Representation and Decision Making in the Immune System

C. H. McEwan

A thesis submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy

Edinburgh Napier University
December 2010

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification at this or any other university or learning institution.

External examiner	Dr. Andy Hone
Internal examiner	Dr. Alistair Armitage

Director of studies	Prof. Emma Hart
Second supervisor	Prof. Ben Paechter

Copyright

Abstract

The immune system has long been attributed cognitive capacities such as *recognition* of pathogenic agents; *memory* of previous infections; *regulation* of a cavalry of detector and effector cells; and *adaptation* to a changing environment and evolving threats. Ostensibly, in preventing disease the immune system must be capable of discriminating states of pathology in the organism; identifying causal agents or "pathogens"; and correctly deploying lethal effector mechanisms. What is more, these behaviours must be *learnt* insomuch as the paternal genes cannot encode the pathogenic environment of the child. Insights into the mechanisms underlying these phenomena are of interest, not only to immunologists, but to computer scientists pushing the envelope of machine autonomy.

This thesis approaches these phenomena from the perspective that immunological processes are inherently inferential processes. By considering the immune system as a statistical decision maker, we attempt to build a bridge between the traditionally distinct fields of biological modelling and statistical modelling. Through a mixture of novel theoretical and empirical analysis we assert the efficacy of *competitive exclusion* as a general principle that benefits both. For the immunologist, the statistical modelling perspective allows us to better determine that which is phenomenologically sufficient from the mass of observational data, providing quantitative insight that may offer relief from existing dichotomies. For the computer scientist, the biological modelling perspective results in a theoretically transparent and empirically effective numerical method that is able to finesse the trade-off between myopic greediness and intractability in domains such as sparse approximation, continuous learning and boosting weak heuristics. Together, we offer this as a modern reformulation of the interface between computer science and immunology, established in the seminal work of Perelson and collaborators, over 20 years ago.

Publications

Portions of this thesis have appeared in the following peer-reviewed forums

Chris McEwan and Emma Hart, *On Clonal Selection*, Theoretical Computer Science, Elsevier, 2010 (in press)

Chris McEwan and Emma Hart, *Clonal Selection from First Principles*, Proceedings of 9th Annual Conference in Artificial Immune Systems, Springer, 2010

Chris McEwan and Emma Hart, *On AIRS and Clonal Selection for Machine Learning*, Proceedings of 8th Annual Conference in Artificial Immune Systems, Springer, 2009

Chris McEwan and Emma Hart, *Representation in the (Artificial) Immune System*, Journal of Mathematical Modelling and Algorithms, 8:125–149, Springer, 2009

Acknowledgements

There are two things no-one tells you before you embark on a Ph.D. One, is just how difficult it is to produce research that is both competent and original. The other, is how characteristically human objective science can be behind the curtain. With respect to both, I am indebted to the guidance of Professor Emma Hart. It is a pleasure to have the opportunity to formally acknowledge this debt, which I know she would understate in person. Professors Jessie Kennedy and Ben Paechter also provided just the right amount of encouragement mixed with gentle scepticism. I realise not all students are so fortunate and am grateful to everyone for their effort and patience while I found my way.

In the wider scientific community, some people deserve special mention. Mark Neal: for getting me into this fine mess. Jon Timmis: for his support and timely advice, without which this thesis would never have been realised. Thomas Stibor: for showing me that good research does not necessarily come from chasing the glorious visions of a research programme. And lastly, Jorge Carneiro: for his patience answering awkward questions about long shelved work, and whose own thesis was the first to inspire in me what would become my thesis.

As always, Harrie and my family provided unwavering love, belief and support in what proved to be difficult times for us all. They are the giants whose shoulders I stood upon.

Thank you, all.

In honour of my mother.
In memory of her father.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

There is no perfect model. It is not possible to maximise simultaneously generality, realism and precision.
R. Levins

1.1 Motivation

The immune system has long been attributed cognitive capacities such as *recognition* of pathogenic agents; *memory* of previous infections; *regulation* of a cavalry of detector and effector cells; and *adaptation* to a changing environment and evolving threats. These are just analogies, but beg the question about just what mechanistic descriptions could account for this seemingly cognitive behaviour. Such descriptions would be insightful, not only to immunologists, but to those seeking to instill similar autonomy into mechanistic computational systems.

Ostensibly, in preventing disease the immune system must be capable of discriminating states of pathology in the organism; identifying causal agents or "pathogens"; and correctly deploying lethal effector mechanisms. What is more, these behaviours must be *learnt* insomuch as the paternal genes cannot encode the pathogenic environment of the child. At a very high level, the immune system can be observed to be responsible for making a decision between immunity or tolerance, with associated costs in the wrong choices. A mathematical immunologist would attempt to assemble experimentally observed phenomena into mechanistic models that are capable of statistically reproducing this high-level behaviour. Our thesis approaches this from the other direction:

> *Examining the immune response from the perspective of a statistical decision function offers insights and abstractions that can be exploited by both computer scientists and immunologists.*

We intend to portray immunological processes as inherently *inferential* processes. We are by no means the first to take such a position, but our approach is different insomuch as we attempt to bridge the gap between biological modelling and statistical modelling, by treating both at the lower level of numerical methods.

1.1.1 Modelling, inference and immunity

The problem of induction, reasoning from the particular to the general, is of course the very stuff of science: extracting natural laws from experimental observations; reasoning about a population based on a sample of its members; even deriving sufficient axioms for deductive formalism. The workhorse of such reasoning in science is modelling and statistical inference. Whether one intends, for example, a logistic regression model of voter demographics or a differential equation model of biological phenomenon, the modelling cycle is much the same:

(1) Formulate (or reformulate) the model.

(2) Optimise model parameters to align with environmental observations.

(3) Interpret the model in terms of explanatory or predictive consequences.

Notice that (1) traditionally requires some expertise, but this is by no means essential. (2) only requires sufficient time and computational power. It is (3) that is the crux of the scientific method: the attribution of meaning and truth through falsification. This requires cognition. However, if we are willing to relax the constraints of science, then iterating (1)-(3) simply describes *inferring the state of nature by whatever means available*. This does not require cognition.

If one accepts that the immune system must (in some sense) be inferring the state of nature in order to produce responses that carry survival advantage, then it follows that there should be insight to be gained from statistically modelling the immune system's environment, rather than mathematically modelling the immune system *per se*. Going further, one can constrain the statistical model to only use components and mechanisms that are qualitatively similar to those that the immune system has at its disposal. Thus, the boundary between statistical and biological modelling blurs. To be clear, we do not expect to find that the immune system embodies a particular method of statistical inference. Rather, we are interested in what any intersection can tell us about the coarse, robust aspects of the immune response. What is phenomenologically sufficient and what is evolutionarily contingent? What is signal and what is noise?

Phenomenological sufficiency is a rather lofty claim, which we temper with the epigraph from mathematical ecologist Richard Levins that opens this chapter. Levins observed [111] that ecologists, like immunologists, tend to favour realism and precision of their models; whereas physicists, like statisticians and computer scientists, tend to favour generality and precision. Like Levins, our goal is *generality and realism.* Generality allows us to move towards determining sufficiency and contingency in the ever growing mass of disconnected experimental observations that make up the immunology literature. Realism ensures that any abstraction retains biological plausibility and provides both a constraint and an inspiration for developing autonomous inferential processes. Any cost to empirical precision may be offset by the benefit of increased *conceptual precision*: something we will argue is lacking in both theoretical models of the immune response and the computational methods that draw inspiration from them. Clarity here would seem a necessary first step to attaining empirical precision.

1.2 Thesis outline

The title of this thesis refers to the two fundamental problems that need to be solved by any autonomous decision maker. To attack any problem one needs a *representation* of that problem. The obviousness of this statement belies the subtle and complex issues that arise in choosing a representation. For statistical inference, a representation is typically chosen *a priori* by the statistician, but truly autonomous systems (including statisticians) must be able to learn their representation as part of inferring the state of nature. Given a representation, one requires a *mechanism* for decision making; for transforming inferred representations of the state of nature into actions. Our research questions are thus

- Can knowledge of the requirements for statistical decision making be applied to develop a plausible model of processes in the immune system?
- If so, does such a perspective offer novel insight that can be exploited by immunologists, computer scientists or statisticians?

Our general approach is to view both representation and decision-making as problems of *approximation.* By first demonstrating that, when suitably formulated, ecological models of interacting “species” are capable of such approximation, we are then able to elaborate this basic dynamical system with decision making behaviours that are plausible, both statistically and biologically.

1.2.1 Chapter summary

In Chapter 2 we review the representational abstractions and mechanistic models of decision making employed in both contemporary immunology and its applied computational derivative. We will raise several issues regarding the scope of these established methods, which in turn will motivate a review in Chapter 3 of representation and decision making in classical and modern statistical inference. The goal here is to develop the necessary first principles from which we can tighten our problem description, critique existing work and formulate our proposed solution. In Chapter 4 we then wield our immunological and statistical knowledge to assess existing work on immune-inspired inference algorithms. We demonstrate both empirically and theoretically that such methods are compromised, in terms of both computational efficacy and biological plausibility, and provide some constructive suggestions for improving this established "paradigm".

We then depart from this paradigm completely. In Chapter 5 we approach the foundational task of learning a *problem representation* that is immunologically plausible and empirically compare our theoretical results with state of the art algorithms. Armed with a representational abstraction, in Chapter 6 we approach the second problem of designing a mechanism for *decision making*. Again, we empirically assess our theoretical work by comparison with the state of the art. Chapter 7 concludes with some final thoughts and future research directions.

1.2.2 Contribution

Briefly, the contribution of this thesis can be stated as the development of a hybrid statistical and biological model of the immune response that

1. Clarifies and strengthens the representational abstractions and models of decision-making employed by immunologists and computer scientists;

2. Offers a simple formalisation of certain immunological phenomena that are currently awkward or impossible to cast under the prevailing methods;

3. Establishes a foundation for immune-inspired computing that is grounded in the underlying numerical methods of statistical and biological modelling, rather than an analogical "mapping" between domains;

4. Produces a general, adaptive numerical method for approximation that is demonstrably competitive with the state of the art.

As a computer scientist, any purported contribution to immunology is the most contentious to defend. As we review later, the interface between computer science and immunology is founded on out-dated models and opaque informal reasoning. In this sense, our contribution is to offer one possible modernisation of this interface, where the opportunity for contribution between domains is improved.

1.2.3 Methods

Both the critical and constructive aspects of this thesis are presented as a mixture of theoretical analysis supported by empirical validation. The formal techniques employed are quite rudimentary, but draw from independent fields that favour their own nomenclature and idioms. We try to shield the reader from these superficial differences as much as possible. We also avoid breaking the flow of text with laborious digressions; the appendix provides sufficient background material to prepare or refresh the reader.

Some mention should be made of our "empirical validation". Clearly, a modelling approach that favours generality and realism over precision may not be well suited to the type of empirical validation expected by a field immunologist. Indeed, at the time of writing, sufficient data on the physical structures involved in our model do not exist to validate against. This is a chicken-egg problem: one cannot expect such data to be collected without first providing a reason to do so. In the meantime, we make do with demonstrating a qualitative phenomenological likeness when our model is applied to synthetic data with similar statistical properties to what might be expected from biological data. This qualitative validation is extended with quantitative comparisons with state of the art algorithms, which provides insight into the efficacy of our model. We will explicitly assume some relationship between statistical efficacy and immunological efficacy. Whether computational efficacy can be reconciled too is difficult to say because the biological and computational substrates are so different. To validate more than this would require collaboration with immunologists.

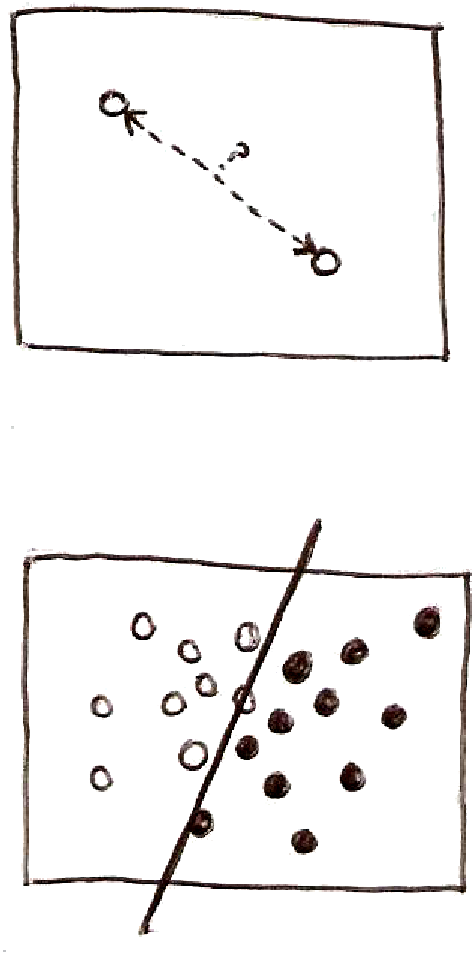


Figure 1.1: An intuitive description of the two fundamental problems in inference and prediction. **Top:** *Representation* of elements of the problem, here represented as points on the plane a distance metric quantifying some relationship between them. **Bottom:** *Discrimination* between elements by some decision-making mechanism, here represented as a linear decision boundary discriminating points on one side (black) from those on the other (white).

Chapter 2

The Real and Artificial Immune Systems

I find it astonishing . . . that this cognitive system has evolved and functions without assistance from the brain.
N. K. Jerne

In this chapter we introduce the necessary background material to frame our problem statement and its eventual solution. We review the appropriate theoretical immunology and how that has been translated into the computational domain. The crux of our argument will be that neither the theoretical immunology nor the computational analogue has an appropriate formal structure to elucidate mechanistic inferential behaviour in the immune system.

2.1 The Real Immune System

Our environment is filled with persistent and novel pathogenic agents with the capacity to invoke illness, disease and death. Once these pathogens have penetrated physical barriers to the body, an elaborate irrigation network drains debris from the tissues to small glands distributed across the body: the *lymph nodes*. It is here that the lion's share of the immune response is initiated and maintained.

The agents that identify and eradicate pathogenic agents are a subset of the white blood cells, or *leukocytes*. Dendritic cells are responsible for sampling debris from the tissues and delivering this "information" to the lymph nodes, where they present fragments of debris on their surface for later inspection. They are referred to as *antigen presenting cells* – "antigen" being a generic term for any chemical structure capable of invoking an immune response. The cells that antigen are presented to, the *lymphocytes*, are the central components of the adaptive immune response. A human adult's immune system contains of the order of 10^{12} lymphocytes, with a daily turnover of around 10^6 [94]. The reason for this constant turnover is that each lymphocyte is differentiated by a single cell-surface receptor configuration. Of the order of 10^5 identical receptors coat the surface of the lymphocyte and are able to bind, more or less, to particular antigenic structures that physico-chemically complement them. It is believed that the variability of receptor configurations is of the order of 10^{16}. Lymphocytes are generated with an essentially random receptor configuration by the translation and manipulation of genes encoding the receptor's binding regions.

