Chapter 10

Metabolic diversity in cell populations: probability densities over the flux polytope

Andrea De Martino and Marcelo Rivas-Astroza

Chapter highlights

Even in clonal populations, cells appear to be strongly heterogeneous in terms of, e.g., protein levels, RNA levels, sizes at birth or division, interdivision times and elongation rates. Part of this variability is likely due to the inherent stochasticity of gene expression at the level of single cells. It is however known that heterogeneous populations may possess an evolutionary advantage, for instance in variable environments or under stress. Despite appearing to be at odds with the idea of optimality presented in the previous chapters, metabolic diversity can be described and modeled within the constraint-based framework introduced in the previous chapters. Specifically, a statistical representation of heterogeneous populations can be obtained by defining suitable probability distributions on the flux polytope. This chapter addresses

- the different sources of variation that affect microbial metabolism along with the mechanisms that may favor higher variability,
- the methods devised to represent heterogeneous microbial populations within the framework of constraint-based models, and
- how these approaches connect to the optimality scenario presented in the previous chapters.

10.1 Introduction

The theory of cellular metabolism developed up to this point through constrained-based models (CBMs) relies crucially on some type of optimality assumption: among all viable flux states encoded in the flux polytope by mass-balance, thermodynamic and regulatory constraints, cells strive for those that maximise a physiologically motivated objective function. For *E. coli* cells growing on carbon-limited substrates, for instance, it is reasonable to take such a function to be the (specific) growth rate. At the very least, these optimal states provide reference points to gauge cellular behaviour. In this respect, having a good grasp of what makes a configuration of fluxes through the network 'optimal' with respect to a certain objective is rather important from a theoretical viewpoint. On the other hand, it is not easy to prove directly in an experiment that a certain function is *actually* being optimised (in any physical system, let alone in a microbe or a microbial population). The optimality assumption can usually be corroborated *a posteriori*, e.g. by comparing optimality predictions to experimentally measured fluxes or growth rates [342, 343], or indirectly, e.g. by showing that, in a given growth medium, certain metabolic enzymes are expressed at just the level ensuring maximal growth [344]. By

looking at the behaviour of individual cells in a population, however, one cannot help but notice a salient feature: their diversity. Individual cells are macroscopically heterogeneous in terms of parameters like interdivision times, elongation rates, sizes at birth or division, etc. This suggests that a corresponding diversity is present at the level of intracellular processes like cell cycle, gene expression and, of course, metabolism. Quantitative experiments probing populations at single-cell resolution (see Experimental method 10.1) can nowadays characterise such a diversity in some detail. Among the remarkable outcomes of these studies is that, when analysed with a lens that accounts for diversity, bacterial growth displays signatures of universality [345, 346, 347], suggesting the existence of general, system- and condition-independent control mechanisms (e.g. of cell division and growth) that do not change with specifics like strain, quality of medium, etc. Identifying these mechanisms yields robust insight (and predictive capacity) into the physiology of microbial systems (see also Chapter on *Control of cell division and coordination with other cell-cycle processes*).

Experimental methods 10.A: Quantitative methods for single-cell analysis of microbial systems

At the very minimum, quantitative experimental characterisation of cell-to-cell diversity in microbial populations requires (i) the possibility of achieving steady-state cell growth in controlled environments, and (ii) the possibility of identifying individual cells within a population. The two setups that are most important for the present chapter (and most widely used in general for the study of cell-to-cell heterogeneity in microbial systems) are the following.

- High-resolution optical microscopy of bacteria growing on agarose pads. Optical microscopy is the first and still most used technique to address cellular individuality [348]. Besides giving direct information about the macroscopic growth dynamics of individual cells [349, 347], it can be used in conjunction with gene expression reporters like fluorescent proteins to quantify diversity in gene expression levels [350] and dynamics [351]. Optical means usually allow to reliably follow the expression of a relatively small number of genes. In addition, however, they can also provide information about many other aspects of bacterial physiology, like motility, chemotaxis or the spatial self-organisation of colonies.
- Microfluidic 'lab-on-a-chip' devices. In essence, these techniques allow to confine single cells or small lineages thereof in controlled environments for long-term data acquisition [352]. A well-known example is the 'mother machine' [353]. In a mother machine cells grow in narrow (ca. 1 μm) microfluidic dead-end channels such that (a) all cells in the same channel are daughters of a mother cell stuck at the closed side of the channel; (b) a main feeding channel carries away cells that grow out of the length of dead-end channels (which suffice to contain a few cells, usually 5 to 10); and (c) nutrient in-flow and waste out-flow from the feeding channel ensure a constant medium in all dead-end channels via diffusion. This setup effectively keeps the population size fixed. Growing bacteria can then be imaged and analysed by standard means like time-lapse microscopy to obtain the statistics of quantities like the interdivision time or the size at birth at stationarity [346].

The setup of mother machines has the advantage that cells can be followed for many more generations than on agarose pads, since the latter tend to become overcrowded after a limited number of rounds of divisions. On the other hand, agarose pads offer a more natural environment for cell division.
In addition to these, a host of other techniques are being increasingly refined and used to probe single-cell properties and behaviour in bacterial populations, including single-cell metabolomics by mass spectrometry [354], nanoscale secondary ion mass spectrometry (nanoSIMS) [355], and single-cell transcriptomics [356].

It is not hard to guess why a bunch of identical cells sharing the same medium would, say, elongate at different rates. For one, gene expression has a stochastic component, from e.g. the random diffusion of transcription factors to targets to the thermal noise driving the on/off dynamics of transcription events. We also know that the cell cycle can be highly variable [357]. And other 'natural' sources of variance can be found in the dynamics of expression in genetic circuits, ageing, asymmetric partitioning of cellular resources at division, inter-cellular interactions, and epigenetic modifications [358]. In other terms, a degree of variability across a population is to be expected. The question, however, is, how can variability be reconciled with the optimality picture? And related to this: can we explain cell-to-cell differences in terms of some other, perhaps more involved, optimality criterion? Are there cases in which variability is optimised? Can we describe quantitatively a microbial population in ways that account for inter-cellular diversity? Note that cell-to-cell variability is inherently a population-level concept. Addressing it therefore requires a framework that is capable of clearly distinguishing single-cell properties from population-level ones.

It is definitely possible to explain cell-to-cell variability in the optimality framework (see Chapter on *Solutions of constraint-based metabolic models*). For example, one could say that, in appropriate conditions, all microbes in a population are

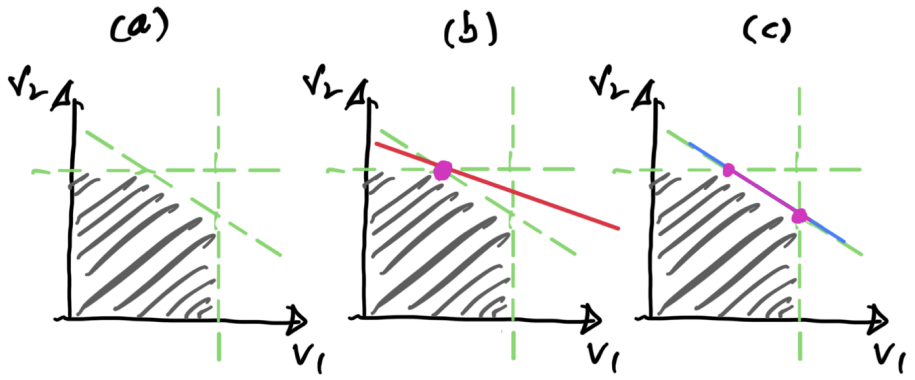

Figure 10.1: One versus multiple optima in the flux polytope. (a) A two-dimensional flux polytope (shaded area) with non-negative fluxes and the defining constraints shown as dashed green lines. (b) The linear objective function represented by the red line has a unique maximum (magenta dot). (c) The linear objective function represented by the blue line has infinitely many maxima, lying on the highlighted magenta segment.

optimal, but the optima are slightly different for different cells. As a matter of fact, optimal states in CBMs need not be isolated points belonging to the flux polytope. There can in fact be infinitely many flux vectors that maximise an objective function (this happens, for instance, when an objective function attains its maxima on one of the edges or faces of the polytope, see Figure 10.1). This implies that identical cells subject to the same constraints and sharing the same objective may end up having different metabolic profiles despite carrying the same value for the objective function. In this scenario, diversity is induced by a very special feature of the objective function and, unless some other ingredient is brought into the game to lift the degeneracy, all optimal states would be equally likely for cells. If having an objective function of this type seems unlikely in a high-dimensional setup such as metabolism, one may imagine a scenario in which all cells optimise the same objective but with slightly different constraints (i.e. in a slightly different polytope, e.g. due to small variations in regulatory constraints, energy demands, or nutrient uptakes). In this case, each cell would solve its own optimisation problem, ending up having, along with a different metabolic profile, a slightly different value of the objective function. Metabolic diversity is therefore induced by variability in the constraints. But it is also possible that, if cells are subject to fluctuating exogenous constraints (e.g. variable nutrient levels), they would prefer to maximise their, say, growth rate *averaged over conditions*, especially if fluctuations occur on faster timescales than those over which metabolic reactions equilibrate. In such a case, the average growth rate would be maximum (given the external variability), but other than that every cell could carry a different growth rate and a different metabolic profile. In this respect, one can say that diversity is now being optimally adapted to external conditions. Or one may even think that different cells have different objective functions. This scenario, possibly unrealistic for growing microbial populations but not for other cell types (think for instance of the mixture of neurons with high energy demands and glia with low energy demands in the brain), would also lead to heterogeneous flux profiles and objectives. And so on.

