi Afuwape and R.D.Gomez A. *Epidemic Thresholds in SIR and SIIR Models Applying an Algorithmic Method*.

- [HS] M.W. Hirsch and S.Smale. *Differential Equations, Dynamical Systems, and Linear Algebra*. Academic Press.
- [MBK] B.D. MacCluer, P.S. Bourdon and T.L. Kriete. *Differential Equations : Techniques Theory and Applications*, American Mathematical Society, Cap. 10.
- [P] C.Procesi. 2002. URL: <http://www1.mat.uniroma1.it/people/procesi/lavori/capit3.pdf>
- [PS] L. Pachter, B. Sturmfels. *Algebraic Statistics for Computational Biology*. Cambridge University Press, 2005.

# Ringraziamenti

*A tutti voi, così diversi ma così importanti, ognuno per ragioni uniche e speciali, voglio esprimere la mia più assoluta gratitudine.*

*Al mio **relatore**, il Prof. Angelo Felice Lopez. Vorrei farle i miei più sentiti ringraziamenti per avermi aiutato a portare a termine questo percorso universitario. Un pò severo, rigido, apparentemente distaccato ma in questi mesi non mi ha mai lasciata sola. Mi ha spronato ogni giorno a chiedermi di più, a pretendere di più. La ringrazio profondamente perchè oltre ad avermi aiutata nella stesura della tesi, mi ha sostenuta e ha creduto profondamente in me. Per tutto questo non potrò mai esserle abbastanza riconoscente.*

*Vorrei inoltre ringraziare il Prof. G. Gentile. Il suo aiuto mi ha permesso di perfezionare la parte di analisi relativa alla tesi.*

*A **Mamma** e **Papà**. Sì, proprio così, con le lettere maiuscole. Ci vuole poco a essere una mamma e un papà ma fare i genitori è un'altra storia. Mi avete sopportata, sopportata, calmata e incoraggiata sempre. Non mi avete mai fatto mancare niente, mi siete sempre stati accanto e mi avete sempre sostenuta in tutte le mie scelte, anche se a volte non rispecchiavano le vostre. Ogni mio successo o insuccesso è allo stesso tempo una vostra gioia o un vostro dispiacere.*

*A te, mamma, voglio dire che sei la mia roccia, mi hai insegnato a non arrendermi mai ma, soprattutto, “a usare la testa e non i piedi”. A te che sei sempre dalla mia parte voglio dire che sei il mio esempio.*

*A te papà, voglio dire che mi hai insegnato a pensare positivo, ad essere ottimista ed a sorridere alla vita.*

*Non vi ho mai detto quanto io sia grata di quello che mi avete insegnato e di quanto io sia orgogliosa e fortunata ad avere dei genitori come voi. Vorrei ringraziarvi dal più profondo del mio cuore. Vi amo follemente e questo traguardo è anche vostro.*

*A mia sorella, **Sara**. A lei che ancora deve perdonarmi che ai suoi 3 anni le ho rubato il giorno del compleanno. Siamo il giorno e la notte, a volte sembra che apparteniamo a due mondi separati eppure per me sei più di una sorella. Anche se non ci sentiamo ogni giorno, tu ci sei stata, ci sei e ci sarai sempre per me. A volte è difficile sopportarti viste le mie “malattie per l’ordine” ma sappi che non c’è nulla che non farei per te.*

*Alla mia Famiglia, a tutti i miei parenti ma soprattutto ai miei nonni, **Aldo e Gino**, che sarebbero stati orgogliosi di me, e alle mie nonne Marisa e Pia.*

*A te, **nonna Marisa**, voglio dire grazie perchè sei sempre stata presente, non ti sei mai persa niente, da quando ero piccola ad oggi, dallo sport alla scuola. A volte ti ricordi più cose tu di papà.. (A proposito.. Ora papà starà pensando: “ecco ora dovrebbe dire anche una cosa positiva”). Ti voglio bene.*

*A te, **nonna Pia**, voglio dire che sei la semplicità e la dolcezza in persona. Anche se a volte non ci puoi essere perchè la salute non lo permette, so che mi sei stata e sei vicina lo stesso. Ti voglio bene!*

*Alle mia maestra delle elementari e alle Professoresse delle medie e del Liceo che mi hanno trasmesso la loro passione e hanno fatto sì che la matematica non fosse solo una materia per me ma un “modo per affrontare la vita”.*

*Vorrei ringraziare tutte le mie Amiche, le mie sorelle acquisite: da quelle che porto nel mio cuore dall’infanzia a quelle che ho incontrato nel mio cammino e che hanno deciso di camminare al mio fianco e di portarmi fino a qui.*