The lymphocytes are further sub-divided into B-cells and a suite of T-cells. B- and T-cells are distinguished by both the form of their receptors and the roles each play in the immune response. There are several variants of T-cells, but for our purposes it is sufficient to note that they are believed to fulfil a co-ordination role in the immune response. In contrast, activation of B-cells results in the massive secretion of soluble versions of their receptor, *antibodies*, that traverse the vascular system and tissues, binding to antigen and signalling their eventual destruction. For our purposes, the exact process of antigen destruction is less relevant than how these basic components, randomly generated in massive quantities, are able to orchestrate an exquisite systemic response; balance aggressive immunity with destructive auto-immunity; and discriminate between pathogenic and benign substances that are made from the same raw materials. This problem is referred to as *self/non-self discrimination* and is the principle phenomenon to be explained (or dismissed) by any mechanistic description of immunity [108, 171].

2.1.1 Self/Non-self discrimination

In 1957, Burnet proposed[1] the clonal selection theory [25] whereby antigen *select* their responding lymphocytes, by virtue of binding between the cell-surface receptor and antigenic fragments of the pathogen, such as surface proteins. Induction of the lymphocyte by receptor binding is a function of binding strength, which in turn is determined by how well the antigen-receptor complexes complement each other physio-chemically. The stronger this *affinity*, the stronger and more prolonged the induction of the lymphocyte, leading to proliferation into a *clone* of lymphocytes with similarly configured receptors. Thus, a repertoire of individual cells gives way to the growth and decay of clonotypes.

During proliferation, daughter cells undergo *somatic hyper-mutation* of the

[1] A similar proposal was presented by Talmage around the same time and both were a cell-based refinement of the natural selection theory of Jerne. See [25] and references therein.

receptor encoding genes, resulting in their being low-fidelity copies of the mother. Thus, daughter cells may also bind the same antigen, but with more or less affinity. As daughter cells compete for antigenic binding and subsequent proliferation, the clone evolves towards a high-affinity receptor configuration – a process called *affinity maturation*. The overarching result is that randomly generated receptors evolve into clones of antigen specific detectors under asexual Darwinian selection.

It is no exaggeration to say that clonal selection is the keystone of modern immunology and it is backed up by considerable experimental evidence [105, 161]. But it *does not* account for self/non-self discrimination (a term actually coined by Burnet). Consider that nowhere in the above description is there a semantic difference between antigen produced by the host and those scavenged from an invading pathogen. We now briefly review the evolution of self/non-self discrimination models in the immunology literature (see also Fig. 2.1).

Negative and positive selection

Burnet's solution to the failure of clonal selection to account for self/non-self discrimination was to posit a mechanism whereby newly created self-reactive lymphocytes were eradicated prior to release into the periphery [94] (i.e. negative selection). Although there is evidence that such processes do occur at some level, such as the early selection of T-cells in the thymus, this more than smacks of teleology – *the immune system does not react to self because it removes all components that react to self.* There are a number of logistical problems with this proposal. First, how is "the self" systematically checked for each lymphocyte prior to release, given the large multitude of differentiated cell types and antigen in the body – particularly as the self changes during the life history of the individual, such as puberty and pregnancy [127]. Second, if cell receptors undergo mutation in the periphery, then what is to stop mutation from a non-self-reactive receptor into a self-reactive receptor? Even if such checks do occur, it might reasonably be expected that the periphery would still contain self-reactive lymphocytes. In fact, experimental evidence shows that self-reactive lymphocytes are abundant in healthy human and mouse immune systems [28, 108].

Two-signal models

The greater appreciation of the independent roles of T- and B-cells lead to several models that make explicit use of this fact. Cohn et al. [22] proposed several models where antigen binding is only the first stage of B-cell activation; full activation requiring a second signal from a *helper* T-cell that may (or may not,

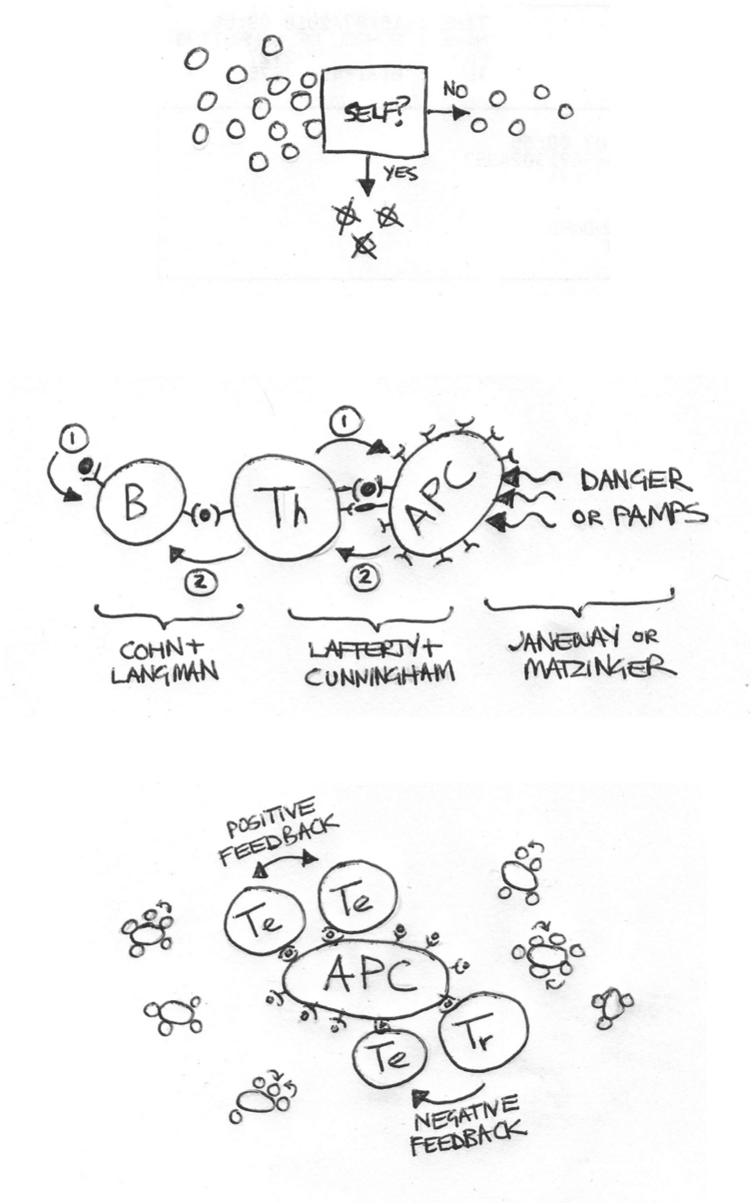


Figure 2.1: A schematic depiction of the principle models of discrimination in immunology. **Top:** Burnet's negative selection posits self-reactive clonotypes are eliminated prior to release into the periphery. **Middle:** The evolution of two-signal models from Cohn and Langman to Matzinger's Danger theory. Responsibility for the self/non-self distinction is delegated further along the chain of interactions until reaching the innate immune system. **Bottom:** Carneiro et al's cross-regulation model. Rather than posit a "switch" for the immune response, they assert that a response is the emergent result of systemic dynamics between pro-response effector and anti-response regulatory T-cells conjugated on the surface of antigen presenting cells.

depending on the model) recognise the same antigen. Here, the self/non-self decision is driven by the presence or absence of helper interactions. The absence of help was suggested to cause permanent inactivation of a responding clone; the idea being that initially responding clones, during foetal development, would tolerise to the relatively self-only environment in the womb due to the absence of pre-existing help. This is not entirely supported, experimentally. A more experimentally valid two-signal model was proposed by Lafferty and Cunningham [103] in which T-Helper cells require both an antigen-binding signal and an antigen-agnostic *co-stimulation* signal from the antigen presenting cell. Such co-stimulation has been experimentally verified. Again, the idea is that the lack of such co-stimulation inactivates T-Helper cells, curtailing the remaining chain of events that would ultimately result in an immune response. The problem here is that antigen presenting cells present *both* self and nonself antigen, so the distinction problem essentially remains unanswered – only delegated to a different cell type. Recent attempts to resolve this conundrum assert the fundamental role of the innate, evolutionarily ancient, leukocyte components for co-ordinating the vertebrate adaptive response [106, 93, 107].

Janeway and Medzhitov [133] proposed that recognition of so-called *Pathogen Associated Molecular Patterns* by germline-encoded receptors on antigen presenting cells would provide the necessary second signal that had the correct self/non-self semantics. That is, the second signal is an assertion that evolutionarily conserved signatures of pathogenicity have also been observed while presenting. This somewhat implies that pathogens enter the body pre-labelled as toxic. Like negative selection, one might expect this to be true in certain cases, but given that pathogens evolve significantly faster than the host germline, PAMPs would not seem to provide a definitive resolution of the self/non-self problem. For example, it is well known that the immune system is able to respond to synthetic, man-made antigen that could never exist in nature [164, 99].

In contrast, Matzinger [124, 125, 126, 127] proposed a variant on this idea where the second signal does not come from signatures on the pathogen themselves, but from so-called *Danger Signals*, such as heat-shock proteins, that are produced by somatic cells undergoing unnatural stress or death. In this case, the second signal is that cell death or stress was also occurring during collection of antigenic debris in the tissues. This is a very elegant explanation: it only depends on the pathogenic effect of pathogen, rather than their physical form; and such somatic signals will evolve at the same rate as the host species germline. Rather than self/non-self discrimination, there is only danger/non-danger which, Matzinger argues [6], is not simply a relabelling of terms. We accept this distinction, but the notion that the evolutionarily ancient innate immune system drives the adaptive immune response is not entirely satisfactory. If the innate immune system is capable of self/non-self discrimination, then what is the evolutionary advantage of the vertebrate adaptive immune system?

Systemic models

A lineage of work in immunology has shunned the reductionist idea that a single "switch" for immunity can even be located. Rather they take a more systemic view that the immunological decision emerges as a result of the dynamical interactions of self-reactive and nonself-reactive lymphocytes.

Parts of this work can trace its history back to N. K. Jerne's seminal *Idiotypic network* theory [98]. Briefly: Jerne observed that given the fact that B-cell surface receptors are naturally the correct size to bind other cell-surface receptors, the lymphocyte repertoire could form a network of interacting clonotypes. He proposed that this network would be self-regulating under a combination of stimulatory and suppressive interactions. Such inter-cell interactions have since been experimentally verified, but the grand self-regulating network is generally considered implausible. Jerne's ideas were developed and formalised mathematically by other giants of theoretical immunology, notably Antonio Coutinho, leading to the so-called "second-generation immune networks" [179, 57]. These incorporated the fundamental role of clonal selection in the immune response, but gave responsibility for *tolerance* to a network of self-reactive clones which did not react to self by virtue of being caught up in the dynamics of network interactions. With characteristic rhetorical flourish, Varela described this network as the immune system's "internal image" of the self; referring to this paradigm as *self-assertion* [165], because the network of self-reactive clones dominated the capacity for the immune system to invoke its default response behaviour. In 1996, Jorge Carneiro elucidated mathematically a mechanism that would allow such self-assertion dynamics to occur [28]. This model also relaxed the reliance on the implausible idiotypic network, although it was still firmly rooted in B-cell interactions. With the developing understanding of T-cell phenotypes it was a small step to recognise that Carneiro's basic mechanism was in fact more plausible when interpreted as a model of T-cell interactions with antigen presenting cells. This led to the so-called *cross-regulation model* [110, 27, 31], which retains much of the systemic spirit of self-assertion, but with a more concise and biologically plausible interpretation. This work is backed up with experimental evidence that *dominant tolerance* – the induction of tolerance by transferring T-cells from tolerant donors – does in fact occur [104, 45]. That is, immunological decisions can be reversed. The fundamental aspect of this model to appreciate here is that the "switch" between tolerance and immunity is now due to a bistable dynamical regime between T-regulatory and T-effector cells bound to the surface of antigen presenting cells. Leon et al. argue [110] that only bistability can account for dominant tolerance, which is anomalous under classical models.

2.1.2 The shape-space abstraction

We have discussed the evolution of theoretical models of mechanisms for self/non-self discrimination. Although this is a central driving force behind any immunological model, if one is to move beyond qualitative conceptual models, then one must formally quantify receptor-antigen interactions. That is, one has to commit to a representation.

Shape-space

Perelson and Oster introduced the shape-space as a simple quantitative model of the immune repertoire [147]. In shape-space, receptors and their ligands are represented as points in an abstract "binding parameter" space, with an isotropic *recognition volume* surrounding each point to account for imperfect matching. Ligands and receptors that have intersecting volumes are said to have affinity – i.e. binding strength is a function of distance in "generalised shape".

The original purpose was to answer questions such as "*given m receptors, what is the probability that a random antigen is recognised?*" [148]. Assuming a recognition region of volume v_i and the total volume of shape-space V, then the probability p that an antigen is recognised by a single clonotype is $p = \frac{v_i}{V}$. It follows that the probability that an antigen is *not* recognised by one of m receptors is $P = (1 - p)^m$. Experimental results estimate that $p \approx 10^{-5}$ of the immune repertoire respond to any given ligand, making P well approximated by a Poisson distribution e^{-mp}. This suggests that a value of $m = 10^6$ would be sufficient to ensure negligible chance of any antigen escaping detection. Such a repertoire would be "complete". This value for m is in agreement with experimental estimation of the smallest known immune system which, Perelson suggests, is because a smaller immune system would offer little protective advantage, e.g. if $m = 10^5$, then $P = e^{-1} \approx 0.37$. A key point to appreciate here is that this model is a heuristic that does not attempt to define the parameters of the space – it only assumes that they could be defined in principle. In particular, p is based on an experimental measurement, not a geometric derivation based on volume ratios.

Perelson and Oster, followed by many others, then went on to provide explicit representations of shape-space, e.g. n-bit binary representations, thus making shape-space an n-dimensional space with 2^n possible shapes. With an explicit representation and affinity function it became possible to computationally simulate an immune repertoire and quantify (in some more-or-less biologically plausible sense) the efficacy of particular models. Although biologically simplistic, for theoretical immunologists the shape-space has a certain heuristic value in quantifying gross properties of the immune repertoire, away from the complex bio-chemical process of protein binding. This seminal work also created the common ground for computer scientists and immunologists. We now present the main criticisms of shape space that influence this thesis.