It is clear from these examples that, in order to represent heterogeneity within CBMs, one must, first and foremost, clarify the origin of heterogeneity as much as possible. Next, it is necessary to shift from the language of individual flux vectors belonging to the flux polytope to that of *ensembles* of flux vectors or, more reasonably for large populations, of *probability densities* defined on the flux polytope. This transition is less trivial and more momentous than it sounds and, together with the causes of variability, is the core subject of the present chapter. We shall begin by giving a more precise characterisation of the different types and sources of diversity that can be considered when modelling metabolic networks. Next, we shall introduce probability densities on the flux polytope and briefly discuss a few simple examples. We shall then address the general problem of using probability densities to model heterogeneity and uncertainty, most notably that seen in empirical data. Finally, we will show how these ideas can be used to generalise the notion of optimality to heterogeneous populations.

10.2 Sources of variability and uncertainty in metabolism

Metabolic heterogeneity is widespread among clonal populations of prokaryotic and eukaryotic cells. Populations of *Escherichia coli* display diverse cell-to-cell conversion yields of glucose into final products, such as fatty acids and tyrosine [359]; not surprisingly, the intracellular concentration of co-factors, including ATP, also vary significantly between cells [360]. *Saccharomyces cerevisiae* metabolic states have been observed to divert over time for each cell. For instance, a single budding yeast does uptake oxygen before the duplicating its genetic material, but it changes to an anaerobic metabolism once DNA duplication starts to prevent mutations related to free-radicals [361, 362]. Animal cells within a single tissue also show heterogeneous metabolisms. Non-small cell lung cancer show a remarkable diversity of preferred carbon sources. Within the tumour, some cells consume glucose and produce lactate, whereas others divert their metabolism to consume lactate as a carbon source [363].

The root cause of this metabolic heterogeneity is manifold, including uneven distribution of nutrients in the environment, asymmetric cell partitioning at division, and noise in gene expression [364, 365]. These effects are stochastic, preventing the determination of each cell's metabolic state in advance. This type of uncertainties are rooted in the nature of metabolism itself, thus we say it is objective uncertainty.

There is another type of uncertainty, the one that comes from our models of metabolism. There can be missing reactions in the metabolic network reconstruction [366], some reactions' directionality can not be anticipated under *in vivo* conditions [367], and experimental errors in the techniques used to measure metabolic quantities –such as exchange fluxes, or the weights of the biomass reaction– does not provide precise values but approximated intervals [368]. Even when using a *bona fide* metabolic network conditioned by precisely measured parameters, the use of optimality principles –e.g. maximisation of biomass growth– are typically implemented as under-determined mathematical formulations, whose output is not a single flux vector but rather a reduction of the viable polytope [369] (see Chapter on *Solutions of constraint-based metabolic models*). This is exemplified in Fig. 10.2.A, where the maximisation of the network fourth flux, v_4, can only reduce the fluxome space to the subspace defined by the line shown in Fig. 10.2.B. Uncertainties that are the result of modelling uncertainties can be categorized as subjective, as they arise solely from an observer's lack of knowledge.

Exercise 10.1 Use the model in Fig. 10.2 to study different objective functions, specifically combinations of fluxes. Can you find cases in which the optimum is not unique?

Although objective and subjective uncertainties arise from different sources, both can be modelled using probability theory as we will see in the following section.

10.3 Probability densities over the flux polytope

In what follows, we shall denote the convex flux polytope by $\mathcal{P}$ and a generic flux configuration in $\mathcal{P}$ by $\mathbf{v} = \{v_i\}_{i=1}^N$. A probability density p defined on $\mathcal{P}$ is any non-negative function such that

$$\int_{\mathcal{P}} p(\mathbf{v}) dv_1 \cdots dv_N \equiv \int_{\mathcal{P}} p(\mathbf{v}) d\mathbf{v} = 1 \quad . \tag{10.1}$$

Notice that the integral over $\mathcal{P}$ implicitly encodes two types of constraints: mass-balance equations (i.e. $\mathbf{Sv} = \mathbf{0}$) and ranges of variability of the form $v_{i,\min} \leq v_i \leq v_{i,\max}$ (see Chapter on *Solutions of constraint-based metabolic models*). The quantity $\int_{\mathcal{P}} d\mathbf{v}$ represents therefore the *a priori* volume of $\mathcal{P}$ (which, understandably, is far from simple to calculate for high-dimensional polytopes like those corresponding to genome-scale metabolic network reconstructions). As usual, $p(\mathbf{v})$ can be interpreted as the relative likelihood of flux configuration $\mathbf{v}$: if we imagine that a cell is assigned a flux

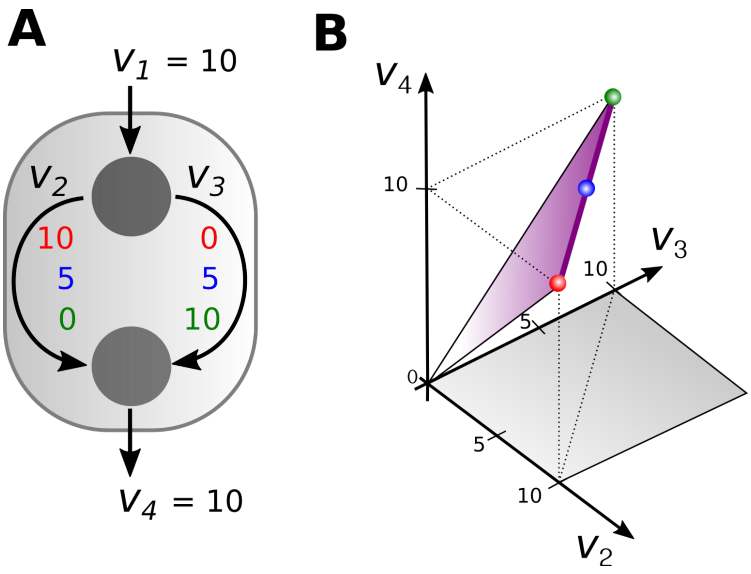

Figure 10.2: Minimal metabolic network with multiple optima. Toy network where the top metabolite has one producing reaction flux v_1 and two reactions fluxes, v_2 and v_3, that convert it into the lower metabolite, which is excreted via v_4. When fixing $v_1 = 10$, the maximisation of v_4 –a proxy of biomass growth rate– results in $v_4 = 10$. Even in this case, there are infinitely many points that are coherent with this solutions, among them the colour-coded as red blue and green (A). The subspace of solutions form a line (B).

configuration by "randomly sampling it from $\mathcal{P}$" using the rule described by p, then $p(\mathbf{v})d\mathbf{v}$ represents the probability that the cell's flux configuration will lie in a small volume $d\mathbf{v}$ around $\mathbf{v}$. It is clear then that probability densities on $\mathcal{P}$ provide a mathematically convenient way of describing the metabolic state of large populations (or ensembles) of cells at a given time, provided one can assume that cells have the same metabolic network and are subject to the same constraints, so that $\mathcal{P}$ is the same for all of them. For the population of cells described by p, the probability density clearly contains all the statistics of metabolic fluxes, from mean values to variances to correlations. For instance, by integrating p over all fluxes except the i-th, one obtains the marginal probability density of flux v_i, i.e.

$$\int_{\mathcal{P}} p(\mathbf{v}) d\mathbf{v}_{\setminus i} = p_i(v_i) \quad , \tag{10.2}$$

where the index $\setminus i$ corresponds to 'except for the flux of index i' (so $d\mathbf{v}_{\setminus i} = dv_1 \cdots dv_{i-1} dv_{i+1} \cdots dv_N$). And from p_i we can immediately retrieve the statistical features of flux v_i (e.g. mean value, variance, etc).

Let us make a few simple examples.