*A te **Veronica**. A te che sono ben 15 anni che mi sei vicina, a te che posso dire di conoscerti da quasi tutta la vita. Dal mini-volley al matrimonio, rende l’idea? Quanto saremmo state alte? Un po’ più di un metro? Eppure siamo sempre le stesse, solo ancora più legate e qualche kilo in più. Abbiamo condiviso tutto, segreti di tutti i tipi e avventure di ogni genere. Con te so di poter parlare di tutto e non sentirmi mai giudicata. Mi hai dimostrato che tutto può succedere, un giorno qualcuno arriva e ti stravolge la vita nel migliore dei modi. Tante persone al tuo posto sarebbero cambiate invece tu hai dimostrato chi sei davvero, rimanendo con i “piedi per terra”. Non sei cambiata nonostante tutto e mi sei stata sempre accanto facendomi sentire onorata con il ruolo di “testimone di nozze”. Hai dovuto imparare ad essere distaccata e diffidente ma, io che ho la fortuna di conoscerti, lo urlerei al mondo che sei una persona d’oro.*

A te **Elisa**. “A te che sei una roccia, una pianta, un uragano, sei l’orizzonte che mi accoglie quando mi allontano”. Sono 10 anni che ci conosciamo, ne hai passate tante ma, lo scrivo e non lo nego, ne abbiamo passate tante insieme. Nessuno alla nostra età può capire cosa significa ed è per questo che ci siamo legate così, come due sorelle. Sei stata, sei e sarai il mio “diario segreto”, custodendo i miei segreti più profondi, belli e brutti che siano. Mi hai fatto sfogare nei momenti bui e mi hai aiutata a tirarmi su. Con te ho trovato una sorella, una certezza. Sister mi hai regalato anche un nipotino favoloso, non potevo chiedere di meglio.

A **Lorena**. Non sei stata una semplice istruttrice, per me sei una confidente speciale. Un’amica con cui uscire ma anche una sorella più grande, pronta a farmi ragionare e insegnarmi. Mi hai sempre dimostrato la tua stima per questo mio percorso di laurea. Oltre ad essere un’amica, sei stata e sei una mia grande fan. Ti voglio bene!

A **Martina D**. A te che mi hai accompagnato in tutti gli anni del Liceo e non solo. Ricordo tutti i pomeriggi che venivi a casa e ci preparavamo per la verifica di matematica insieme fra una risata e l’altra. Per non parlare di quando studiavamo a casa tua in piscina. Sei stata fra le prime persone che ho aiutato davvero in matematica. Mi hai insegnato a mettermi in gioco e dimostrare le mie capacità. Forse non sono riuscita a farti diventare un piccolo genio in matematica è vero, però siamo state una bella squadra! Ti voglio bene roscia.

A **Valentina e Serena**. Siete state il mio punto di riferimento in questi 5 anni. Vi ho conosciute il primo giorno e abbiamo condiviso paure, ansie pre-esame e felicità post-esame. Centinaia di ore passate insieme in facoltà, a lezione, ai seminari e non solo. Ci siamo aiutate tanto, tutti gli esami li abbiamo preparati insieme. Arrivare alla laurea senza di voi non avrebbe avuto senso. “Squadra vincente non si cambia!” Ecco voi siete la mia squadra, per me l’università siete voi. Questo traguardo è anche vostro.

A **Irene**. Una bellissima persona con cui condividere le pause all’università e non solo. Tu che sei così piccina ma dal cuore grande. Mi hai dimostrato che non bisogna mai arrendersi e che con costanza e impegno si possono ottenere buonissimi risultati. Ti voglio bene!

A **Giorgia**. A te che sei una bomba di energia, che con la tua semplicità riesci sempre a sorridere e farmi sorridere. Dalla vacanza a Malaga ci siamo rese conto di quanto siamo simili e non ci siamo più separate. In questi pochi anni ci siamo ascoltate, incoraggiate e supportate e sei diventata così speciale per me.

A **Chiara**. *Per me non sei solo una compagna di palestra e passeggiate. E' vero siamo diverse, io sono quella più istintiva mentre tu sei quella più razionale. Per non parlare del fatto che io sono quella dolce e affettuosa mentre tu sei quella fredda e distaccata, che quasi ti senti male se ti do un abbraccio. Siamo la dimostrazione che "gli opposti si attraggono" e insieme ci completiamo. Nonostante ciò, di persone come te ce ne sono poche. Dietro quel faccino da angelo si nasconde una ragazza che con la sua semplicità e naturalezza potrebbe conquistare il mondo. Tutti dovrebbero avere un'amica come te. Ti voglio bene acidella!*

A **Ilaria**. *In poco tempo mi hai dimostrato che non si può rinunciare a passare del tempo con te. Che devo dirti? "Tu si a fine do' munno." Ti voglio bene!*

Alle "**Gatte**". *Le ultime ma non le meno importanti. A voi che in così poco tempo siete diventate Completamente speciali. "Siete arrivate all'improvviso e mi avete presa sempre più, come un uragano di vita. Mi avete presa dal mio sonno, scosso e riattivato il cuore." Abbiamo condiviso momenti belli e brutti ma sempre insieme. Voi ci siete sempre. Sempre pronte a tirarmi su in ogni momento. Auguro a tutti di costruire un legame con qualcuno come il nostro.*

*Infine, vorrei ringraziare tutte le persone che in un modo o nell' altro sono entrate a far parte della mia vita, soprattutto nel periodo universitario. Mi avete insegnato a crescere e avete contribuito a farmi diventare quella che sono oggi.*

*A tutti voi voglio dire GRAZIE.  
Questo traguardo è anche vostro.  
Vi voglio bene.*

Valeria