Theoretical arguments against shape-space

The issues with shape-space as a theoretical abstraction were most notably asserted by Carneiro and Stewart [32]. Their argument is straight-forward: for a theoretical immunologist, deriving an affinity function and its dimensions from the limited experimental knowledge of known binding relationships is clearly ill-posed and data-dependent. Alternately, experimentally validating the parameters of the real shape-space is a "remote goal", which would likely result in a "highly complex, irregular and discontinuous" affinity function. Carneiro and Stewart criticise theoreticians' tendency to not distinguish clearly between these two, quite different, interpretations of shape-space, and thus, avoid the obvious difficulties with either. Furthermore, Carneiro and Stewart's experimental work suggests that shape complementarity is a necessary, but not sufficient, condition for recognition – there is a "relational aspect", not accounted for by the classical lock-and-key metaphor. Carneiro suggests that immunological models should be robust to the exact nature of the affinity relationship. In his own work, this took the form of binding occurring probabilistically without regard for position in shape-space. Receptors bind to multiple antigen that have no geometric relationship to each other. As such, the resulting model's dynamics are not bound to, or a side-effect of, any topological properties of the space it operates in [29, 30, 109].

Experimental arguments against shape-space

More recently, experimental evidence has been growing against the validity of shape-space as an abstraction. It has been increasingly recognised that lymphocyte receptors can bind many *distinct* ligands (*poly-recognition*) and, similarly, a ligand can select many clonotype "specificities" (*poly-clonality*). The general term for this phenomenon is *degeneracy* [189, 134] and there are several authors, in immunology and biology in general, who embrace degeneracy as an important feature of biological systems [62, 61, 174, 135, 187]. For immunology in particular, the most pressing question is how the high-level of specificity in immune responses emerges from these degenerate interactions [89, 188]. It logically follows that if receptors can bind many *distinct* ligands, then the notion of an isotropic recognition volume and affinity as metric distance is inappropriate. Indeed, the shape-free probabilistic binding model of Carneiro becomes a more accurate representation of the actual biology, rather than just good modelling practice.

There is some ambiguity in the literature as to what the shape-space represents. It is used to abstract both T-cell and B-cell binding models, but a crucial biological detail is that both have morphologically different receptors and bind to entirely different structures during the course of a response. At best, one could argue that T- and B-cells just "live in" different shape spaces. But how different?

For B-cells, it is important to realise that a binding site (epitope) is not a predefined object. It is an arbitrary discontinuous region on the three-dimensional surface of a molecule. It comes into being as an epitope by virtue of binding to a receptor, that is, in the context of a particular interaction [85]. The whole surface may have, so to speak, "epitope potential". To appreciate what makes up the binding site, it is useful to elaborate on the basics of protein structure. A protein is a long chain of shorter structures, called peptides, which are themselves chains of amino acids. Laid out as a long chain, this is referred to as the protein's *primary structure*. During synthesis, the protein undergoes a complex folding process which, ultimately, results in a three-dimensional *tertiary structure* where some peptides are buried inside the structure and others are brought together on the surface (see Fig. 2.2). The significance of this is that B- and T-cells sense different aspects of the protein [94, Sect. 3.11]:

> *Antigen recognition by T-cell receptors clearly differs from recognition by B-cell receptors and antibodies. Antigen recognition by B cells involves direct binding of immunoglobulin to the intact antigen and [...] antibodies typically bind to the surface of protein antigens, contacting amino acids that are discontinuous in the primary structure but are brought together in the folded protein. T cells, on the other hand, were found to respond to short contiguous amino acid sequences in proteins. These sequences were often buried within the native structure of the protein and thus could not be recognised directly by T-cell receptors unless some unfolding of the protein antigen and its 'processing' into peptide fragments had occurred.*

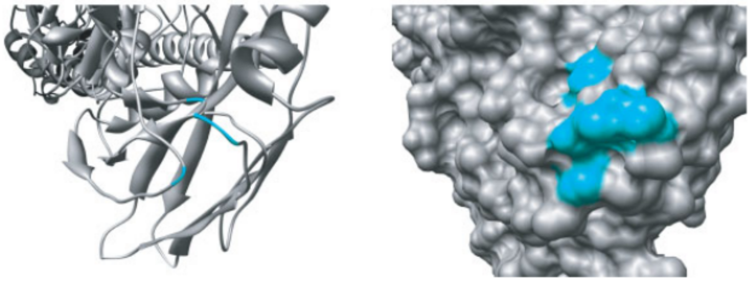


Figure 2.2: A discontinuous epitope on a protein consists of residues that are distant in the primary sequence, but close when the protein is folded into its native three-dimensional structure. All of the residues are required for recognition by the antibody and thus are not epitopes on their own. Approximately 90% of ligands are discontinuous. Reproduced, in part, with permission from [85].

Philosophical arguments against shape-space

With self/non-self discrimination and cognitive analogies abound, it is little wonder that some immunologists take occasion to ponder the more philosophical aspects of their muse [171, 170, 91, 40, 41]. Here we concentrate on two influential propositions that are not so much philosophical arguments *against* shape-space, so much as compelling proposals that are unrealisable under this abstraction.

Francisco Varela was an influential cyberneticist, cognitive scientist and theoretical immunologist. His ideas were largely driven by his phenomenological philosophical leanings and his early work with Maturana on the so-called autopoietic theory of the biology of cognition and behaviour [123]. Varela often referred to the immune system as a “cognitive network” [178] much like the neural system – though an order of magnitude larger and inherently mobile. A recurring theme in his theoretical immunology was that the immune system *constructs* its own internal representation of “the self”. We find this argument compelling, less for philosophical reasons, but because this is a fundamental task for autonomous inference. Irun Cohen is an experimental and theoretical immunologist who has expressed several radical ideas that have generated interest in the computational community [44, 42]. The relation is quite natural: Cohen commonly refers to the immune system as a “computational system” (and also as a cognitive system, though this interpretation of cognition seems weaker than Varela’s). Essentially, he sees the immune system as performing a non-classical distributed computation on the state of the body, with feedback mechanisms that govern the computation’s evolution [43, 92, 39]. The purpose of this computation is *maintenance* – inflammation, healing, garbage collection, and so on – with the immune response reduced to an extremal form of this maintenance. One of his most influential ideas is *co-respondence* – how coherent system-wide responses emerge from the local interactions of diverse, contradictory components with limited sensing and effecting capabilities [42]. Note that neither author provides explicit mechanistic explanations of these mysterious phenomena. Later we will offer precise quantitative interpretations of these philosophical proposals.

Our assertion that shape-space cannot realise these ideas will become formally clearer in later chapters. For now, an intuitive argument may suffice. Conceptually, the shape-space portrays the immune repertoire as a collection of abstract points spread out in an abstract space, with binding affinity a function of distance or pattern-matching. Observe that, in order to *construct* a representation one needs *building blocks*, not point-wise comparisons between atomic entities. Observe also that, by definition, locality in shape-space is anathema to degenerate, contradictory, systemic interactions. Any response from such a localised repertoire of receptors implies a decision function biased by the rule that "like begets like" – that nearby points have similar self-ness or nonself-ness. Given that both self and non-self must prefer proteomic forms that are functional over forms that are "close", such an inductive bias would seem physiologically limited.

Computational arguments against shape-space

The implicit assumption behind computational models in immunology is that "shape" can be abstracted from its physico-chemical reality without affecting the logical behaviour of the model, i.e. self/nonself discrimination. First, we make a general observation about n-dimensional spaces. In an n-dimensional shape-space, the search space for the immune repertoire is of the order $O(c^n)$ where c is a constant. Such exponential scaling is computationally abhorrent for even small c and moderately sized n. Let us assume, like Perelson and Oster, that $p = \frac{v_i}{V}$ represents the probability that a receptor binds antigen. But now, let us also introduce an explicit representation. Without loss of generality, we model both v_i and V as cuboid regions with sides of length l and L, respectively[2]. That is, $v_i = l^n$ and $V = L^n$ and thus $p = (\frac{l}{L})^n$, which clearly shrinks exponentially in n. To provide a sufficient covering of the space for repertoire "completeness" requires of the order $(\frac{L}{l})^n$ receptors. If one is to avoid the necessary exponential increase in repertoire size, the only alternative is to increase the volume of recognition by increasing l; that is, to *decrease* receptor specificity. Let us assume it desirable to retain a fixed value of p, such as that derived experimentally by Perelson. For

[2] In fact, extending this analysis to spherical volumes produces even *worse* results. A well known result is that the volume of a hyper-sphere approaches zero as dimensionality increases! See e.g. [3, 58, 90, 16, 167] regarding the break down of geometric intuition in high-dimensions.

fixed p, as n is increased $\frac{l}{L} = p^{\frac{1}{n}} \rightarrow 1$ very rapidly. In other words, the necessary length of l to retain a given value of p rapidly approaches L; that is, each receptor can bind to almost all of the space.

We will revisit this phenomenon of dimensionality in Chapter 3. The point we wish to make here is that the dimensionality of the shape-space cannot be abstracted away. The same logical functionality may not be retained in arbitrary large shape-spaces. Some authors (e.g. [32, 163]) have speculated that the "true" dimensions of shape-space may be anywhere from 5-20 dimensions. Taking Perelson's value of $p = 10^{-5}$, Fig. 2.3 shows that values between 5 and 20 dimensions would still entail very low specificity to attain completeness.

2.2 The Artificial Immune System

On the back of our immunology review, it will be convenient to introduce the foundations of immune-inspired computing. We will provide deeper theoretical and experimental analysis of immune-inspired algorithms in Chapter 4.

2.2.1 From *in vivo* to *in silico*

As we have previously noted, once receptors and antigen are given an explicit form, nomatter how abstract, then large-scale computational simulation becomes a viable alternative to solving minimal mathematical models of elements that "bind" in some unspecified way. The seminal work at the interface of computing and immunology was carried out by several notable researchers: Forrest [70] explored the similarity between clonal selection and natural selection using the methods of evolutionary computing; Farmer, Packard and Perelson [66] followed a similar line of research, noting similarities between aspects of immunological processes and Holland's Learning Classifier Systems from Operations Research; Varela and Bersini [15, 14, 13] took a more cybernetic approach, applying immunological ideas to reinforcement learning and control problems.

Forrest et al's early work, together with Perelson's shape-space abstraction, proved to be particularly influential. Their position that *"the genetic algorithm without cross-over is a reasonable model of clonal selection"* produced a dogma that largely persists to this day. The intuitive mapping from the immunology to population-based algorithms in metric-space was sufficiently compelling: Cutello and Nicosia [33] developed the pattern recognition aspect of Forrest et al's work; soon after, de Castro and Von Zuben proposed their seminal data-analysis and optimisation algorithm [54]. But it is in the black-box stochastic optimisation

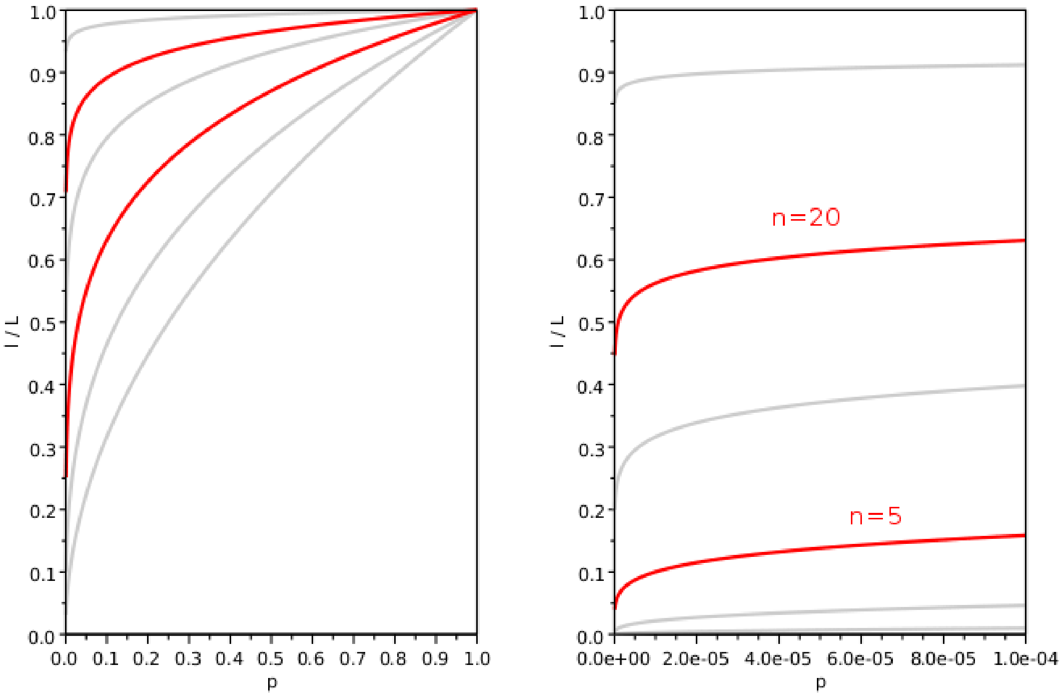


Figure 2.3: The effect of shape-space dimensionality on the probability of antigen recognition and the width of recognition volumes. Curves represent 2, 3, 5, 10, 20 and 100 dimensions, with 5 and 20 highlighted as speculated plausible values. The coverage of each dimension l for a given recognition volume rapidly approaches the width of the space L. On the right we zoom in on the range of Perelson's experimentally derived value $p = 10^{-5}$, which still entails low receptor specificity to attain repertoire completeness.

setting where this work has particularly flourished [52, 47, 48, 46]. Recently, there have been some promising theoretical developments in this programme (see e.g. [49, 173]), but ultimately, it is difficult to assert that this is not just a variation on the well established theme of evolutionary computing [141].

2.2.2 Constructing artificial immune systems

Today, this line of research tends to come under the banner of *Artificial Immune Systems*, a term established by Timmis and de Castro in their comprehensive review and unification of disparate work preceding the turn of the millennium [55]. In addition to bolstering the field, the textbook of Timmis and de Castro prescribed a framework for producing and communicating such artificial immune systems, which has been adopted widely in the field. They propose three sufficient components that make up such a system:

(1) A representation of immunological elements, e.g. receptors and antigen

(2) A set of functions that quantify element interactions, e.g. affinity

(3) A set of algorithms derived from theoretical models and observed immunological phenomena, e.g. clonal selection, danger theory etc.

This framework is really too general to defend or criticise. But what can be criticised is how computer scientists have chosen to interpret it.

Deconstructing artificial immune systems

In Timmis and de Castro's framework, notice that (1) and (2) will always be intimately related, simply because quantifiable functions must operate on representations. Through a mixture of pragmatism, familiarity and ostensible validity, "shape-space" has become synonymous with the metric spaces $\mathbb{R}^n$, $\mathbb{Z}^n$, and $\{0,1\}^n$ and "affinity" synonymous with their accompanying metrics, such as Hamming or Euclidean distance. It then follows that the responsibility for novel computational methods lies entirely in (3), simply because computational problems are traditionally cast in these same metric spaces. There are two points of contention lurking here: (*i*) that metric shape-space can sufficiently represent immunological phenomena; and (*ii*) that shape can be generalised to the high-dimensional spaces of multivariate data without compromising the functionality of such phenomena. We have already argued that both are wrong, leading to a contradiction in the reliance on (3) for computational novelty and efficacy.