- If we assume that all cells in the population maximize the same objective function, and that there is no degeneracy in the optimal state, then
$$p(\mathbf{v}) = \delta(\mathbf{v} - \mathbf{v}^\star) \quad , \tag{10.3}$$
where $\mathbf{v}^\star$ denotes the (unique) objective-maximising flux vector and $\delta(x)$ denotes Dirac's δ-distribution.
- If we can make no assumption on the cells' metabolic activity other than it has to be compatible with the constraints encoded by $\mathcal{P}$, then any flux vector $\mathbf{v} \in \mathcal{P}$ is equally likely to occur in a population. This means that p is constant on $\mathcal{P}$. Specifically, its value must be equal to the inverse of the volume of $\mathcal{P}$:
$$p(\mathbf{v}) = \left(\int_{\mathcal{P}} d\mathbf{v}' \right)^{-1} \qquad (\mathbf{v} \in \mathcal{P}) \quad . \tag{10.7}$$
For any given flux polytope, this distribution can be sampled in practice using the methods described in the Chapter *The space of metabolic flux distributions*.
- Imagine having a dataset derived from a ^{13}C labelling experiment (mass spectrometry) that gives the mean value $\overline{v}_i$ of *every* flux in the network (the average being over the population of cells used in the experiment), together

Mathematical details 10.A: Dirac's δ-distribution

For our purposes, the defining property of the δ-distribution in one dimension is the following: if a variable x is δ-distributed around the finite value $x^\star$, then, for any continuous function f,

$$\int_{-\infty}^{+\infty} f(x)\delta(x-x^\star)dx = f(x^\star) \quad . \tag{10.4}$$

This means that, intuitively, $\delta(x) = 0$ everywhere on the real axis except at $x^\star$, where its value is $+\infty$. Such a function only makes sense under an integral sign. In this respect, (10.3) should be seen as an abuse of notation, albeit a convenient one. There are however several ways to represent the δ-distribution that comply with the above requirement. For example, one can define

$$\begin{aligned}\int_{-\infty}^{+\infty} f(x)\delta(x-x^\star)dx &:= \lim_{\sigma\to 0}\int_{-\infty}^{+\infty} f(x)\frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{(x-x^\star)^2}{2\sigma^2}}dx \\ &= \lim_{\sigma\to 0}\int_{-\infty}^{+\infty} f(x^\star+\sigma y)\frac{1}{\sqrt{2\pi}}\,e^{-\frac{y^2}{2}}dy = f(x^\star) \quad . \end{aligned} \tag{10.5}$$

The generalisation to $n > 1$ dimensions is obtained by straightforwardly assuming $\delta(\mathbf{x}-\mathbf{x}^\star) = \prod_{i=1}^n \delta(x_i - x_i^\star)$, so that

$$\int_{\mathbb{R}^n} f(\mathbf{x})\delta(\mathbf{x}-\mathbf{x}^\star)d\mathbf{x} = f(\mathbf{x}^\star) \quad . \tag{10.6}$$

Because the δ-distribution effectively has non-zero probability mass only at a single point, it is reasonable to expect (10.6) to hold also if the integral is carried out over a compact domain D, provided $\mathbf{x}^\star$ belongs to D. This is indeed the case, although the proof requires some work. For a quick guide to the many other interesting and useful properties of the δ-distribution that are beyond our current scopes, see [370].

with an experimental error σ_i (which likely conflates different sources of uncertainty of which we may know very little, if anything at all), such that the experimental population-level estimate of v_i is $\overline{v}_i \pm \sigma_i$. Let us assume that we know enough about the experiment to be able to define a flux polytope for the cell type ($\mathcal{P}$), and that all empirically measured (averages and errors) are in $\mathcal{P}$. Then, if we want to describe the population by a density in $\mathcal{P}$ that is uniform over the domain defined by experimental estimates, we can set

$$p(\mathbf{v}) = \prod_{i=1}^N \frac{\theta(\overline{v}_i + \sigma_i - v_i)\theta(v_i - \overline{v}_i + \sigma_i)}{2\sigma_i} \qquad (\mathbf{v}\in\mathcal{P}) \quad , \tag{10.8}$$

where $\theta(x)$ denotes the Heaviside (step) function.

Mathematical details 10.B: The step function

The Heaviside function is defined as

$$\theta(x) = \begin{cases} 1 & \text{for } x > 0 \\ 0 & \text{for } x < 0 \end{cases} \quad . \tag{10.9}$$

Exercise 10.2 Show that, for a real variable x, a continuous function f and upon integration over $\mathbb{R}$, $\frac{d}{dx}\theta(x) = \delta(x)$. **Hint.** Use the fact that $\frac{d}{dx}[\theta(x)f(x)] = \theta'(x)f(x) + \theta(x)f'(x)$.

- (Boltzmann distribution) Let $f(\mathbf{v})$ denote a generic function of the flux vector, such as $f(\mathbf{v}) = \sum_{i=1}^N c_i v_i$, with c_i prescribed constants. The Boltzmann distribution is defined as

$$p(\mathbf{v}) = \frac{1}{Z(\beta)}\,e^{\beta f(\mathbf{v})} \qquad (\mathbf{v}\in\mathcal{P}) \quad , \tag{10.10}$$

where β is a constant and Z is a factor ensuring normalisation (i.e. (10.1)), namely $Z(\beta) = \int_{\mathcal{P}} e^{\beta f(\mathbf{v})}d\mathbf{v}$. The behaviour of p is simple to grasp in three limits.

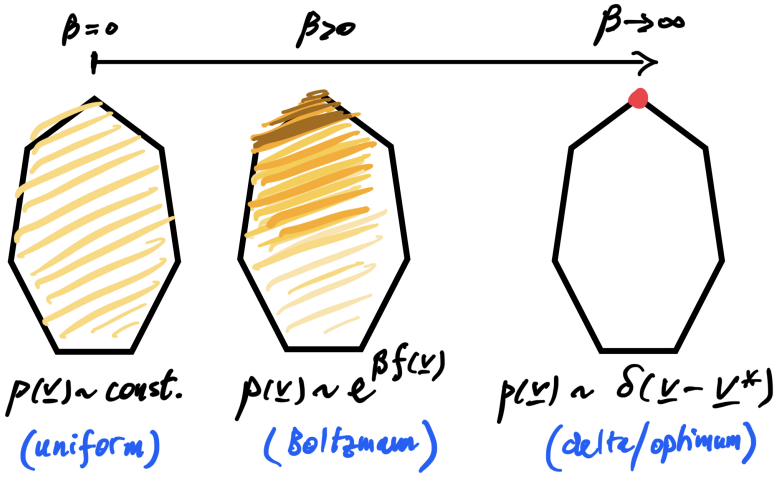

Figure 10.3: Boltzmann distribution on the flux polytope. The Boltzmann distribution, Eqn (10.10), morphs from a uniform probability density to a δ-distribution concentrated on the flux vector that maximises the function f as β varies from 0 to $+\infty$.

1. For $\beta \to 0$, (10.10) reduces to (10.7): in other words, p becomes uniform over $\mathcal{P}$ (and therefore insensitive to f).
2. For $\beta \to +\infty$, p effectively concentrates on the flux vector $\mathbf{v}^\star$ that maximises f (which for simplicity we assume to be unique). To see this at a heuristic level, it suffices to notice that, for any $\mathbf{v} \neq \mathbf{v}^\star$, the ratio

$$\frac{p(\mathbf{v}^\star)}{p(\mathbf{v})} = e^{\beta[f(\mathbf{v}^\star)-f(\mathbf{v})]} \tag{10.11}$$

   increases exponentially as β increases. Because densities are normalised, when this ratio becomes large, $p(\mathbf{v})$ must become very small. Hence, when integrated over $\mathcal{P}$, the larger is β, the closer to $\mathbf{v}^\star$ must flux vectors be in order to give a significant contribution to the integral. For $\beta \to +\infty$, the only relevant contribution comes from $\mathbf{v}^\star$, so that, effectively, $p(\mathbf{v}) \simeq \delta(\mathbf{v}-\mathbf{v}^\star)$. This conclusion can be reached more precisely using Laplace's method (a.k.a. saddle-point approximation) to evaluate integrals of the form $\int_{\mathbb{R}^n} e^{\beta g(\mathbf{x})} d\mathbf{x}$ in the limit $\beta \to \infty$ for fixed n (see e.g. [371], Ch. 27).
3. By a similar reasoning, for $\beta \to -\infty$ the only relevant contribution to integrals involving p comes from the (unique, by assumption) flux vector $\mathbf{v}_\star$ that *minimizes* f, so that, effectively, $p(\mathbf{v}) \simeq \delta(\mathbf{v}-\mathbf{v}_\star)$.

When β varies, things depend strongly on the form of f and can become rather complicated when f is non-linear, especially if it involves terms that involve the product of two or more fluxes ('high-order interactions'). However, in the simple case in which f is linear (as outlined above), then the probability density gradually morphs from a uniform distribution over $\mathcal{P}$ to a δ-distribution around the maximum of f as β increases from 0 to $+\infty$ as shown in Fig. 10.3 (and likewise when β decreases from 0 to $-\infty$). In this respect, the parameter β can be seen simply as a 'degree of optimisation': the closer a population is to optimising f, the higher the value of β. For reasons that will become clear in the next section, the Boltzmann distribution plays an especially important role in this chapter.

Exercise 10.3 Well- versus ill-defined flux spaces. Using the sampling methods introduced in the Chapter (*The space of metabolic flux distributions*) and a linear objective function of your choice, write a program that will sample a toy two-dimensional flux polytope according to (10.10), and check the outcome for a few values of β. Then try changing the shape of the polytope in different ways by changing the constraints. What features of the polytope can make sampling harder and/or less accurate (i.e. require a larger number of samples)? Can you work out a modification of the sampling algorithms that alleviates these problems?