2.3 Conclusion

We have described the key decision mechanisms and representational abstraction employed by immunologists and exploited by computer scientists. At this stage, it is difficult to rigorously assess the different self/non-self discrimination mechanisms, though we have been able to build up a quite cogent argument against the shape-space as a representational abstraction. But metric space is a powerful conceptual tool, not to be given up lightly. Resolving this impedance mismatch is a central aim of Chapter 5. In the next chapter we introduce the statistical approach to representation and decision making. This will allow us to better assess the theoretical immunology, critique existing work in immune-inspired inference, and motivate our contribution to the state of the art.

Chapter 3

Statistical Inference

Be approximately right, rather than exactly wrong.
John Tukey

In our consideration of mechanistic descriptions for the immune system's seeming inferential and predictive behaviour, we now turn our attention to the study of mechanistic inference and decision making: statistics. Here we intend to go beyond the superficial decision function analogy and expose the internal workings of these mechanisms. This numerical perspective will provide the foundation for assessing existing work and proposing an alternative.

3.1 How to make an optimal decision

Statistical decision theory provides the foundational framework that unifies much of the field of statistics [97, 190]. Much of the methodological fragmentation and controversy that actually makes up the field of statistics can be seen as different responses to the utter intractability of decision theory; so it is a good place to start. Decision theoretic problems are specified by way of four spaces,

1. The possible "states of nature" $\theta \in \Theta$
2. The possible experimental outcomes or "observations", $x \in \mathcal{X}$
3. The possible "actions" to be taken $\omega \in \Omega$
4. The possible functions $f \in \mathcal{F}$ for choosing actions $f : \mathcal{X} \to \Omega$

and an additional *loss function* ℓ for quantifying the cost in taking action $f(x)$ when the state of nature was θ, i.e. $\ell : \Omega \times \Theta \to \mathbb{R}$. In the case of *estimation* and *prediction*, the case we will be considering, our "decisions" are about the state of nature, and thus $\Omega = \Theta$ and $\omega = \hat{\theta}$, our estimate of θ. This reflects the fact that we are interested in how the immune system is able to infer the self/nonself-ness of a particular pathogen, i.e. $\Theta = \{\textit{self}, \textit{nonself}\}$ and $\mathcal{X}$ is the space of ligands.

An inconvenient practicality is that the states of nature and the experimental observations are random quantities. The optimal decisions (on average) are those from the decision function that minimises the *expected loss*, or *risk* $\mathcal{R}(f)$

$$f = \underset{f}{\operatorname{argmin}}\, \mathcal{R}(f) = \underset{f}{\operatorname{argmin}}\, \mathbb{E}\left[\ell(f(x), \theta)\right] \tag{3.1}$$

In principle, this expectation should be computed with respect to the joint distribution $P(\mathcal{X}, \Theta)$ over observations and states of nature

$$\mathcal{R}_{bayes} = \sum_{x \in \mathcal{X}} \sum_{\theta \in \Theta} p(x, \theta) \ell(f(x), \theta) \tag{3.2}$$

leading to the *Bayes optimal* decision function. In practice, this is not possible with non-trivial $\mathcal{X}$ and Θ. How one chooses to proceed is a semi-religious decision that has been argued for over one hundred years in the statistics literature. The crux of the debate is the justification for holding either $x \in \mathcal{X}$ or $\theta \in \Theta$ fixed and averaging over the other[1], leading to either of

$$\mathcal{R}_{frequentist} = \sum_{x \in \mathcal{X}} p(x|\theta) \ell(f(x), \theta) \tag{3.3}$$

$$\mathcal{R}_{bayesian} = \sum_{\theta \in \Theta} p(\theta|x) \ell(f(x), \theta) \tag{3.4}$$

In practice, both of these formulations may still be intractable. Another inconvenience is that the space of decision functions $\mathcal{F}$ is infinitely large. Typically, the statistician suggests a smaller family of functions $\hat{\mathcal{F}} \subset \mathcal{F}$ where the search will be concentrated. This makes it highly likely that the optimal $f \notin \hat{\mathcal{F}}$. Instead, we seek the best $\hat{f} \in \hat{\mathcal{F}}$ and acknowledge a necessary cost in *approximation error* due to $\mathcal{R}(\hat{f}) > \mathcal{R}(f)$. Further, recall that we do not know the distributions $P(\mathcal{X}, \Theta)$, $P(\mathcal{X})$ or $P(\Theta)$. The risk must be estimated from n observations

$$\mathcal{R}_{empirical} = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), \theta_i) \tag{3.5}$$

[1] Notice how Eq. (3.3) asserts the Frequentist philosophy that θ may be unknown but it is not random; and that a function should provide performance guarantees over any *potential* sample from $\mathcal{X}$. In contrast, Eq. (3.4) asserts the Bayesian philosophy that θ is indeed random; and a function should be performant on *actual* data, not hypothetical data. Each must adhere to two "ideals" encoded in the equations, all of which are in fact compromises.

and extrapolations made from empirical performance on the observed sample. Unfortunately, we cannot guard against the possibility that the sample may mislead us into choosing a sub-optimal $\tilde{f} \neq \hat{f}$. Thus, we accept an additional cost in *estimation error* due to $\mathcal{R}(\tilde{f}) > \mathcal{R}(\hat{f})$. Lastly, any finite sample from $\mathcal{X}$ may partition $\hat{\mathcal{F}}$ into disjoint subsets that perform identically on this sample. We have no principled way to choose the "best" from the best performing subset.

3.1.1 How to make a pragmatic decision

In classical statistics, *approximation error* is called *bias* (towards an overly simplistic model) and *estimation error* is called *variance* (due to sampling from $\mathcal{X}$). Ideally, a decision function would exhibit low bias and variance. In practice, this is not possible as decreasing the bias, by using a more complicated function, will increase the variance by *over-fitting* to the observed sample – rather than *generalising* to unobserved data [88]. Balancing complexity and performance would allow us to choose good decision functions and help us differentiate between empirically identical functions. We do this by minimising the *regularised* risk

$$\mathcal{R}_{regularised}(f) \quad = \quad \mathcal{R}_{empirical}(f) + \lambda\mathcal{C}(f) \tag{3.6}$$

where $\mathcal{C}$ is some measure of the complexity of f and λ controls the trade-off in performance and complexity. We will see explicit instantiations of Eq. (3.6) later, here we simply note that it quantifies the folk wisdom behind *Ockham's Razor*: the simplest model that performs adequately is preferred.

Two approaches to regularisation

How the modeller proceeds next depends on her intentions for f. If she intends that f be an *explanatory model*, it is wise to start with simple $\hat{\mathcal{F}}$ and add complexity until performance is adequate. This is the classical statistician's approach. If, however, she only cares about producing a good *predictive model*, then it may be better to start with a highly complex $\hat{\mathcal{F}}$, that can describe almost any function, and constrain the model by adding bias and reducing variance [81]. This is the classical computer scientist's approach.

Notice that the statistician's approach is also the traditional mathematical modeller's approach: the theoretical immunologist wants her model to explain some aspect of the immune system. But for the immune system itself, only predictive power carries survival advantage. That is, the immune system's capacity

	$\theta = +$	$\theta = -$	
$\hat{\theta} = +$	TP	FP	$\hat{P}$
$\hat{\theta} = -$	FN	TN	$\hat{N}$
	P	N	T

Table 3.1: The contingency table for $\{\theta, \hat{\theta}\}$ pairs.

for inferring the state of nature may have more in common with the computer scientist's approach, than with theoretical immunologist's.

3.1.2 Comparing decision makers

Before introducing specific predictive models, we briefly address the methodology for assessing and comparing models. This follows on smoothly from the introductory theory and will be necessary background material for interpreting our empirical results in later chapters. Recall that our space of actions, or predictions of the state of nature, consists of two classes, e.g. $\Theta = \{$*self*, *nonself*$\}$. Conveniently, this is the simplest and most well-developed case for assessing decisions between e.g. *rejecting* or *failing to reject* the null hypothesis; labelling observations as *positive* or *negative* examples of a concept; determining *probable* or *improbable* samples from a distribution; or detecting the *presence* or *absence* of a stimuli in a noisy signal; and so on.

For a finite Θ, any given observation $x_t \in \mathcal{X}$ has a finite number of possible $\{\theta_t, \hat{\theta}_t\}$ pairs, which we collect in a contingency table much like Table 3.1. The count data in this contingency table allows us to derive several metrics to evaluate and compare performance. Using the standard terms from signal detection, we name possible pairs as *true positives* (TP), *true negatives* (TN), *false positives* (FP) and *false negatives* (FN). The terms *true* and *false* refer to the correctness of the decision; *positive* and *negative* refer to the predicted classes.

Particular metrics derived from this table are domain specific, but of general use are *accuracy* $\frac{TP+TN}{T}$, *sensitivity* $\frac{TP}{P}$, *specificity* $\frac{TN}{N}$, and *precision* $\frac{TP}{\hat{P}}$. Notice that metrics that span columns of the table, such as *accuracy* and *precision*, are dependent on the true underlying class distribution. This can be an important factor where, for example, if $p(\theta = +) = 0.01$ then a decision function that always decides $\hat{\theta} = f(x) = -$, regardless of x, will have a seemingly impressive accuracy of 99%. In contrast, the columnar metrics *specificity* and *sensitivity* would be 1.0 and 0.0, respectively, indicating a constant decision function.

A classic result, typically cast in the terminology of statistical hypothesis testing, is the Neyman-Pearson lemma [143]. For a given acceptable probability of a *false positive* ($1 -$ *specificity*, or "significance") we seek a decision function that has maximal *sensitivity* (or "power"). If the distribution $P(\Theta)$ is known, then the Neyman-Pearson lemma states that the optimal decision is

$$\hat{\theta} = f(x) = \begin{cases} + & \text{if } \frac{p(x|\theta=-)}{p(x|\theta=+)} < 1 - \textit{specificity} \\ - & \text{otherwise} \end{cases}$$

that is, where $p(x|\theta = +)p(\hat{\theta} = +|\theta = -) > p(x|\theta = -)$. But of course, as the reader might have anticipated, in practice $P(\Theta)$ is never known. Any empirical estimate of it could be led astray by the finite sample it was based on.

Uncertainty, confidence and significance

The frequentist solution to the uncertainty of the optimal decision function attempts to establish the distribution of a metric or statistic (e.g. *accuracy*) that can be used to select the "optimal" decision function by empirical comparison. In this sense, optimality means that one can provide performance guarantees over arbitrary samples from $\mathcal{X}$. A practical problem is that we typically only have one sample X drawn from $\mathcal{X}$. However, by repeatedly sub-sampling (with replacement) from this single sample, we can generate a series of *synthetic distributions* over $\mathcal{X}$ that *could have* produced our observed sample. A fortunate result [64] is that the distribution of the metric over these potential distributions gives us insight into the metric's distribution over $\mathcal{X}$. Given such a distribution, one can assert with $c\%$ confidence that the metric's true value lies in $c\%$ of the volume of the density function. We can now empirically compare decision functions: if the $c\%$ volumes of the metric distributions for two different decision functions are non-overlapping, then we can be $c\%$ *confident* that these differences are not due to chance. That is, the difference is *statistically significant* with probability of error $p = 1 - \frac{c}{100}$. By convention, the choice $p = 0.05$ is often made.

3.2 The alpha and omega of inference

We now make the decision theory more concrete by discussing two fundamental models that typify the extremes of a spectrum of inferential methods and lay the groundwork for the more advanced methods employed later. We assume an $n \times m$ matrix X of n-dimensional observations drawn from $\mathcal{X}$ and an m-dimensional vector θ of the states of nature corresponding to each observation. We also assume $\theta_i \in [-1, +1]$ to blur any distinction between classification and regression. Correct decisions occur when $\text{sign}(\theta) = \text{sign}(f(x))$.

Parametric methods and least squares

The simplest non-trivial functional relationship f that could exist between $\mathcal{X}$ and Θ is linear, $\theta = X'f$, where f is an n-dimensional vector and $'$ indicates transposition. This equation is solved in the same manner as high-school algebra $f = (X^{-1})'\theta$, with the slight complication that inverting a matrix is not as straight-forward as inverting a scalar. In practice, we replace X^{-1} with the Moore-Penrose pseudo-inverse X^{+}, which reduces to the proper inverse for a fully determined system, but also behaves sensibly in over and under-determined systems: solving $\text{argmin}_f \|\theta - X'f\|_2^2$ when there are no solutions (hence the name, least squares); and $\text{argmin}_f \|f\|_2$ s.t. $\theta = X'f$ when there are infinite solutions.

In the typical case of $n \neq m$, the pseudo-inverse is essentially a low rank inversion, where $\text{rank}(X) \leq \min(n, m)$. Appendix A has more technical details. Using the notion $(\cdot)_k^{-1}$ for a rank k inversion, the least squares solution is

$$\begin{aligned} f &= (X^{+})'\theta \\ &= (XX')_k^{-1}X\theta && (3.7) \\ &= X(X'X)_k^{-1}\theta && (3.8) \end{aligned}$$

Typically, one of (XX') or $(X'X)$ will be of rank k, depending on whether $n > m$ or vice-versa. The decision function is then $\hat{\theta} = f(x) = \langle x|f\rangle$ and the decision boundary a hyperplane $\langle \cdot|f\rangle = 0$ where $\hat{\theta}$ changes sign[2]. Given an observation $\hat{x}$ we predict the state of nature $\hat{\theta}$ as

$$\begin{aligned} \hat{\theta} &= \langle \hat{x}|f\rangle && (3.9) \\ &= \left\langle \hat{x}|(XX')_k^{-1}X\theta\right\rangle \\ &= \left\langle \hat{x}|X(X'X)_k^{-1}\theta\right\rangle \end{aligned}$$

This decision function is *parametric* insomuch as it assumes an explicit parametric form for f and attempts to optimise those parameters. The assumption here of linearity makes this decision function strongly *biased*, but this also means that f is not significantly perturbed by random variation in X (see Fig. 3.1(a)).

[2] Technically, as presented this boundary always passes through the origin. An additional intercept term removes this restriction and is trivially incorporated by adding a redundant $x_0 = 1$ component to each vector x. The associated component f_0 will then provide the intercept allowing boundary translation, in addition to rotatation.

Non-parametric methods and nearest-neighbour

Rather than assume a fixed parametric form of the decision function f, one might attempt to directly estimate $P(\mathcal{X}, \Theta)$ using the observed X and θ. Typically one does this by *smoothing* the point mass of each observed x to attribute some probability mass to nearby, but never observed, points, assuming they would have the same θ as the observed x. This is done by way of a *kernel function*, e.g. a Gaussian distribution centred on each observed x. Controlling the level of smoothing is achieved by controlling σ^2, the variance of the Gaussian. To make a decision, we simply compute the expected value of θ conditioned on X

$$\hat{\theta} = f(\hat{x}) = \mathbb{E}\left[\theta|X\right] = \sum_{x_i \in X} p_{\mathcal{N}}(\hat{x}\ ;\ x_i, \sigma_i^2)\, \theta_i \tag{3.10}$$

where $p_{\mathcal{N}}(\cdot\ ;\ \mu, \sigma^2)$ is a Gaussian density parametrised by μ and σ^2. A common variation on this idea is to make predictions based on the observed states of nature for $x \in X$ that are simply *nearest* to the observation $\hat{x}$, e.g.

$$\hat{\theta} = f(\hat{x}) = \sum_{x_i \in X} \mathbf{1}\left[\|x_i - \hat{x}\|_2 < \epsilon\right] \theta_i \tag{3.11}$$

where $\mathbf{1}[\cdot] \in \{0, 1\}$ is an indicator function and ϵ defines the size of the neighbourhood surrounding $\hat{x}$. Alternatively, one can use a fixed number k of nearest-neighbours; this is equivalent to adapting ϵ to the space around $\hat{x}$ to ensure each decision averages the same number of points.

These decision functions are *non-parametric* insomuch as they make no assumption of a parametric form of f. Localisation introduces non-linear decision boundaries shaped by the proportion of labels in any given region of the space $\mathcal{X}$ (see Fig. 3.1(b)). This is much less *biased* than parametric methods, but also much more dependent on the quality and amount of data. Thus the range of f can alter dramatically with even minor variation in X. Controlling the size of the neighbourhood can be used to trade-off bias against variance: small neighbourhoods have low bias, high variance; larger neighbourhoods have large bias, but lower variance. When using k neighbours, variance is fixed and bias adjusted to the local properties of the space.

An (almost) dualistic relationship

We gain better insight into why the non-parametric, low bias, non-linear nearest-neighbour and the parametric, low variance, linear least squares methods represent opposite ends of the same spectrum, by examining the dual relationship in

Eq. (3.9). If we use the dot-product to measure neighbourhood locality[3] then starting from the linear decision function

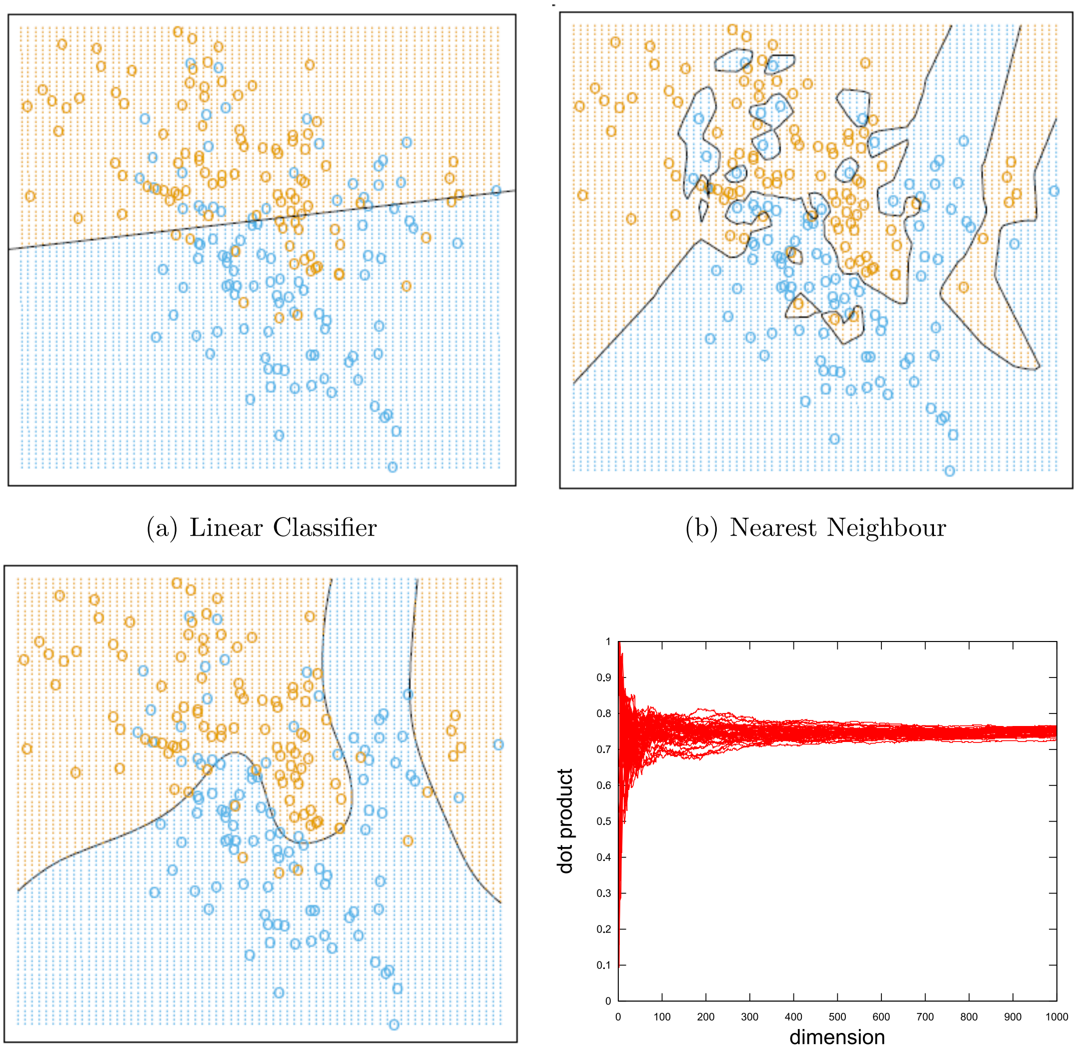
$$\begin{aligned}
\hat{\theta} &= \langle \hat{x} | f \rangle && (3.12)\\
&= \left\langle \hat{x} | X (X'X)_k^{-1} \theta \right\rangle \\
&= \langle \hat{x} | X \alpha \rangle \\
&= \left\langle \hat{x} \middle| \sum_{x_i \in X} \alpha_i \left| x_i \right\rangle \right\rangle \\
&= \sum_{x_i \in X} \alpha_i \langle x_i | \hat{x} \rangle \\
&\approx \sum_{x_i \in X} \theta_i \langle x_i | \hat{x} \rangle && (3.13)\\
&\approx \sum_{\langle x_i | \hat{x} \rangle < \epsilon} \theta_i \langle x_i | \hat{x} \rangle && (3.14)
\end{aligned}$$

we observe that the nearest-neighbour decision, when using the entire X as a neighbourhood, is an approximation to the linear classifier that omits the inversion of $(X'X)$. That is, Eq. (3.13) is “solving” $\theta = X'f$ as $f = X\theta$, rather than $f = X^{-1}\theta$. The benefit of locality-based non-linear decision boundaries in Eq. (3.14) can overcome the cost of this omission (cf. Figs. 3.1(b) and 3.1(c)) but, unfortunately, this reasoning is not as universally applicable as one might hope.

3.3 The Curse of Dimensionality

Understanding the spectrum of inferential methods now puts us in a position to understand a significant obstacle that underlies much of the variations on the alpha and omega in the literature; and is a key aspect of this thesis. The so-called “curse of dimensionality” [11] refers to various practical and theoretical issues that arise due to the simple fact that *as the dimensionality of the space $\mathcal{X}$ increases, its volume increases exponentially faster.* The original problem that led Bellman to coin the wonderful name was computational in nature: the number of grid-based function evaluations in an optimisation problem quickly becomes infeasible with even a modest number of degrees of freedom. But the curse has many different faces and more subtle consequences.

[3] This is less common in practice, but sound. The dot-product is closely related to both the Euclidean distance $\|a-b\|_2^2 = \langle a|a \rangle + \langle b|b \rangle - 2 \langle a|b \rangle$ and (cosine of) the angle $\frac{\langle a|b \rangle}{\|a\|_2 \|b\|_2}$ between vectors, which are more common measures of neighbourhood locality.



(a) Linear Classifier

(b) Nearest Neighbour

(c) Bayes Optimal

Figure 3.1: The theoretical spectrum of inferential models, their decision boundaries and associated errors. **(a)** The linear classifier has high bias (linearity) but low variance as variation in sampled observations does not radically alter the decision boundary; **(b)** The nearest neighbour classifier has low bias but high variance as the decision boundary is explicitly dependent on observed data; **(c)** The Bayes optimal decision boundary derived from the underlying model used to generate the data. Notice that Bayes optimality does not mean 100% accuracy. Although in low-dimensional space the nearest-neighbour appears closer to the Bayes optimal decision function, in high-dimensional space this intuition breaks down, in part, due to effects such as **(d)** the convergence of distances as dimensionality increases. Figures **(a)**, **(b)**, and **(c)** are taken from [88] with permission.

3.3.1 Distance and density in high-dimensional space

In Chapter 2 we introduced some of the unintuitive behaviour of density in high-dimensional space in the context of shape-space. Here we elaborate on the statistical consequences of high dimensionality. These consequences effect all immune-inspired algorithms that are based on shape-space and assume that shape can be generalised to represent multivariate data.

Notice the similarity of Perelson's recognition volume v_i and the neighbourhoods ϵ (or σ^2) used in non-parametric statistics. Recall that a fixed-size volume grows more slowly than the space as dimensionality is increased. For a non-parametric decision function, this results in an increase of the *variance* of estimations made in that neighbourhood because the neighbourhood covers less of the space. In the worst case, there are no neighbours in any fixed neighbourhood, and thus, no generalisation from observations. Again, the only solution is to increase ϵ which, as shown in Fig. 2.3, rapidly approaches the width of the space. This dramatically increases the *bias* of our decision function as desirable non-linearities from localisation are lost. As discussed previously, Eq. (3.12) shows that as ϵ is increased the nearest-neighbour "solution" approaches $\theta = X'f$.

One may attempt to finesse this problem by fixing k the number of neighbours, rather than fixing ϵ, the size of the neighbourhood. In principle, this would allow us to fix variance at some cost to bias. Unfortunately, this strategy does not quite work as expected [4, 16]. A high-dimensional space has 2^n corners where most of the volume is concentrated. The practical consequence of this is that any metric defined across the space becomes increasingly meaningless as dimensionality increases because all data points tend to become equidistant! Figure 3.1(d) illustrates the difference in pairwise distances between 10 uniformly distributed points as the dimensionality of the unit-space is increased. It is clear that all distances are converging. This results in a very fine distinction between an ϵ-neighbourhood capturing either *all or none* of the data-points. Although one can still select the k "nearest" neighbours; they will not be significantly nearer than the other $(n-k)$ neighbours – an implicit low-dimensional assumption behind this strategy. The situation is worse if the data are noisy: even an inconsequentially small amount of noise, once applied to many dimensions, can cause a significant displacement from the original position. In high dimensions, a scalar metric cannot differentiate between two vectors that differ negligibly across all dimensions (and may be the same after accounting for noise) or differ significantly on just a few dimensions (and are clearly different in some respect). The discriminatory power of pairwise distance has been lost.

3.3.2 Dimensionality and under-determinism

High dimensionality does not only affect decision functions that rely explicitly on geometric reasoning. Increasing the dimensions of $\mathcal{X}$, without increasing the number of observations X, affects the nature of the solution to the least squares criterion, making it *under-determined* and numerically *ill-conditioned.* The result of ill-conditioning is a linear model that is nevertheless highly variable due to numerical instability in its solution. In conjunction with under-determinism, the *"least"* squares solution $\text{argmin}_f \|\theta - X'f\|_2^2$ will be non-unique. That is, many linear f will have minimum error. In effect, the linear decision function is too complex; not because of inherent complexity in the model, but because of the paucity of observational data. Using regularisation, we can artificially introduce bias to reduce the variance and reclaim a unique solution that minimises our objective function. The seminal method to do this was called *Tikhonov regularisation* in approximation theory and *Ridge Regression* in statistics [19] and the solution is given by

$$\underset{f}{\text{argmin}}\, \|\theta - X'f\|_2^2 + \lambda\|f\|_2^2 \tag{3.15}$$

where the parameter λ can be used to trade-off accuracy against complexity, quantified in Eq. (3.15) as Euclidean norm. Low norm solutions is our bias. That both measures use the same norm allows us to rewrite the solution as a minor variation on the Moore-Penrose pseudo-inverse $f = (XX' + \lambda\mathbb{I})_k^{-1}X\theta$. This further clarifies how Eq. (3.15) finesses numerical instabilities in inverting XX' by adding a small constant λ to the diagonal components. Variations on Eq. 3.15 will return in later chapters of the thesis.

3.4 The State of the Art

All of the preceding material leads us to a lesson learnt in contemporary statistical learning that is really at the heart of this thesis: *finding good decision functions is, in principle, simple; but crafting a representation that makes finding a good decision function simple can be, in practice, difficult.* This lesson has consequences for anyone conceptualising the immunological decision in shape-space, regardless of whether the intention is biological or computational. We will see these consequences empirically in the next chapter, specifically for immune-inspired inference algorithms. Here we go on to consider the state of the art methods in statistical prediction and machine learning that attempt to finesse the curse of dimensionality and naive representations.

3.4.1 Simple decisions, powerful representations

Recall from Eq. (3.9) the dualism between row-space and column-space derived projections of θ onto $\hat{\theta}$. It is a small notational and conceptual step to go from

$$\hat{\theta} = \sum_{i}^{n} f_i \hat{x}_i = \sum_{j}^{m} \alpha_j \langle x_j | \hat{x} \rangle \tag{3.16}$$

to a more abstract and richer functional form

$$\hat{\theta} = \sum_{i}^{n} f_i \beta_i(\hat{x}) = \sum_{j}^{m} \alpha_j \mathcal{K}(x_j, \hat{x}) \tag{3.17}$$

Both of these generalisations allow us to introduce non-linear transformations in the representation (where we need them) but maintain linearity in the parameters (where it is analytically convenient to do so). The practical significance of either abstraction is that simple linear decision boundaries in the transformed space can produce complex, non-linear boundaries in the original space.

Kernel methods

The kernel function $\mathcal{K}$ on the right-hand-side of Eq. (3.17) returns an inner-product in *some* unspecified, higher-dimensional, non-linearly transformed space [160]. Assuming that $\mathcal{K}$ satisfies some basic properties not relevant here, the representer theorem [182, 157] assures us that variants of Eq. (3.15) can always be represented as a linear combination of the observations X. The main research thrust surrounding this strategy is the so-called *kernel trick* – deriving high(er)-dimensional inner-products in terms of the low(er)-dimensional untransformed vectors[4]. This ingenious trick avoids the computational burden of explicitly working with a higher-dimensional representation and can be used in any context where one would normally use $X'X$, e.g. non-linear principle component analysis [158]. However, this method is better suited to transforming low-dimensional, non-linearly separable observations into high dimensional, linearly separable observations. If the original observations are already high-dimensional, then any measure derived from their dot-product or Euclidean distance may already be cursed, in the sense discussed earlier.