- In Constrained Allocation FBA [343] (see the Chapter *Large cell models*), one considers an ensemble of growth-rate maximisation problems constructed by sampling (from a prescribed probability density) a family of random variables

representing the proteome fraction to be invested in each metabolic enzyme per unit flux of the corresponding reaction. The idea in CAFBA is that different sets of parameters effectively correspond to different cells, reflecting the cell-to-cell variability in e.g. transcription levels and protein abundances. The population-level behaviour is then obtained by averaging over different choices of these parameters (i.e. over a population of heterogeneous cells). An alternative interpretation is however possible, namely that different parameters reflect the different environmental conditions that a species can encounter over its life process history. By averaging over parameters one obtains a growth strategy that levels out this environmental variability. Such a strategy may be the one that cells prefer to implement e.g. when environmental fluctuations are fast (faster than regulatory timescales). In either case, in CAFBA, randomness in a family of parameters related to the optimisation problem induces randomness in the solutions, and therefore a probability density over the feasible space. This probability is unfortunately hard to write down explicitly in the case of CAFBA due to the complexity of the optimization problem. Its marginal distributions are however easy to calculate numerically. Two of them, specifically for the single-cell growth rate and acetate excretion fluxes, are shown in Fig. 2 in [343].

We could provide more examples but the key message of this section should already be visible: probability densities on the flux polytope are useful (a) when one wants to explicitly represent how uncertainties, experimental knowledge (with errors), or variability in parameters impact our knowledge of what part of the flux space $\mathcal{P}$ is occupied by the metabolic states that occur in a true microbial population; and (b) when one is interested in representing an *optimal* (in some sense) population in a way that explicitly accounts for heterogeneity. If one has data (with errors), a probability density can provide a representation of the data, as in (10.8). It can likewise describe the solution to a population-level optimisation problem, and therefore a purely theoretical prediction, as in (10.3). Or the solution to an optimisation problem with uncertainty, i.e., partial knowledge or variability in some of the parameters, in which case it represents an 'informed' theoretical prediction (as in the CAFBA example, where the 'information' injected into the problem comes from the probability density from which parameters are sampled). Or it can simply be a tool to interpolate between extreme cases when we are unsure about how well a certain function is being optimised (as in (10.10)). Notice how, in our examples, different motivations activate different theoretical routes, all of which lead to working with probability densities that have *a priori* different origins and meanings even though they can be formally the same.

The two broad motivations for working with probability densities on $\mathcal{P}$ outlined above [i.e. (a) representing uncertainty and (b) representing optimal heterogeneous populations], pose fundamentally different modelling challenges. In the first case, the key question is one of model selection: given some empirical knowledge, what is the probability density on $\mathcal{P}$ that best represents our residual uncertainty? For instance: how good of a choice for p is (10.8) given the data we had? Are there criteria that can guide our choice of a probability density? We will briefly consider these issues in the upcoming Sec. 10.4. When attempting to model optimal heterogeneous populations at the theoretical level, instead, one basically has to generalise the problem tackled by CBMs models like FBA to the case in which an optimal probability density is searched for instead of an optimal flux configuration. We will see how this can be done in Sec. 10.5.

10.4 Representing heterogeneity and uncertainty

10.4.1 ML, MAP and Bayesian inference

We have seen that probability densities on $\mathcal{P}$ can represent, under certain assumptions, populations of microbes whose metabolism can be described by the same flux polytope, and that different probability densities can be surmised to model the distribution of $\mathbf{v} \in \mathcal{P}$ when some external information (e.g. experimental data) is available. Here, we will address the following question: how can one choose the p that best represents our knowledge about the metabolic state of a population in presence of these external data?

To summarise the huge and highly involved set of problems behind the above (very general) question [371] in a way that is useful for the purposes of this chapter, we can start by assuming we have *a priori* chosen a form of p that depends on certain free parameters and ask how to tailor parameters so that p 'optimally' matches the empirical evidence. To

be concrete, let us denote by $\boldsymbol{\psi}$ the vector of parameters of p, and by $\mathbf{W} = \{\mathbf{w}^1, \mathbf{w}^2, ..., \mathbf{w}^R\}$ a set of R experimental samples of $\mathbf{v}$. Each measurement, $\mathbf{w}$, is a vector of metabolic fluxes that ideally should include all the reactions of a metabolic network. In practice, a $\mathbf{w}$ typically spans only a subset of all the reactions of the metabolic network, those that are amenable to ^{13}C labeling (TCA, glycolysis, and pentose phosphate pathways) or that correspond to exchange fluxes that can be reliably measured (glucose and oxygen consumption, or lactate and ethanol, to name a few). According to Bayes' rule (we assume all variables to be continuous), the quantities

- $p(\boldsymbol{\psi}|\mathbf{W})$: the conditional probability density of the parameters given the observations (a.k.a. the *posterior*);
- $p(\mathbf{W}|\boldsymbol{\psi})$: the conditional probability density of the observations given the parameters (a.k.a. the *likelihood*);
- $p(\boldsymbol{\psi})$: the prior probability density of parameters (a.k.a. the *prior*);
- $p(\mathbf{W})$: the (marginal) probability density of observations (a.k.a. the *evidence*)

are related by the formula

$$p(\boldsymbol{\psi}|\mathbf{W}) = \frac{p(\mathbf{W}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{p(\mathbf{W})} \quad . \tag{10.12}$$

Ideally, what one would like to know in order to 'optimally' set the parameters of p is how likely a parameter set is given the data, i.e. the full posterior $p(\boldsymbol{\psi}|\mathbf{W})$, as it allows to quantify our uncertainty on the model itself. One may however also consider different (less ambitious) ways to choose parameters. The three best known methods are the following:

- Maximum Likelihood (ML) inference aims at finding the parameter vector that maximises the likelihood:

$$\boldsymbol{\psi}_{\mathrm{ML}} = \arg\ \max_{\boldsymbol{\psi}} p(\mathbf{W}|\boldsymbol{\psi}) \quad . \tag{10.13}$$

  In standard cases, this produces a single 'optimal' vector $\boldsymbol{\psi}$ (hence it is called a 'point estimator'), resulting in a p that models -in a context-specific manner- the metabolic heterogeneity within the cellular population.
- Maximum a Posteriori (MAP) inference aims instead at finding the parameter vector that maximises the posterior:

$$\boldsymbol{\psi}_{\mathrm{MAP}} = \arg\ \max_{\boldsymbol{\psi}} p(\boldsymbol{\psi}|\mathbf{W}) \equiv \arg\ \max_{\boldsymbol{\psi}} p(\mathbf{W}|\boldsymbol{\psi})p(\boldsymbol{\psi}) \quad , \tag{10.14}$$

  where the last equality follows from the fact that $p(\mathbf{W})$ does not depend on $\boldsymbol{\psi}$. Again, the MAP estimator is a point estimator.
- Bayesian inference aims finally at computing the full posterior distribution $p(\boldsymbol{\psi}|\mathbf{W})$. It is therefore a 'distribution estimator' rather than a point estimator.

The following exercise should clarify the way in which point estimators differ from (and are less informative than) distribution estimators in practice.

Exercise 10.4 MAP inference versus Bayesian inference. Consider a Bernoulli random variable with parameter ψ, i.e. such that the probability of having k successes in n trials given ψ is

$$p(k|\psi) = \binom{n}{k}\psi^k(1-\psi)^{n-k} \quad , \tag{10.15}$$

and assume that the prior for ψ is a β-distribution with parameters a and b, i.e.

$$p(\psi) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\psi^{a-1}(1-\psi)^{b-1} \quad . \tag{10.16}$$

Calculate the full posterior $p(\psi|k)$ and the MAP estimator for ψ as a function of k, n, a and b. Then assume $a = b = 2$ and compare the following situations: (i) a Bernoulli process that returned 2 successes in 3 trials; (ii) a Bernoulli process that returned 20 successes in 33 trials. Show that the MAP estimator for ψ is 60% for both (i) and (ii) (so the two processes are indistinguishable to MAP), while the posterior is different. Knowing the posterior, which process would

you pick if you were asked to point to the one that is more likely to have $\psi = 0.6$?

Mathematical details 10.C: Inference in metabolic network modeling

ML is the most commonly used point estimation method. As said above, the estimated parameters, $\hat{\boldsymbol{\psi}}$, are computed as the argument that maximizes the likelihood of the observed data, i.e.

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} p(\mathbf{W}|\boldsymbol{\psi}) = \arg\max_{\boldsymbol{\psi}} \prod_{i=1}^{R} p(\mathbf{w}^{(i)}|\boldsymbol{\psi}) \quad . \tag{10.17}$$

ML takes a familiar form if one follows, for instance, Theorell *et al.* [372] in modelling $\mathbf{v}$ according to a multivariate normal distribution:

$$\mathbf{v} \sim N(\mathbf{v}|\boldsymbol{\psi}) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} e^{\left(-\frac{1}{2}(\bar{\mathbf{v}}-\mathbf{v})^T \Sigma^{-1} (\bar{\mathbf{v}}-\mathbf{v})\right)} \tag{10.18}$$

The parameters encompass the mean flux vector, $\bar{\mathbf{v}}$, and the covariance matrix, $\boldsymbol{\Sigma}$. This is, $\boldsymbol{\psi} = [\bar{\mathbf{v}}, \boldsymbol{\Sigma}]$. Accordingly, each of the i terms of the rightmost term of Eq. 10.17 can be expressed as

$$p(\mathbf{w}^{(i)}|\boldsymbol{\psi}) = N(\mathbf{w}^{(i)}|\boldsymbol{\psi}) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} e^{\left(-\frac{1}{2}(\bar{\mathbf{v}}-\mathbf{w}^{(i)})^T \Sigma^{-1} (\bar{\mathbf{v}}-\mathbf{w}^{(i)})\right)} \quad , \tag{10.19}$$

so that their product can be transformed into a summation by applying the log function (without changing the final outcome):

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} \sum_{i=1}^{R} \log[p(\mathbf{w}^{(i)}|\boldsymbol{\psi})] \tag{10.20}$$

$$= \arg\max_{\bar{\mathbf{v}},\boldsymbol{\Sigma}} \sum_{i=1}^{R} \left[-\frac{1}{2}(\bar{\mathbf{v}} - \mathbf{w}^{(i)})^T \Sigma^{-1} (\bar{\mathbf{v}} - \mathbf{w}^{(i)}) - \log\left(\sqrt{(2\pi)^N |\Sigma|}\right) \right] \tag{10.21}$$

The second term of the summation of Eq. (10.21) is a constant and can be omitted from the maximisation. We are therefore left with

$$\hat{\boldsymbol{\psi}} = [\hat{\bar{\mathbf{v}}}, \hat{\boldsymbol{\Sigma}}] = \operatorname*{argmin}_{\bar{\mathbf{v}},\boldsymbol{\Sigma}} \sum_{i=1}^{R} (\bar{\mathbf{v}} - \mathbf{w}^{(i)})^T \Sigma^{-1} (\bar{\mathbf{v}} - \mathbf{w}^{(i)}) \quad . \tag{10.22}$$

Eq. (10.22) corresponds to the well-known weighted least squares estimator. The resulting $\hat{\bar{\mathbf{v}}}$ corresponds to the mode of p, which can be interpreted as the flux vector with the highest frequency within $\mathcal{P}$. With $\hat{\bar{\mathbf{v}}}$ and $\bar{\boldsymbol{\Sigma}}$, the frequency of any flux configuration can be computed from $N(\mathbf{v}|\hat{\bar{\mathbf{v}}}, \hat{\boldsymbol{\Sigma}})$.

Standard techniques, such as confidence intervals, can be applied to assess the precision of $\hat{\boldsymbol{\psi}}$. The larger the number of samples, R, the smaller the uncertainty in $\hat{\boldsymbol{\psi}}$. Unfortunately, measurements of flux vectors typically include only a handful of points [373], which may lead to $\hat{\boldsymbol{\psi}}$ over-fitted to the sample set. One way to overcome limited sample sizes is to regularise the estimation procedure by incorporating *prior* information on $\boldsymbol{\psi}$ via the MAP estimation method (10.14). The evidence $p(\boldsymbol{\psi})$ in MAP can be used to encode the distribution of $\mathbf{v}$ values observed in previous experiments or formulated as a plausible non-informative probability distribution. For example, Heinonen *et al.* [374] formulated $p(\boldsymbol{\psi})$ as a multivariate normal distribution with mean values equal to zero, and variances for each flux adjusted to prevent fluxes extending beyond their lower and upper bounds defined in $\mathcal{P}$. MAP estimation can be considered as an ML estimation whose objective function has been augmented by the prior distribution of $p(\boldsymbol{\psi})$. In this sense, MAP estimation is a 'regularised' ML estimation, which helps prevent overfitting.

MAP estimation however does not exploit the capacity of Bayes' theorem to explore the full set of values that the parameters can achieve. By producing a distribution estimation of the parameters, Bayesian inference allows quantifying the parameters' variability. Compared to point estimation methods, though, Bayesian inference is

computationally expensive as it requires to asses how different values of $p(\boldsymbol{\psi})$ affect $p(\mathbf{W}|\boldsymbol{\psi})$. Fortunately, some families of p are susceptible to methods such as Gibbs sampling or Markov Chain Monte Carlo that offer an efficient way to compute the posterior numerically [375]. This is the case, for instance, for the truncated multivariate normal distributions that Heinonen *et al.* [374] used for the likelihood and prior functions appearing in (10.12). The posterior can then be used to derive statistical features of quantities that depend on $\boldsymbol{\psi}$, e.g. metabolic fluxes.

In practice, most parameters underlying the mechanisms that govern cellular metabolism -e.g., enzymes' allosteric regulation or the local conditions within cells' organelles- remain unknown. Various hypotheses can be advanced to close this knowledge gap. Alas, it is not uncommon to have conflicting scenarios. For instance, to explain overflow metabolism in *S. cerevisiae* and *E. coli* [376, 377, 378], numerous plausible explanations have been pushed forward, including ATP savings for the production of non-oxidative enzymes (which by being smaller, compared to their oxidative counterparts, require less ATP in their synthesis) [379, 380], limited uptake rates capacity [381], and an upper limit on the dissipation of Gibbs energy [382]. (See [383] for an excellent review of optimisation-based explanations.) Because each mechanism can be encoded through a different prior, it is clear that the choice of the prior is a delicate matter in Bayesian inference. Generally speaking, the choice of the prior becomes less and less problematic the more data we have, i.e. the better sampling we have of the state space of the system. However, if data is scant, the prior will leave an important imprint on the resulting posterior. In these cases, a careful selection of the prior is paramount. Among the methods most commonly employed are (a) the construction of empirical priors (namely priors that encode previous knowledge about parameters), (b) the use of so-called "non-informative priors" (i.e. priors that reflect 'vague knowledge' about parameters, like the fact that a certain parameter is non-negative) [384], and (c) the selection of priors based on the Maximum Entropy principle (see below) [385, 386].

10.4.2 MaxEnt inference

According to the principle of Maximum Entropy (MaxEnt) [387, 388], among all probability densities that are consistent with given prior knowledge or data, the one having the largest value of the entropy

$$H[p] = -\int_{\mathcal{P}} p(\mathbf{v}) \ln p(\mathbf{v}) d\mathbf{v} \tag{10.23}$$

is the one that best represents our knowledge about the system. A classical intuitive justification of the MaxEnt principle is most easily given for discrete variables [389].

Mathematical details 10.D: Entropy

Consider N cells, each of which can be found in any of K states (what precisely defines a state is immaterial for this reasoning). Let an assignment $\mathbf{n} = \{n(i)\}$ be given, such that $n(i)$ denotes the number of cells in state i (with $\sum_{i=1}^{K} n(i) = N$). Because we can always exchange the states of two cells without changing $\mathbf{n}$, there are multiple 'microscopic' ways to realise an assignment $\mathbf{n}$. Combinatorics tells us that the number of different microscopic realisations of an assignment $\mathbf{n}$ is given by

$$\mathcal{N}(\mathbf{n}) = \frac{N!}{\prod_{i=1}^{K} n(i)!} \quad . \tag{10.24}$$

If all $n(i)$'s are large enough, we can use Stirling's approximation ($n! \simeq (n/e)^n$) to see that

$$\mathcal{N}(\mathbf{n}) \simeq e^{NH(\mathbf{n})} \quad , \quad H(\mathbf{n}) = -\sum_{i=1}^{K} \frac{n(i)}{N} \ln \frac{n(i)}{N} \equiv -\sum_{i=1}^{K} p(i) \ln p(i) \equiv H(\mathbf{p}) \quad , \tag{10.25}$$

where $p(i)$ denotes the fraction of cells in state i (or, equivalently for us, the probability to find a cell in state i). H is the *entropy* of the assignment $\mathbf{n}$, and is in essence a measure of the microscopic degeneracy that underlies a

macroscopic arrangement. The distribution $\mathbf{p} = \{p(i)\}$ carrying the largest entropy subject to certain constraints is therefore the one having the largest underlying microscopic degeneracy given those constraints. So, if one were to randomly pick a microscopic state given those constraints, the most likely macroscopic state would be the maximum entropy distribution. In other terms, the MaxEnt distribution is the least biased distribution compatible with the constraints, as any other distribution satisfying the same constraints would correspond to a smaller underlying degeneracy, thereby neglecting some feasible (i.e. constraint-satisfying) microscopic configurations. In this respect, a MaxEnt distribution requires the least information besides prior knowledge (i.e. constraints). (A more detailed justification for using the MaxEnt principle as an inference tool is given e.g. in [389].)

If for instance cells are assigned to states in a completely random way, the MaxEnt distribution is the solution of

$$\max_{\mathbf{p}} -\sum_{i=1}^{K} p(i) \ln p(i) \quad \text{subject to} \quad \sum_{i=1}^{K} p(i) = 1 \quad , \tag{10.26}$$

which can be found via the method of Lagrange multipliers.