[4] For example, the polynomial kernel $\mathcal{K}(x, y) = (1 + \langle x, y \rangle)^p$ only uses a regular dot-product; but the exponent p returns the same result as if all p-order component products were features, e.g. when $p = 2$ then $x_i \cdot x_j$ is an implicit dimension in addition to the original x_i and x_j. There are many more exotic kernel functions than the polynomial kernel, see [160].

Generalised transformations

The nonlinear transformation implied by a kernel function can be explicitly represented in Eq. (3.17) by $\beta_i(\hat{x})$. The obvious problem here is that we suffer the computational burden of working explicitly in the higher-dimensional space. But the left-hand-side of Eq. (3.17) is much more general. The β_i can represent *any* transformation, with $\beta_i(\hat{x}) = \hat{x}_i$ reducing to the standard linear model. The generalisation of linear models with non-linear functions is common statistical practice (see e.g. [88, Chapter 9]). Using T to represent a transformation that takes an arbitrary vector z to $\tilde{z}$, observe that in the linear case

$$\hat{\theta} = f(x) = \langle f|Tx\rangle = \langle T'f|x\rangle \tag{3.18}$$

that is, $\langle f|\tilde{x}\rangle = \left\langle \tilde{f}|x \right\rangle$. It is the non-linearity of $T(x)$ that breaks this equivalence.

3.4.2 Many simple decision functions

Continuing with the left-hand-side of Eq. (3.17) it is another small step to note that, if β_i can be arbitrary, then why not let it be a full decision function – a transformation $\mathcal{X} \rightarrow \Theta$, rather than between different observation spaces. Thus our representation of $\hat{x}$ becomes the decisions of an *ensemble* of decision functions and the original decision function f becomes a higher-level weighted integration of these low-level decisions [78]. The general goal of ensemble methods is to improve the performance of a single decision function, not by increasing its complexity, but by integrating the results of many diverse decision functions. *Diversity is crucial*, and can be specified in different ways: different function families; different parametrisation of the same family; functions trained on different subsets of the observations, and so on (see e.g. [162]).

Boosting [73, 101] has emerged as one of the most radical and successful instantiations of ensemble learning. The radical aspect is the formal demonstration of the equivalency of *weak* and *strong* learnability[5]: a “weak learner”, performing only marginally better than random guessing, can be aggregated into an arbitrarily strong learning algorithm

$$\hat{\theta} = f_{strong}(x) = \sum_i \alpha_i f_{weak_i}(x)$$

[5] This formal demonstration only holds in the PAC learning framework (see [101, 156] for background and proofs), though the same intuition has been applied very successfully in general.

```
Initialise empty ensemble and uniform distribution over data
$\mathcal{E} = \emptyset$
$D_i = |X|^{-1} \;\; i = 1 \ldots m$
for $t = 1 \ldots T$ do
    Generate a new weak learner with current distribution
    $f_t = \operatorname{argmin}_{f \in \mathcal{F}} \; \Pr_D [f(X) \neq \theta]$
    $\epsilon_t = \ell(X, \theta, f_t)$
    if $\epsilon_t \geq 0.5$ then
        break
    end
    Weight learner based on performance
    $\alpha_t = \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$
    Reweight (and normalise) data distribution based on performance
    $D_i = D_i \exp(-\alpha_t \theta_i f_t(x_i)) \;\; i = 1 \ldots m$
    $D_i = \frac{D_i}{\sum_j D_j} \;\; i = 1 \ldots m$
    Add learner to ensemble
    $\mathcal{E} = \mathcal{E} \cup \alpha_t f_t$
end
```

Algorithm 1: The canonical Adaboost.M1 algorithm

Intuitively, boosting can be seen as the integration of many cheap heuristics that will often fail – but have some edge over random guessing in some circumstances – rather than the integration of a few, strong classifiers as employed by ensemble methods in general. There are still gaps in the theoretical understanding of boosting (see e.g. [75, 132]), but it is well established that a key aspect of its success is that, during training, diversity is induced by re-weighting the observations (see Alg. 1). After each iteration of weak learner construction, successfully classified observations have their weight decreased (and vice-versa), forcing weak learners in later iterations to compensate any predecessor's *bias* and concentrate on observations that are causing continued training error. During decision making, integration across the ensemble increases the confidence in any particular decision, by averaging out the *variance* of individual weak learners. Somewhat contradicting the bias-variance trade-off, we see both a decrease in bias, through diversity, *and* a decrease in variance, through integration (Alg. 1).

3.4.3 Successive approximation

The culture surrounding boosting has a different origin than the rest of the material presented in this chapter; but the statistics community have also provided insightful analysis that is particularly clean [75]. This perspective allows us to introduce much the same ideas without getting bogged down in foreign nomenclature. Indeed, we have essentially already done the introduction in Eq. (3.17).

The statistical view of boosting is simply as an additive expansion in a set of "basis functions" $\beta_i(\cdot)$. Quite literally, boosting attempts to approximate the label vector θ using a linear combination of functions of X. This is quite a subtle difference and deserves some reflection. Ostensibly, all decision functions want the distance (or loss) between inferences $\hat{\theta}$ and the truth θ to be as small as possible. The way this is usually cast, as was done in this chapter, is as *learning* the underlying function $f : \mathcal{X} \to \Theta$. In contrast, boosting directly attempts to approximate θ with a seemingly arbitrary collection of functions of X. Recall, predictive power does not necessarily imply any explanatory power.

Fitting residuals

When we measure the loss of a decision function, we are measuring the distance between its decisions $\hat{\theta}$ and the truth θ, e.g. $\|\theta - \hat{\theta}\|_2^2$. The vector $\mathbb{R} = \theta - \hat{\theta}$ is called the *residual.* In the case of a *linear* least squares, the residual $\mathbb{R} = \theta - X'(XX')_k^{-1}X\theta$ cannot be explained any further because $X\mathbb{R} = 0$. That is, the residual is outside of the vector space spanned by the observations – it is in the null-space of X. However, the same does not hold for a nonlinear least-squares model $f(X;\theta)$. What remains in its residual may profitably be used to develop a second model $f'(X;\theta - f(X;\theta))$. Combining these models captures aspects of the observations missed by the first, but picked up by the second. In principle, this repeated fitting of the residuals can construct arbitrarily precise, though possibly overfit, *successive approximations* of θ. Starting from the residual $\mathbb{R}_0 = \theta$

$$\begin{aligned}
\mathbb{R}_{t+1} &= \theta - F_t(x) &(3.19)\\
&= \theta - \sum_{i=1}^{t} f_i(X;\mathbb{R}_i) \\
&= \theta - f_1(X;\mathbb{R}_1) - f_2(X;\mathbb{R}_2) - \ldots - f_t(X;\mathbb{R}_t) \\
&= \theta - [F_{t-1}(X) + f_t(X;\mathbb{R}_t)] \\
&= [\theta - F_{t-1}(X)] - f_t(X;\mathbb{R}_t) \\
&= \mathbb{R}_t - f_t(X;\mathbb{R}_t) &(3.20)
\end{aligned}$$

Note that the dependence on $\mathbb{R}_i$ forces a sequential nature on this approach. The above method is called ℓ_2 boosting [23, 118] (Alg. 2). Notice that if $\theta \in [-\infty, \infty]$ then the subtraction increments components in $\mathbb{R}$ whenever $\text{sign}(\theta) \neq \text{sign}(\hat{\theta})$; and vice-versa. Thus, observations are re-weighted based on their difficulty. Algorithm (1) uses a similar mechanism except that instead of the *residual,*

```
*Initialise empty ensemble and residual*
$\mathbb{R} = \theta$
$F = \emptyset$
**for** $t = 1 \dots T$ **do**
    *Generate a new weak learner with current residual*
    $f_t = \mathrm{argmin}_{f \in \mathcal{F}} \|\mathbb{R} - f(X;\mathbb{R})\|_2^2$
    *Recompute the new residual*
    $\mathbb{R} = \mathbb{R} - f_t(X)$
    *Add learner to ensemble*
    $F = F + f_t$
**end**
```

Algorithm 2: ℓ_2 boosting. The least squares perspective identifies boosting as a variation on stagewise fitting residuals.

```
*Initialise the ensemble and gradient*
$\nabla \ell = \theta$
$F = \emptyset$
**for** $t = 1 \dots T$ **do**
    *Generate weak learner most correlated with negative gradient*
    $f_t = \mathrm{argmin}_{f \in \mathcal{F}} \| - \nabla \ell - f(X)\|_2^2$
    *Weight learner based on performance on actual loss function*
    $\alpha_t = \mathrm{argmin}_{\alpha}\, \ell(\theta, F_{t-1}(X) + \alpha f_t(X))$
    *Add learner to ensemble*
    $F = F + \alpha_t f_t$
    *Recompute the gradient*
    $\nabla \ell_i = -\left[\frac{\partial \ell(\theta_i, F(x_i))}{\partial F(x_i)}\right] \quad i = 1 : n$
**end**
```

Algorithm 3: Gradient boosting. Observing that the residual is the gradient of the squared-error loss function, we can generalise to boosting-*like* procedures for arbitrary (convex) loss functions.

it uses the negative exponential of the so-called *margin* $\exp[-\theta f(X)]$. This also decays same-sign factors and vice-versa, albeit more smoothly.

Using only two fittings was proposed in 1977 by Tukey under the characteristically humorous name, *twicing* [177]. This "stagewise fitting" was not historically popular amongst statisticians, who naturally preferred well-designed models [24]. However, under the severe conditions of modern data analysis, of which Tukey is the grandfather, well-designed models are difficult to construct. Mildly effective heuristics are not. That many weak heuristics can be combined into a strong model, is the conceptual and theoretical crux of boosting.

Generalised residual fitting

One of the key differences between Adaboost (Alg. 1) and ℓ_2 boosting (Alg. 2) is that they are optimising different loss functions – exponential and squared loss, respectively. Friedman [79], Breiman [21] and Mason [122] introduce a generalisation that can be applied to convex loss functions.

Consider an arbitrary initial guess f_0 of a *linear* decision function's parameters. Let f^* represent the optimal function parameters. Then the best movement away from this initial guess would be

$$f_{t+1} = f_t + (f^* - f_t) \tag{3.21}$$

except for the snag that taking this step would require knowing f^*, which is what we are attempting to find! One might attempt to find the optima analytically by solving $\frac{\partial \ell(\theta,\hat{\theta})}{\partial f} = 0$ but it is more instructive here to reach the same solution algebraically. Now, we know that the linear solution is $f^* = (XX')^{-1}X\theta$ and it follows from this that $X\theta = (XX')f^*$. That is, even though we do not know f^*, we do know that its projection over (XX') will be $X\theta$. We also know that f^* must lie in the span of the observations. For any f, the product $(XX')f$ only loses information that is in the null-space of X. Plugging into Eq. (3.21) gives

$$\begin{aligned} f_{t+1} &= f_t + (f^* - f_t) \\ &\approx f_t + ((XX')f^* - (XX')f_t) \\ &= f_t + (X\theta - (XX')f_t) \\ &= f_t + X(\theta - X'f_t) \\ &= f_t + X(\theta - \hat{\theta}_t) \\ &= f_t + X\mathbb{R}_t \end{aligned} \tag{3.22}$$

and we arrive at the same conclusion as numerically solving $\frac{\partial \ell(\theta,\hat{\theta})}{\partial f} = 0$. That is, $X\mathbb{R}_t$ is the (negative) gradient of the the squared error loss function with respect to f_t, and we follow this gradient until it is zero. If we followed this gradient for T steps, then by linearity the accumulation of steps can be left as individual f_t

$$\hat{\theta} = f(x) = \langle x|f_T\rangle = \left\langle x|\sum_{t=1}^{T} f_t \right\rangle = \sum_{t=1}^{T} \langle x|f_t\rangle = \sum_{t=1}^{T} f_t(x) \tag{3.23}$$

where $f_t = X\mathbb{R}_t = -\frac{\partial \ell(\theta,\hat{\theta}_t)}{\partial f}$. With a slight abuse of notation, we can generalise this gradient descent in parameter space, to a non-parametric gradient descent in function space

$$\hat{\theta} = F_T(X) = \sum_{t=1}^{T} f_t(X) = F_{T-1}(X) + f_t(X) = \hat{\theta}_{T-1} - \frac{\partial \ell(\theta, \hat{\theta}_{T-1})}{\partial \hat{\theta}_{T-1}} \tag{3.24}$$

that is, the best f_t is the one that's decisions follow the negative gradient of the loss function in Θ^m. One caveat is that we are constrained to choose functions from $\hat{\mathcal{F}}$, so we choose the f_t that's decisions are closest to this gradient

$$f_t = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \| - \frac{\partial \ell(\theta, \hat{\theta}_{T-1})}{\partial \hat{\theta}_{T-1}} - f(X)\|_2^2 \tag{3.25}$$

which, for squared error loss, $-\nabla\ell = X\mathbb{R}$ and $f(X) = Xf$ and we recover ℓ_2 boosting. This generalised procedure can be plugged back into the boosting framework and applied to arbitrary (convex) loss functions (see Alg. 3).

3.5 Conclusion

We have reviewed theoretical and numerical statistical inference: from abstract decision theory; through the spectrum of (non-)parametric, (non-)linear decision functions; the bias/variance trade-off and curse of dimensionality; to state of the art methods to address these issues.

Coming from the field of Computational Learning Theory, boosting shares its foundations with the seminal work of Littlestone and Warmuth et al. on *Online Learning* and *Weighted Majority Learning* [115, 114]. The boosting process has also been shown to have a game theoretic interpretation as learning an optimal mixed-strategy in iterated games [74]. Intuitively, at least, these ideas would seem relevant to notions of decision making in the immune system. In particular, compare the definition of weak learnability and Cohen's *co-respondence*.

In the next chapter, we turn our attention to existing work in using immunological metaphors to produce novel inference algorithms. These methods tend to revolve around variations of the ideas embodied in Eq. (3.11) thus we can conclude *a priori* that they will not be particularly sophisticated from a statistical perspective. It might be argued that the immunological metaphor contributes something that compensates for the lack of statistical sophistication. We will have to wait and see if that is true.

Chapter 4

Critical Analysis of Prior Work

Passing peer review is better understood as saying a paper is not obviously wrong, redundant or boring, rather than as saying it is correct, innovative and important.
Cosma Shalizi

Having established our statistical foundations we are now in a position to better assess prior work in applying immune-inspired computational methods in the domain of inference and prediction. We will find these methods to be compromised and will offer a reformulation for improving their biological and statistical plausibility. In the remaining chapters, we move away from this paradigm entirely.