Exercise 10.5 Show that the solution of the above maximisation problem is the uniform distribution $p(i) = 1/K$ for all i.

If other constraints are imposed, though, the MaxEnt distribution will clearly change.

Exercise 10.6 MaxEnt distribution in different cases. Assume that a certain real variable x takes values $x(i)$ in the K states (one can for instance think of $x(i)$ as the growth rate of cells in state i). Show that the MaxEnt distributions for constraints imposed on (i) normalisation of the distribution, (ii) normalisation and mean value of x, (iii) normalisation, mean value of x and second moment of x, and (iv) normalisation and mean of the logarithm of x, are, respectively, uniform, exponential, Gaussian, and power-law.

For our purposes, the continuous case with entropy given by (10.23) can be seen as a straightforward generalisation of the discrete one.

To get some grasp of the scenario that the MaxEnt rule provides within CBMs, let us work out one especially noteworthy case, namely the MaxEnt probability density of flux configurations with a given mean value of a generic function f of the fluxes. This probability density is the solution of

$$\max_{p(\mathbf{v})} -\int_{\mathcal{P}} p(\mathbf{v}) \ln p(\mathbf{v}) d\mathbf{v} \quad \text{subject to} \quad \int_{\mathcal{P}} p(\mathbf{v}) d\mathbf{v} = 1 \quad \text{and} \int_{\mathcal{P}} f(\mathbf{v}) p(\mathbf{v}) d\mathbf{v} = \overline{f} \quad . \tag{10.27}$$

To find it, we construct the functional

$$\mathcal{L}[p] = H[p] + \alpha \left[\int_{\mathcal{P}} p(\mathbf{v}) d\mathbf{v} - 1 \right] + \beta \left[\int_{\mathcal{P}} f(\mathbf{v}) p(\mathbf{v}) d\mathbf{v} - \overline{f} \right] \quad , \tag{10.28}$$

where α and β are Lagrange multipliers for the normalisation and the mean-value-of-f constraints, respectively. Variation of $\mathcal{L}$ with respect to p yields the maximum condition

$$1 - \ln p(\mathbf{v}) + \alpha + \beta f(\mathbf{v}) = 0 \quad . \tag{10.29}$$

Solving for p results in

$$p(\mathbf{v}) = \frac{e^{\beta f(\mathbf{v})}}{e^{-(\alpha+1)}} \quad . \tag{10.30}$$

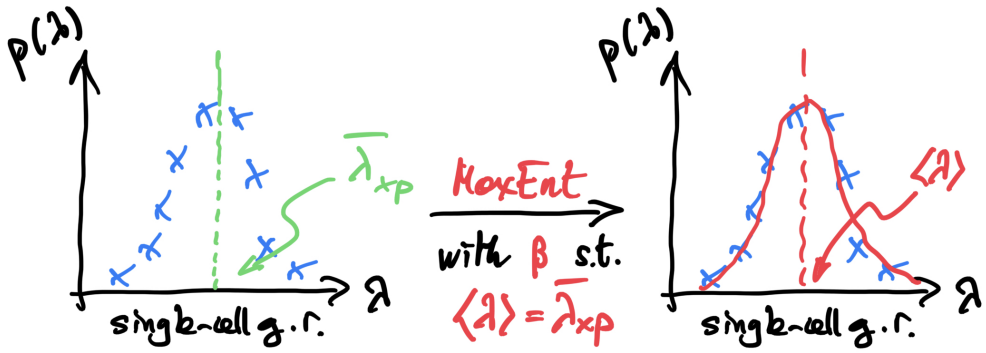

Figure 10.4: MaxEnt modelling of single-cell growth rate distributions. Empirical distributions are reproduced by a MaxEnt assumption where the mean growth rate is constrained, leading to a Boltzmann distribution over the flux polytope (Eq. (10.34)).

The normalisation condition however implies

$$\int_{\mathcal{P}} e^{\beta f(\mathbf{v})} d\mathbf{v} = e^{-(\alpha+1)} \equiv Z(\beta) \quad , \tag{10.31}$$

so that, finally,

$$p(\mathbf{v}) = \frac{1}{Z(\beta)} e^{\beta f(\mathbf{v})} \qquad (\mathbf{v} \in \mathcal{P}) \quad . \tag{10.32}$$

The value of β must be determined from second constraint, namely

$$\frac{1}{Z(\beta)} \int_{\mathcal{P}} f(\mathbf{v}) e^{\beta f(\mathbf{v})} d\mathbf{v} = \overline{f} \quad . \tag{10.33}$$

Notice that the result is nothing but Boltzmann's distribution (10.10). We have therefore found that (10.10) is the MaxEnt distribution for a given mean value of the function f. This means that if we have a dataset returning the empirical mean value of an observable f over a population of cells, our knowledge is best represented by assuming that $p(\mathbf{v})$ is of the form (10.10), with β ensuring the matching of empirical and theoretical means.

This suggests a possible way to represent single-cell growth-rate distributions [390], such as the *E. coli* populations growing in rich media studied e.g. in [391, 346, 347] (see Figure 10.4). Let us assume that all cells in the population can be described by the same flux polytope $\mathcal{P}$ and let $\lambda(\mathbf{v})$ denote the growth rate associated to flux configuration $\mathbf{v}$. We can ask the following question: what is the $p(\mathbf{v})$ on $\mathcal{P}$ that best represents our knowledge that the mean growth rate of cells is $\overline{\lambda}$ (empirical)? The answer is

$$p(\mathbf{v}) = \frac{1}{Z(\beta)} e^{\beta \lambda(\mathbf{v})} \qquad (\mathbf{v} \in \mathcal{P}) \quad , \tag{10.34}$$

where $Z(\beta) = \int_{\mathcal{P}} e^{\beta \lambda(\mathbf{v})} d\mathbf{v}$, and where β is set so that the empirical mean growth rate ($\overline{\lambda}$) matches the theoretical mean, i.e.

$$\frac{1}{Z(\beta)} \int_{\mathcal{P}} \lambda(\mathbf{v}) e^{\beta \lambda(\mathbf{v})} \, d\mathbf{v} = \overline{\lambda} \quad . \tag{10.35}$$

We can therefore solve the above equation (numerically) and analyse the resulting distribution. One sees from (10.34) that β has a 'natural' unit given by $\lambda_{\max}$, the maximum growth rate achievable in $\mathcal{P}$ (which is easily computed by LP). In the populations analysed in [390], the value of β that ensures the matching condition ranges from $190/\lambda_{\max}$ to $300/\lambda_{\max}$, suggesting that indeed the degree of optimisation of λ is significant. The most remarkable result, however is that the marginal distribution of the growth rate computed from (10.34), namely

$$p(\lambda) = \int_{\mathcal{P}} \delta(\lambda - \lambda(\mathbf{v})) p(\mathbf{v}) d\mathbf{v} \quad , \tag{10.36}$$

matches the overall empirical growth-rate distributions. In other words, if one adjusts the parameter of (10.34) so that the theoretical mean growth rate and the empirical one coincide, then (10.36) reproduces the entire empirical growth-rate distribution. This observation confirms the empirical fact that the variance of single-cell growth-rate distributions is a function of the mean, such that, if growth rates are re-scaled by the mean, distributions roughly collapse on 'universal curves' that turn out to only depend on the degree of optimisation (i.e. on β). In addition, the analysis of [392] has shown that predictions for individual fluxes obtained from (10.34) (i.e. mean values plus standard deviations) provide a better fit to experimentally measured fluxes than growth-rate maximising fluxes obtained from FBA. This is especially important as it suggests that, despite the relatively high degree of optimisation, the cell-to-cell variability underlied by (10.34) is biologically relevant.

In the following section we will use this observation as a springboard to study optimal heterogeneous populations.

10.5 Representing optimal populations

Let us start from a rather abstract question. Suppose that an organism is actually maximising a certain function F, unknown to us, which depends on metabolic fluxes $\mathbf{v}$ as well as on a set of other variables $\mathbf{w}$ that are not part of metabolism: $F \equiv F(\mathbf{v}, \mathbf{w})$. We shall denote by $(\mathbf{v}^\star, \mathbf{w}^\star)$ the (supposedly unique) configuration of variables where F attains its maximum. Let's furthermore say that we have a guess for what the organism's objective function might be, and that this guess is only a function of metabolic fluxes, which we denote by $f \equiv f(\mathbf{v})$. If we trust our guess, and if f is maximised by the (supposedly unique) flux vector $\widehat{\mathbf{v}}$, our prediction for the fluxes would be $\widehat{\mathbf{v}}$. Question: what is the probability that $\widehat{\mathbf{v}}$ is the true optimum, i.e. that $\widehat{\mathbf{v}} = \mathbf{v}^\star$? Note that $f(\mathbf{v}^\star) \equiv f^\star < \widehat{f} \equiv f(\widehat{\mathbf{v}})$ (i.e. at the 'true' optimum the value of f is bound to be smaller than the maximum value of f).