4.1 Clonal selection as algorithm

For non-parametric statistics, the mapping between immunological and statistical domains is quite intuitive. For a sample of antigen (data) distributed across shape-space (possibly self and non-self depending on the application) the goal is to generate a repertoire of receptors (prototypes) that are sufficient to capture salient features of the antigen distribution and generalise to unseen antigen. The shape-space abstraction certainly makes it seems plausible that the immune system needs to achieve a similar goal; but it is readily apparent that the Pattern Recognition, Machine Learning and Signal Processing literature abounds with variations on this basic idea. The question we are concerned with in this chapter, is whether the immunological inspiration contributes anything substantive on top of these standard metric-space methods. Of course, published benchmarks claim that they do; yet few offer any real reason as to why that would be so.

4.1.1 The state of the art

We base our analysis on two algorithms for the following reasons: they are widely cited; have been used by more than one research group in published work; and are representative of the approach used by other more proof-of-concept work. Freitas and Timmis [71] review 17 immune-inspired algorithms that all fit this general description. The reader is directed to the primary references for detailed discussion of these algorithms; our analysis relies upon none of those details.

aiNET

For *unsupervised* learning, the seminal clonal selection algorithm is de Castro and Von Zuben's CLONALG (Alg. 4). This work is only of historic interest, later developing into aiNET (Alg. 5) from the same authors [54, 53]. Both CLONALG and aiNET have spawned many derivative algorithms, usually with an application-specific focus; often employing hybridisation with classical methods. Such *ad hoc* domain specific hybridisations are not relevant here. The principle idea behind Algs. (4) and (5) is to use *affinity maturation* to distribute receptors in antigen space. It is apparent from inspection that the major thrust of both algorithms is very similar. Indeed, a large portion of the inner-loop of aiNET *is* CLONALG. What aiNET adds is the suppressive effects of inter-clonal interactions, purported to allow the repertoire to regulate its own size without *a priori* parametrisation.

AIRS

For *supervised* learning, Watkin's clonal selection based algorithm AIRS has garnered significant attention in the literature [183, 83, 184, 185, 84, 186]. The supervised moniker is a red herring – AIRS models each class independently using an unsupervised process (Alg. 6). During decision making, AIRS makes no appeal to immunology and simply selects the k-nearest receptors from any class, using a majority vote to determine the predicted classification. The only significant differences between aiNET and AIRS is that the latter's inner proliferation-mutation loop iterates until the population reaches a desired average fitness. For what follows, these differences will be inconsequential.

4.1.2 Theoretical and empirical analysis

Let us start with a very simple observation, that can be derived from just the loop structure in the pseudo-code directly. These algorithms all process data-

```
$\mathcal{R} = RandomRepertoire()$
for $x \in \mathcal{X}$ do
    $\mathcal{P} = \emptyset$
    // proliferation and mutation
    for $\mu_t \in Fittest(\mathcal{R}, x)$ do
        $\mathcal{Q} = Daughters(\mu_t, ||\mu_t - x||_2)$
        $\mathcal{P} = \mathcal{P} \cup Fittest(\mathcal{Q}, x)$;
    end
    // clonal selection
    $\mathcal{R} = \mathcal{R} \cup Fittest(\mathcal{P}, x)$
    $\mathcal{R} = \mathcal{R} \setminus Weakest(\mathcal{R}, x)$
end
```

Algorithm 4: CLONALG. The set $\mathcal{R}$ of prototypes (i.e. receptors) μ_t is evolved via mutation, proliferation and selection for each datum (i.e. antigen).

```
$\mathcal{R} = RandomRepertoire()$
while ... do
    for $x \in \mathcal{X}$ do
        $\mathcal{P} = \emptyset$
        for $\mu_t \in Fittest(\mathcal{R}, x)$ do
            $\mathcal{Q} = Daughters(\mu_t, ||\mu_t - x||_2)$
            $\mathcal{P} = \mathcal{P} \cup Fittest(\mathcal{Q}, x)$;
        end
        // Delete clones with low antigen affinity
        $\mathcal{P} = \{\mu_i : \mu_i \in \mathcal{P} \text{ and } ||\mu_i - x||_2 > \epsilon\}$
        // Delete clones with high intra-clonal affinity
        $\mathcal{P} = \mathcal{P} \setminus \{\mu_i, \mu_j : \mu_i, \mu_j \in \mathcal{P} \text{ and } ||\mu_i - \mu_j||_2 < \sigma_{intra}\}$
        $\mathcal{R} = \mathcal{R} \cup \mathcal{P}$
    end
    // Delete clones with high inter-clonal affinity
    $\mathcal{R} = \mathcal{R} \setminus \{\mu_i, \mu_j : \mu_i, \mu_j \in \mathcal{R} \text{ and } ||\mu_i - \mu_j||_2 < \sigma_{inter}\}$
    // Generate fresh components
    $\mathcal{R} = \mathcal{R} \cup RandomRepertoire()$
end
```

Algorithm 5: aiNET. Essentially CLONALG with additional prototype interactions that attempt to self-regulate the size and quality of $\mathcal{R}$.

```
$\mathcal{R} = RandomRepertoire()$
for $\{x, \theta\} \in \mathcal{X}$ do
    $\mu_t = Fittest(x, \theta, \mathcal{R})$
    $\mathcal{P} = \{\mu_t\}$
    while $AvgFitness(\mathcal{P}) < \sigma$ do
        for $\mu_k \in \mathcal{P}$ do
            $\mathcal{P} = \mathcal{P} \cup Daughters(\mu_k, ||\mu_k - x||_2)$
        end
        $Cull(\mathcal{P})$
    end
    $\mu_{t+1} = Fittest(\mathcal{P})$
    if $\mu_{t+1} > \mu_t$ then
        $\mathcal{R} = \mathcal{R} \cup \mu_{t+1}$
        if $||\mu_{t+1} - \mu_t||_2 < \epsilon$ then
            $\mathcal{R} = \mathcal{R} \setminus \mu_t$
        end
    end
end
```

Algorithm 6: AIRS training procedure. Similar in spirit to aiNET but has different implementation details and maintains θ-specific repertoires. The test procedure is simply k-nearest neighbour amongst all prototypes.

```
$\mathcal{R} = RandomRepertoire()$
for $\{x, \theta\} \in \mathcal{X}$ do
    $\mu_t = Fittest(x, \theta, \mathcal{R})$
    $\mu_{t+1} = \frac{\mu_t + x}{2}$
    $\mathcal{R} = \mathcal{R} \cup \mu_{t+1}$
    if $||\mu_{t+1} - \mu_t||_2 < \epsilon$ then
        $\mathcal{R} = \mathcal{R} \setminus \mu_t$
    end
end
```

Algorithm 7: AIRS$^-$. The optimal (one step) candidate is chosen deterministically, rather than via AIRS' many rounds of stochastic mutation and resource competition (c.f. Algorithm 6).

points (antigen) sequentially in their outer loop; and perform stochastic search for fitter prototypes (receptors) in their inner loops. Now, if there is only ever one data-point in the space at any given time, then any fitness landscape induced by a pattern-matching or distance function will be uni-modal, with the mode appearing centred on that data-point. Thus, *stochastic search appears to be entirely redundant* as a learning strategy. With only one antigen, affinity maturation is simply an inefficient hill-climb along a gradient that could be derived from the same quantities used to compute the affinity function.

A simple experiment clarifies. We completely remove the generate-and-filter subroutine from AIRS, replacing it with a trivial, deterministic update which we dub AIRS$^-$ (see Alg. 7). Here, we simply generate one mutant daughter exactly halfway between the datum and the best matching receptor. The rest of the algorithm is unchanged. Table (4.1) reports performance comparisons for several benchmark datasets. The figures validate our observation: *the clonal selection phase of AIRS has almost no positive effect on the algorithm's performance.* Not only is the stochastic search unnecessary, it can be detrimental. AIRS performs significantly worse on all high-dimensional datasets. Indeed, on the *newsgroup* dataset AIRS has the same expected accuracy as producing decisions based on a coin flip! For comparison, on the same task 3-nearest neighbour achieves 75% accuracy, linear regression 80% and Multinomial Naive Bayes 97%.

In deriving the deterministic update rule for AIRS$^-$ we simply performed the logical behaviour that AIRS was indirectly attempting by affinity maturation. Regardless of how this behaviour is implemented, we now ask what is this behaviour achieving during training? In AIRS$^-$ we used the update rule

$$\mu_{t+1} = \gamma(x_t + \mu_t), \tag{4.1}$$

where μ_{t+1} is the deterministically constructed best mutant, μ_t is the best matching existing prototype and $\gamma = 0.5$ was the distance to the boundary of the mutation region used in AIRS. Using this fact, we can express (4.1) as

$$\mu_{t+1} = \mu_t + \gamma(x_t - \mu_t) \tag{4.2}$$

of which there are two points to make. First, to generalise back we note that this has the same form as aiNET's "guided mutation" step, where $\gamma \approx \frac{1}{\|x_t - \mu_t\|_2}$. So, aiNET is not only performing random search in a unimodal space, but performing random search along perturbations of the line between x_t and μ_t. Second, Eq. (4.2) is the well-known update rule for MacQueen's 1967 online k-means algorithm [119]. It is also well known (see e.g. [20]) that this strategy implies stochastic

	dimension	AIRS	AIRS$^-$
elements	2	74.35 ± 7.29	71.95 ± 7.72
iris	4	94.67 ± 5.36	94.47 ± 6.34
balance	5	80.93 ± 4.11 *	77.36 ± 4.83
diabetes	8	71.60 ± 4.40 *	69.45 ± 4.98
breastcancer	9	96.28 ± 2.35	96.35 ± 2.19
heart-statlog	13	78.15 ± 8.63	77.11 ± 7.34
vehicle	18	62.05 ± 4.89 *	57.06 ± 6.04
segment	19	88.21 ± 2.48 *	83.79 ± 2.91
ionosphere	34	84.44 ± 5.18	89.66 ± 5.39 *
sonar	60	67.03 ± 11.60	84.58 ± 7.86 *
newsgroup	3783	51.35 ± 4.60	78.87 ± 14.05 *
* significant at p-value of 0.001			

Table 4.1: Accuracy comparison of AIRS and our deterministic derivative. Experiments were performed using the default algorithm parameters, 10-fold stratified cross-validation and a paired T-test. Most datasets are standard UCI benchmark problems (`http://archive.ics.uci.edu/ml/`). *Newsgroups* is a two-class classification of determining `comp.graphics` from `alt.atheism` posts using a subset of the 20 Newsgroup dataset. *Elements* is a synthetic mixture of Gaussians taken from [88] that we will further use in the remainder.

gradient descent on the loss function

$$\ell(\mu_1 \dots \mu_k | X) = \sum_i^k \sum_{x_j \in \mu_i} \|x_j - \mu_i\|_2^2 \tag{4.3}$$

which is the sum of squared distances from prototypes to their assigned datapoints. Note that for k-means the stochasticity comes from computing the gradient using only a single datum sample. The update is deterministic, which involves (i) explicitly moving μ_t to μ_{t+1}, and (ii) monotonically decreasing γ over time to ensure convergence. In contrast, aiNET and AIRS retain one or both of μ_t and μ_{t+1} depending on their pairwise distance and derive γ per datum as an (inverse) function of distance. It seems unlikely this strategy is implicitly optimising anything: it is k-means with variable k and unmotivated randomness.

Reasoning by analogy is not enough

Based on this observation, we hypothesise that, though smaller in size, the AIRS repertoire does *not* compress or otherwise extract meaningful structure from the dataset. We validate this claim by comparing the loss in Eq. (4.3) against that of k-means with the *same number of prototypes* as AIRS memory cells (see Table 4.2). For non-trivial datasets, AIRS is far from the local optima found by k-means. Alternately, we can find the value $\hat{k}$ for k-means that produces the same performance as AIRS. It is apparent that a significantly larger amount of compression is possible than is achieved by AIRS. A similar result has already been demonstrated by Timmis and Stibor for aiNET [166]. By comparing the Kullback-Leibler divergence between a density estimate based on the original data, and one based on the repertoire of memory cells, they demonstrated that aiNET fails to compress non-uniformly distributed data. Although they did not identify the futility of aiNET's stochastic search, they did identify *another* factor that limits its effectiveness, which also applies to AIRS. By enforcing a uniform, fixed width separation between components, both algorithms fail to represent fine-grained structure in the data occurring at a granularity below this width; likewise, both fail to generalise uniform regions with fewer components (Fig. 4.1).

There have been recent attempts in the literature to address these omissions using an "adaptive radius" [17, 180]. The idea behind these methods is that by varying each receptor's recognition volume in inverse proportion to the density of antigen in that region of shape-space, the repertoire can more accurately reflect dense (resp. sparse) regions by packing more (resp. less) receptors into a given region, as appropriate. Technically, this is sound. But practically, this still contradicts the very notion of "compression" as dense antigenic regions must now be represented by very many receptors. The problem here is not so much fixed-size recognition volumes, but the insistence of non-overlapping recognition. To demonstrate the cost of this constraint, Table (4.3) provides a comparison between classification accuracy of AIRS and a classical Radial Basis Function (RBF) classifier fit via the k-means algorithm. This comparison is not entirely fair, as the RBF was fit in a batch setting and thus benefited from random access to the data. But even if we handicap the RBF to only *two* basis functions (cf. the number of prototypes used by AIRS in Table 4.2) it *still* significantly outperforms AIRS on eight of our datasets.

4.1.3 How "immune inspired" should an "algorithm" be?

Having cut through the immunological rhetoric, it is apparent that any biological influence in these algorithms is in fact very weak. Although the degree of biological fidelity necessary for an algorithm to be "inspired" can be a contentious issue, attending to several rudimentary details would significantly increase the validity of the *immune inspired* moniker. After introducing these details, we will demonstrate that they improve the *algorithm* moniker also.