The answer goes like this: according to the MaxEnt principle, the probability density $p(\mathbf{v})$ for any flux configuration $\mathbf{v}$ to be the true state of the system (i.e. the true optimum) should be undetermined other than by our knowledge that the real optimum has some value of f below $\widehat{f}$. What is the correct constraint to enforce (besides normalisation) if we are to look for such a $p(\mathbf{v})$? We could impose that allowed configurations strictly have some fixed value of $f < \widehat{f}$. This choice would lead to a uniform density over all states with a given value of f. In this way, however, we are assuming that states with a different value of f are inaccessible, a rather strong hypothesis. However, if we impose that only the mean value of f is constrained, MaxEnt will return a probability density with the exact same mean value as the uniform density just described (by construction) but a much larger entropy, just because –intuitively– it will assign a non-zero probability to many more states. Hence, as long as we have no other information, the best prediction we can make for $p(\mathbf{v})$ is given by the probability density that maximises the entropy $H[p]$ subject to the constraint $\langle f \rangle \equiv \int_{\mathcal{P}} p(\mathbf{v}) f(\mathbf{v}) d\mathbf{v} = f^\star$, i.e. by the solution of

$$\max_p \; -\int_{\mathcal{P}} p(\mathbf{v}) \ln p(\mathbf{v}) d\mathbf{v} \quad \text{subject to} \quad \int_{\mathcal{P}} p(\mathbf{v}) d\mathbf{v} = 1 \quad \text{and} \quad \int_{\mathcal{P}} p(\mathbf{v}) f(\mathbf{v}) d\mathbf{v} = f^\star \quad . \tag{10.37}$$

We now know the result to be given by (10.10), i.e.

$$p(\mathbf{v}) = \frac{1}{Z(\beta)} \, e^{\beta f(\mathbf{v})} \qquad (\mathbf{v} \in \mathcal{P}) \quad , \tag{10.38}$$

where β is the Lagrange multiplier enforcing the constraint $\langle f \rangle = f^\star$. What this means in practice is this: if one is modelling a microbe's metabolism and is unsure about the objective function but has a guess (f), information theory suggests that the best one can do is to assume that metabolic flux configurations are selected according to (10.38). Ideally, the value of β for which one obtains the best agreement between predictions based on sampling (10.38) and experiments is the 'degree' to which the system optimises f. If f is the true objective function, then the agreement between theory and experiments will get better and better as β increases. It is important to keep in mind that (i) while we have assumed that the organism is actually maximising something, we didn't really use the fact that F is maximised at $(\mathbf{v}^\star, \mathbf{w}^\star)$ (only that the true state of the system has some value of f below $\widehat{f}$); (ii) this is a totally ideal situation

(for instance, experimental data have errors, so whether comparisons between theory and experiments are informative doesn't only depend on the theory but also on the quality of the data).

The fact that (10.38) is 'optimal' in a rather fundamental sense (*a priori* different from the sense in which f-maximising populations are optimal) encourages to view distributions described by (10.34) through a different lens. When we maximise the entropy at fixed mean growth rate, in practice, we are looking for the 'broadest' probability density (i.e. the most variable population) on $\mathcal{P}$ that is compatible with the given mean. In other terms, we are saying that, given a mean growth rate, the optimal population is the one that has the largest possible variability. To quantify variability in a more readily understandable way, it is convenient to transform it into a measure of the amount of information encoded in p. One can reason as follows: if no prior information is available about the population, uncertainty is maximal and all flux vectors in $\mathcal{P}$ must be considered to be equally likely. This means that, for such a population, the probability density over $\mathcal{P}$ is uniform (see (10.7)). We shall denote the entropy of the uniform distribution over $\mathcal{P}$ by $H(0)$. When we inject information into the problem (e.g. the fact that the population has a certain mean growth rate), then the probability density is no longer uniform but given, say, by (10.34). The uncertainty is therefore reduced by $H(0) - H(\beta)$, where $H(\beta)$ is the entropy of (10.34). (Clearly, $H(0)$ is just the entropy of (10.34) for $\beta = 0$.) The quantity

$$I = \frac{H(0) - H(\beta)}{\ln 2} \tag{10.39}$$

denotes the amount of information (in bits, hence the factor $\ln 2$) injected by a non-zero value of β. Re-formulating our population-level optimisation, we can say that, for any fixed mean growth rate $\langle\lambda\rangle$, the optimal population is the one carrying the smallest value of I. A short calculation shows that $\langle\lambda\rangle$ and I are related by

$$\beta\langle\lambda\rangle = I \ln 2 + \int_0^\beta \langle\lambda\rangle d\beta' \quad , \tag{10.40}$$

where it should be noted that $\langle\lambda\rangle$ is an increasing function of β (as β increases, the density gets more and more concentrated around the growth-rate maximising flux vector, thereby leading to an increase of $\langle\lambda\rangle$).

Exercise 10.7 Retrieve formula (10.40).

The curve $\langle\lambda\rangle$ versus I described by (10.40) can therefore be computed numerically for any metabolic network reconstruction (as the only ingredients required are encoded in the flux polytope $\mathcal{P}$) [390]. The resulting line (see Figure 10.5) separates the $(\langle\lambda\rangle, I)$ plane in a viable (achievable) region and a forbidden region where the mean growth rates are too large for the amount of information encoded in the population. This 'phase diagram' yields, first and foremost, a general prediction linking the mean growth rate (fitness) of a microbial population to its metabolic heterogeneity: all populations must have fitness-heterogeneity values in the viable region. Recent work relying on an advanced statistical inference framework has shown that actual microbial populations indeed lie in the viable part of the plane [393]. In addition, it provides a quantitative definition of an optimal population that accounts for variability: optimal populations have fitness-heterogeneity pairs that lie on the boundary between the viable and the forbidden region. In this respect, results from [390, 392, 393] can be summarised by saying that heterogeneous, faster-growing *E. coli* populations (mean growth rate larger than roughly 1/h, richer growth media) are very close to optimality, while slower-growing ones tend to have mean growth rates and information contents that get more and more sub-optimal the less rich is the growth medium. (Of course, this notion of optimality refers to the growth rate and information content as the key parameters to evaluate a population's performance. It may well be, and this is an issue definitely worth exploring, that slower-growing population are optimal with respect to some other parameter(s).) At any rate, the above definition of optimality coincides with the standard one (growth-rate maximisation) for $\beta \to +\infty$, in which case variability goes strictly speaking to zero as all cells collapse on the same flux configuration. And we now understand how it generalises it: by stressing the way in which heterogeneous populations can be optimal despite growing at sub-maximal rates.

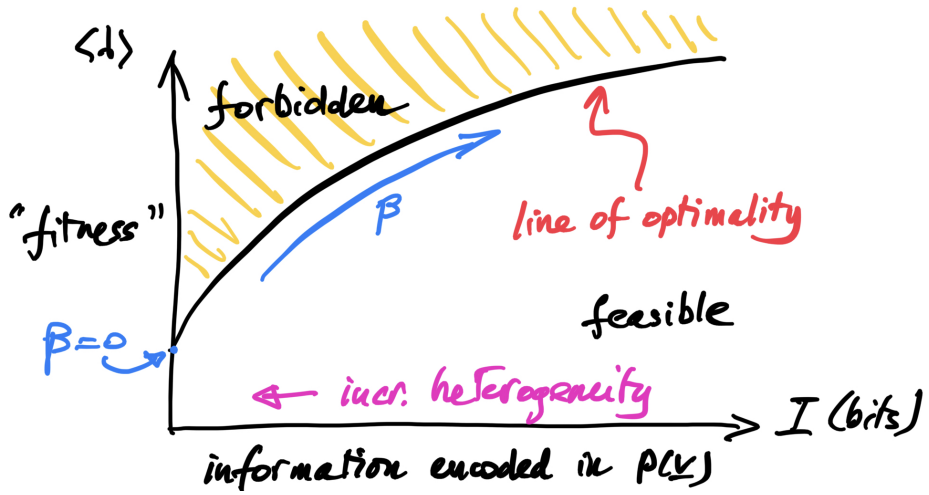

Figure 10.5: Fitness-information bound (general form). The black line encodes the maximum mean growth rate achievable for any given value of the information content I (Eq. (10.39)) of a metabolic flux distribution (or the mininum value of I required to achieve any given mean growth rate).

For later convenience, note that, because the entropy is a convex functional, the solution to the MaxEnt problem is the same as the solution to

$$\max_p \int_{\mathcal{P}} p(\mathbf{v})\lambda(\mathbf{v})d\mathbf{v} \quad \text{subject to} \quad \int_{\mathcal{P}} p(\mathbf{v})d\mathbf{v} = 1 \quad \text{and} \quad -\int_{\mathcal{P}} p(\mathbf{v})\ln p(\mathbf{v})d\mathbf{v} = H^\star \quad .$$

The above problem has perhaps a more direct interpretation: the optimal population is the one that has the largest mean growth rate at fixed variability (entropy) or, equivalently, at fixed information content.

Before moving on, we notice that, in the above setting, optimality of heterogeneous populations has a rather simple mechanistic interpretation in terms of how populations 'occupy' the flux polytope. If one considers the uniform distribution on $\mathcal{P}$, Eq. (10.7), and calculates the marginal distribution for the growth rate (i.e. (10.36)), one finds that the growth-rate landscape in which populations grow is extremely skewed towards slow growth rates: the overwhelming majority of metabolic flux configurations corresponds to slow-growing cells, i.e. with growth rates roughly two orders of magnitude below $\lambda_{\max}$. This implies that, whatever flux vector we are in, a small random change to it is overwhelmingly more likely to decrease our growth rate than increase it. In this respect, slow states have an 'entropic' advantage over fast states. On the other hand, by definition, fast-growing flux configurations replicate faster than slow-growing ones, and therefore have a replicative advantage. It is therefore tempting to interpret the probability density (10.34) as resulting from the balance between these two tendencies. One can for instance imagine that a microbial population grows and evolves in time in $\mathcal{P}$ due to (i) replication events, and (ii) small random changes of the flux vector (due e.g. to gene expression noise). Ref. [390] has indeed shown that such a population naturally evolves toward a distribution very close to (10.34), where the role of β is played by the inverse rate of diffusion of the population in $\mathcal{P}$, that is, by the inverse of the rate at which small random changes occur: fast rate implies small β, and vice versa. (As the mathematical analysis of this scenario requires the toolbox of non-linear Fokker-Planck equations, it is beyond the scopes of this chapter.)

The above theory can be extended in various directions. We shall limit ourselves to one example here, namely that of optimal populations in fluctuating environments [394]. The basic assumption we make is that the growth rate λ is a function of both the flux vector $\mathbf{v}$ and of a single (for simplicity) exogenous variable $s \geq 0$ representing the stress level to which the population is subject: $\lambda \equiv \lambda(\mathbf{v}, s)$. We furthermore assume that s is a random variable with probability density $P(s)$. For any value of s, λ will be maximised by a certain flux vector $\mathbf{v}^\star \equiv \mathbf{v}^\star(s)$. If fluctuations of s are sufficiently slow, then cells may be able to perfectly adapt their metabolic response to every value of s they encounter. But this is unlikely to be possible in rapidly fluctuating environments. In the latter case, it is instead reasonable to assume that cells will try to maximise their average growth rate, where the average is taken over the distribution of s. And the relevant distribution to consider is now the probability to observe a certain flux configuration $\mathbf{v}$ given that the

state of the environment is s: $p(\mathbf{v}|s)$. So the objective function (to be maximised over $p(\mathbf{v}|s)$) is just

$$\langle\lambda\rangle = \int ds\, P(s) \int_{\mathcal{P}} p(\mathbf{v}|s)\lambda(\mathbf{v},s)d\mathbf{v} \quad . \tag{10.41}$$

Now we must specify the constraints. One is simple and concerns normalisation: $\int_{\mathcal{P}} p(\mathbf{v}|s)d\mathbf{v}$ should be equal to one for all s. To introduce the second one, we note that, because one expects $\mathbf{v}$ to encode information about the environment, it is convenient to constrain the mutual information between $\mathbf{v}$ and s, i.e.

$$I(\mathbf{v};s) = \int ds\, P(s) \int_{\mathcal{P}} p(\mathbf{v}|s) \log_2 \frac{p(\mathbf{v},s)}{p(\mathbf{v})P(s)} d\mathbf{v} \quad , \tag{10.42}$$

where $p(\mathbf{v},s) = P(s)p(\mathbf{v}|s)$ is the joint distribution of $\mathbf{v}$ and s, while $p(\mathbf{v}) = \int dsP(s)p(\mathbf{v}|s)$. Clearly, $I = 0$ if $p(\mathbf{v},s)$ factorises over $\mathbf{v}$ and s and it gets larger and larger as $\mathbf{v}$ and s become more and more correlated. Putting these pieces together, we can write the cell's optimisation problem as

$$\max_{p(\mathbf{v}|s)} \int ds\, P(s) \int_{\mathcal{P}} p(\mathbf{v}|s)\lambda(\mathbf{v},s)d\mathbf{v} \quad \text{subject to} \quad \int_{\mathcal{P}} p(\mathbf{v}|s)d\mathbf{v} = 1 \quad (\forall s)$$
$$\text{and} \quad \int ds\, P(s) \int_{\mathcal{P}} p(\mathbf{v}|s) \log_2 \frac{p(\mathbf{v}|s)}{p(\mathbf{v})} d\mathbf{v} = I^\star \quad . \tag{10.43}$$

A comparison with (10.5) should clarify how the above generalises the previously discussed optimisation framework. Again using the method of Lagrange multipliers one finds that the optimal probability density is now given by

$$p(\mathbf{v}|s) = \frac{p(\mathbf{v})}{Z(s,\beta)}\, e^{\beta\lambda(\mathbf{v},s)} \quad , \tag{10.44}$$

where

$$Z(s,\beta) = \int_{\mathcal{P}} d\mathbf{v}\, p(\mathbf{v})\, e^{\beta\lambda(\mathbf{v},s)} \quad , \tag{10.45}$$

while β is a Lagrange multiplier.

Exercise 10.8 Retrieve formula (10.44) (hard).

The meaning of (10.44) is straightforward: when $\beta \to 0$, the metabolic flux configuration $\mathbf{v}$ becomes independent of s, implying $I = 0$. As β increases, $\mathbf{v}$ and s get more and more correlated, while $p^\star(\mathbf{v}|s)$ tends to get more and more sharply peaked around $\mathbf{v}^\star(s)$. In the limit $\beta \to +\infty$ cells respond to each value of s by selecting the exact flux configuration that maximises λ. To achieve this, maximal I is required. A detailed study of the optimal probability density emerging in this case within a highly coarse-grained model of metabolism has been carried out in [394], showing how complex metabolic strategies (including the coexistence of slow-growing, persistent states with fast-growing ones) arise as optimal responses to a fluctuating environment.

10.6 Discussion

Metabolic variability in cell populations has, as we have discussed, multiple origins, both rooted in unavoidable stochastic effects and (possibly) in the fact that, in certain cases, being heterogeneous can be optimal for a microbial population. Models can account for variability by representing (sufficiently large) populations via probability densities defined on the flux polytope. Two main (different) goals can be achieved. First, one can look for the probability density that yields the best (in a precise sense) description of a set of empirical data. Methods like Maximum Likelihood and Maximum

Entropy provide different, albeit related, approaches to this task. Second, one can formulate optimization problems for populations, whose general solution is a probability density rather than a single flux configuration. Solutions to these problems can highlight how fitness and variability are related in optimal populations, providing useful theoretical benchmarks for real microbial systems. While possibly more demanding from a mathematical viewpoint (and certainly more demanding from a computational viewpoint), these approaches significantly expand the scope of CBMs, including in terms of predictive power. In addition, they can refine the notion of optimality and provide insights into the fundamental principles that govern the organization of metabolism across populations. The question of whether variability confers an advantage to microbial populations is however very general, and goes beyond the metabolic level of CBMs on which we focused here. A broader discussion of these aspects is presented in the Chapter *Cell behavior in the face of uncertainty*.

Economic analogy 10.A: Maximum Entropy economic equilibrium

Most of economic theory relies on the assumption that markets are capable of allocating resources optimally, i.e. so that the utilities of each of the participating agents is maximised (an assumption can can be seen as the analog of each cell in a population maximising its growth rate). In order to achieve optimal states (called 'equilibria' in economics), agents endowed with *a priori* different preferences, resources and goals identify the actions that maximise their utilities (e.g. transactions, trade, or production) and carry them out. This process however can become more and more demanding as the number of agents that take part in the market gets larger and larger, because (in short) the set of viable transactions for each agent can become exceedingly large. How can one describe the equilibria that arise from these situations?

A possible approach, used at least since [395] (see also [396, 397, 398]), is based on the Maximum Entropy principle. The idea, in short, is the following. Once every agent has somehow chosen their preferred actions (i.e. once a system-wide 'configuration of individual actions' has been selected), the market as a whole presents a set of transactions to be carried out that aggregate the choices of individual agents. When looked at the aggregate level, though, each set of transactions can correspond to more than one configuration of individual actions. (This can happen, for instance, because agents have a degree overlap in their characteristics which makes them indistinguishable from an economic perspective.) If one assumes that agents choose their actions at random from their set of viable transactions, then some sets of transactions are bound to be more likely than others, simply because they can be realised in more 'microscopic' ways (for instance, by interchanging agents of the same type). It is then reasonable to think that the likelihood of any particular set of transactions will be larger, the larger the number of microscopic ways in which it can be realised. Taking entropy as a measure of multiplicity, the most likely set of transactions, then, is the one that maximises the entropy.

A model of market where the above program is worked out in detail is found in [395]. The 'statistical equilibrium' theory that follows from the use of the Maximum Entropy principle generalises the standard competitive equilibrium discussed in microeconomics by providing a description of optimality in large markets with heterogeneous participants. This line of work has also inspired further developments that explicitly included agents' heterogeneity into the theory of competitive equilibria [399, 400, 401]. To the best of our knowledge, a similar approach has not yet been used to model heterogeneous microbial systems.