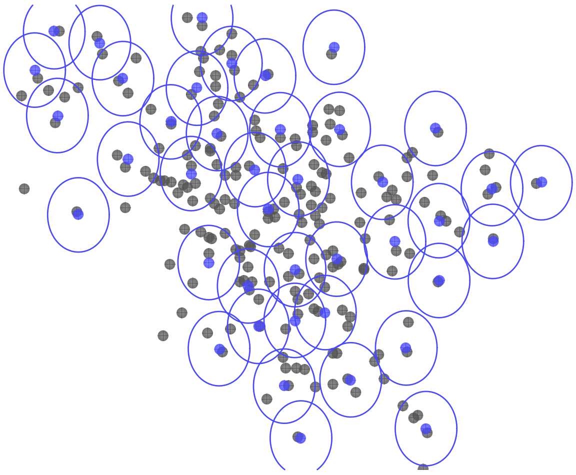


Figure 4.1: Configuration of the aiNET repertoire on the *Elements* dataset, explicitly showing the fixed-width "suppression threshold" used to resolve pairwise competition. It is apparent that although aiNET has fewer prototypes than data, it has not "compressed" the data insomuch as density information has been lost under an essentially uniform tiling. AIRS suffers from exactly the same problem, although the threshold is a hidden parameter in that case. Best viewed in colour.

	k (memory)	AIRS	k-means	$\hat{k}$
iris	47	1.10	0.768	20
balance	295	16.93	13.5	225
diabetes	407	22.81	8.028	125
breastcancer	209	55.22	28.0	100
heart-statlog	209	108.46	9.036	20
vehicle	336	92.50	23.284	25
segment	219	135.81	51.81	45
ionosphere	145	410.66	94.86	12
sonar	143	420.04	38.679	3

Table 4.2: The within-cluster squared distances for AIRS and k-means using the same number of prototypes as AIRS' memory cells. The value $\hat{k}$ is the number of k-means required to produce the same performance as AIRS. These figures suggest that, although smaller than the dataset, the AIRS repertoire has not extracted meaningful structure. This is further illustrated for a two-dimensional dataset in Figure 4.1.

	AIRS	RBF (2)
balance	80.93± 4.11	86.18± 3.76 *
breastcancer	96.40± 2.18	96.18± 2.17
diabetes	71.60± 4.40	74.06± 4.93 *
heart-statlog	78.15± 8.63	83.11± 6.50 *
ionosphere	85.53± 5.51	91.74± 4.62 *
iris	94.67± 5.36	96.00± 4.44 *
segment	88.21± 2.48 *	87.32± 2.15
sonar	67.03±11.60	72.62± 9.91 *
vehicle	62.05± 4.89	65.34± 4.32 *
elements	69.85±10.69	73.80± 10.28 *
* significant at p-value of 0.05		

Table 4.3: Classification accuracy comparison of AIRS and Radial Basis Functions. The RBF is handicapped to only two prototypes per class. Compare this to the AIRS repertoire sizes in Table 4.2 for the same datasets.

1. **Antigen are not processed sequentially.** This is an artifact of the desire for AIS to perform "online". We agree with this sentiment but strictly sequential processing of antigen is of dubious biological validity and the unimodal fitness function renders stochastic search impotent.

2. **Clones are a population.** This is true by definition, but AIS algorithms typically represent them as discretely present or absent. Without clonal growth and decay, notions such as immunological memory and adaptation are trivialised to GA-like elitism. This is an artifact of practitioner bias towards the methods of evolutionary computing [87].

3. **Selection is an anthropomorphism.** With the exception of selective breeding, the survival of species is not determined by fitness *per se*, but by *exclusion* from garnering limited resources for survival. Furthermore, "fitness" is not an inherent property of species, but must be assessed with respect to the environment and the entire population. This distinction is missing from AIS algorithms, again due to practitioner's algorithmic bias.

4. **Cells are adaptive.** Adaptive sensitivity to prolonged stimulation has been explored by Andrews *et al.* [7] in a modelling context, but is yet to be fully integrated into an algorithmic context.

4.2 Rethinking clonal selection as algorithm

Although the results of Sect. 4.1.2 may seem discouraging, we do not consider this to be the final word by any means. The computational properties of the immune system are a rich topic, and it is only natural that seminal work should have erred on the side of simplification. However, we think it apparent that to address these issues requires more than *ad hoc* modifications.

We propose that the uncomfortable mixture of instance-based/non-parametric statistics and evolutionary computing methods be unified under the setting of probabilistic approximation and estimation. This will not address the fundamental metric shape-space problem of Chapter 2 but it will offer a more general and analytically elegant formulation of the traditional AIS[1]. Further, it will address, albeit rudimentarily, the main biological and statistical criticisms raised in the preceding sections. To appreciate the leap we intend to make, we must first understand the workhorse of statistical model fitting: *the EM Algorithm.*

4.2.1 Expectation maximisation

The basic idea behind the EM Algorithm [56, 131] is to solve a difficult "incomplete" data problem with a simpler "complete" data problem. Often, this incompleteness will be a convenient fiction. We dispense with a fully general introduction and cut straight to mixture models, which are particularly apt in this context and are more algorithmically transparent that the abstract EM "algorithm". Our presentation mostly follows that of [18], where the reader is directed for additional details.

In a mixture model, we postulate an underlying generative model for the observed data $x_i \in X \subset \mathcal{X}$ that is a mixture of simpler distributions

$$p(x_i|\Theta) = \sum_{k=1}^{K} p(x_i|\theta_k)p(\theta_k) \tag{4.4}$$

where θ_k parametrises a member of a family of distributions (e.g. multivariate Gaussians with $\theta_k = \{\mu_k, \Sigma_k\}$). The overarching goal is to find a parametrisation of our mixture that maximises the likelihood of observing the given data

[1] In [130] the authors extend this formulation to address stochastic black-box optimisation, but with only partial success. Here we concentrate on the inferential domain only.

$$\begin{aligned}
\underset{\Theta}{\operatorname{argmax}}\, p(X|\Theta) &= \underset{\Theta}{\operatorname{argmax}} \prod_{i=1}^{N} p(x_i|\Theta) \\
&= \underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(x_i|\Theta) \\
&= \underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} p(x_i|\theta_k) p(\theta_k) \right]
\end{aligned} \tag{4.5}$$

If we knew which component generated each x_i the objective would be greatly simplified, so we assume a hidden vector y where $y_i = k$ if x_i was generated by the component parametrised by θ_k. The likelihood becomes

$$p(X|\Theta, y) \simeq \sum_{i=1}^{N} \log p(x_i|\theta_{y_i}) p(\theta_{y_i})$$

Unfortunately, we do not know y, but given some arbitrary y we do know

$$\begin{aligned}
p(y|X, \Theta) &= \prod_{i=1}^{N} p(y_i|x_i, \Theta) \\
&= \prod_{i=1}^{N} \frac{p(x_i|\theta_{y_i}) p(\theta_{y_i})}{\sum_k p(x_i|\theta_k) p(\theta_k)}
\end{aligned}$$

We now have all the quantities we need to invoke the EM Algorithm. Because y is a random quantity, the goal is to maximise the expected (log) likelihood of the now complete data $p(X, y|\Theta)$

$$\begin{aligned}
\mathbb{E}\left[\log p(X, y|\Theta)|X, \Theta\right] &= \sum_{y \in \mathcal{Y}} p(y|X, \Theta) \log p(X|\Theta, y) \\
&= \sum_{y \in \mathcal{Y}} \left[\prod_{i}^{N} p(y_i|x_i, \Theta) \right] \log \left[\prod_{i}^{N} p(x_i|\theta_{y_i}) p(\theta_{y_i}) \right]
\end{aligned}$$

which, after some manipulation, simplifies to

$$\mathbb{E}\left[\log p(X, y|\Theta)|X, \Theta\right] = \sum_{k=1}^{K} \sum_{i=1}^{N} p(y_i = k|x_i, \Theta) \log p(x_i|\theta_k) p(\theta_k) \tag{4.6}$$

Starting from an initial value Θ_0, the EM Algorithm alternates between calculating the distribution for the *expectation*, holding Θ_t fixed; then *maximising* the likelihood, by updating Θ_{t+1} holding $p(y_i = k|x_i, \Theta_t)$ fixed. Hence, the name. The algorithm is guaranteed to increase the likelihood at each step until a local optimum is reached. Algorithm (8) describes the steps for Gaussians mixtures.

4.2.2 The EM algorithm as "simulation"

Looking at Alg. (8) we can identify the rudimentary sense-act loop of a clonal selection simulation. In the E-Step, we first calculate the *demand* on each datum $\sigma_i = \sum_k p(x_i|\theta_k)p(\theta_k)$ before allowing components to *sense* the environment by allocating data proportionally to each component's contribution to the demand $\gamma_{i,k} = \frac{p(x_i|\theta_k)p(\theta_k)}{\sigma_i}$. In the M-Step, each component *acts* by moving μ_k, adapting its distribution Σ_k, and updating its prior π_k. It is this basic connection we will now develop to make the translation to models that may have qualitatively different "actions" from those derived from differentiating the global log-likelihood with respect to the parameters.

Population as prior

Treating the prior π_k as clonotype population carries a particularly attractive connection to dynamical models of evolutionary systems. If one considers a Bayesian update, e.g. $\gamma_{i,k}$ in Alg. (8)

$$\gamma_{i,k} \approx p(\theta_k|x_i) = \frac{p(x_i|\theta_k)p(\theta_k)}{\sum_j p(x_i|\theta_j)p(\theta_j)}$$

then it has already been observed [159] that this has the same form as the discrete replicator equation

$$x_{k(t+1)} = \left(\frac{f_k(x_{(t)})}{\sum_j x_{j(t)} f_j(x_{(t)})} \right) x_{k(t)} \tag{4.7}$$

where f_k is the replicator's fitness, which we associate with the likelihood $p(x_i|\theta_k)$, and x_k is the replicators population size, which we associate with prior $p(\theta_k)$. The essential dynamics of Eq. (4.7) are that replicators with above average fitness (the denominator) grow, while others decay. For Algorithm (8), a component's prior π_k aggregates this measure over all data points, where each α_k is the sum of individual replicator updates $\alpha_k = \sum_i \left(\frac{p(x_i|\theta_k)}{p(x_i|\Theta)} \right) p(\theta_k)$. Intuitively, components with consistently higher likelihood are rewarded by having their prior (in the next time step) increased. There are two interesting deviations from traditional Bayesian

```
while likelihood not converged do
    $\sigma = \emptyset$
    $\alpha = \emptyset$
    $\gamma = \emptyset$
    E-Step: compute probabilities for the expectation
    for $\mu_k \in \mathcal{K}$ do
        for $x_i \in \mathcal{X}$ do
            $\sigma_i = \sigma_i + p(x_i|\mu_k)p(\mu_k)$
        end
    end
    for $\mu_k \in \mathcal{K}$ do
        for $x_i \in \mathcal{X}$ do
            $\gamma_{k,i} = \frac{p(x_i|\mu_k)p(\mu_k)}{\sigma_i}$
            $\alpha_k = \alpha_k + \gamma_{k,i}$
        end
    end
    M-Step: Update the parameters to maximise the expectation
    for $\mu_k \in \mathcal{K}$ do
        // Update location (mean) of component
        $\mu_k = 0$
        for $x_i \in \mathcal{X}$ do
            $\mu_k = \mu_k + \frac{\gamma_{k,i}}{\alpha_k}|x_i\rangle$
        end
        // Update covariance of component
        $\Sigma_k = 0$
        for $x_i \in \mathcal{X}$ do
            $\Sigma_k = \Sigma_k + \frac{\gamma_{k,i}}{\alpha_k}|x_i - \mu_k\rangle\langle x_i - \mu_k|$
        end
        // Update prior of component
        $\pi_k = \frac{\alpha_k}{\sum_j \alpha j}$
    end
end
```

Algorithm 8: The EM Algorithm for Gaussian mixtures: $p(y_i = k|x_i, \Theta) \approx \gamma_{k,i}$, $p(x_i) \approx \sigma_i$ and $p(\theta_k) \approx \pi_k$. The maximisation of the likelihood has a closed-form solution for Gaussians, where $\theta_k = \{\mu_k, \Sigma_k\}$.

statistics: we are considering iterations where it is the parameters of the model, rather than the data, that is changing; and replicator fitness is typically a function of the population fitness, whereas mixture components do not traditionally interact with each other directly.

Clonal Selection as E-Step

The first contribution is largely *from* the EM algorithm. The key quantity is $p(\theta_k|x_i) \propto p(x_i|\theta_k)p(\theta_k)$. Ignoring the normalising denominator for a moment, this equation states, in words, that the probability that a datum should be assigned to a particular component (cf. clonal selection), is proportional to the probability assigned to that point in space by the component (cf. affinity) multiplied by the prior probability of that component (cf. population). This naturally incorporates the intuition that fitness is a function of *both* binding strength and abundance. Further, the probabilistic interpretation hides awkward geometric notions of affinity, accommodating either biologically or application driven measures. This formulation allows us to address several of the short-comings of existing clonal selection algorithms discussed earlier. By using more than a single datum we now have a complex fitness landscape suitable for stochastic search. Adaptive control of the local bandwidth of component distributions may represent adaptive stimulation. Clones have a rudimentary population and competition dynamic that acknowledges classical models from mathematical biology. We find this to be a compelling list of benefits, which come essentially for free.

Affinity Maturation as M-Step

The analogy continues with affinity maturation insomuch as the overarching goal is to “reparameterise the mixture” in order to optimise *some* quantity. Here the immunological perspective departs from both the regular EM Algorithm and evolutionary approaches to maximising likelihood. If our components are multivariate Gaussians, then by definition the weighted mean is an intuitive location to move a component (this is the M-Step in Alg. 8). But in affinity maturation *the components do not move*: daughter clones spread out into the space; some coming to dominate their parent and siblings. Reparameterise the mixture, for affinity maturation, is not just an update of $\Theta_t \rightarrow \Theta_{t+1}$ but a partial redefinition of the model: components enter stochastically and leave in accord with selective pressures. This further distinguishes clonal selection and affinity maturation from black-box optimising the log-likelihood with an evolutionary algorithm. An evolutionary algorithm’s population would each search for a global optimum of

```
while not converged do
    Sample new components from the current mixture
    $\Theta = \Theta + \{\theta_i : \theta_i \sim \Theta\} \;\; i = 1 \dots k$
    Fit the new mixture model without updating means
    $(\ell, \Theta) = \mathrm{EM}(X, \Theta)$
    Evaluate and remove poor components
    $\Theta = \Theta - \{\theta_i : \theta_i \in \Theta \text{ and } p(\theta_i) < \epsilon_1 \text{ or } \det(\Sigma_i) < \epsilon_2\}$
end
```

Algorithm 9: A modified EM Algorithm for Gaussian Mixtures which uses sampling and exclusion of components instead of relocating existing components. This can be considered as adding a very rudimentary "meta-dynamics" to the EM Algorithm: there is no *a priori* model; poor components are eradicated; and proliferation is proportional to fitness.

$p(X, y|\Theta)$ in Θ-space. In contrast, during affinity maturation each member of the population is searching for its own optima of $p(X|\theta_k)$ in $\mathcal{X}$-space. Any optimisation of $P(X, y|\Theta)$ is implicit in optimising its factors.

4.2.3 Empirical analysis

There is much existing work in the statistics literature on stochastic variants of the M-Step (see e.g. [131, 34, 95]). Much like the stochastic k-means introduced earlier, these methods tend to involve deterministic updates based on a sample of the data, rather than stochastic updates *per se*. However, we are now in a position to make use of stochastic search as our fitness landscape is no longer unimodal. The obvious question is whether an EM-*like* algorithm with proliferation and mutation is a valid technique for fitting mixture models?

We assess this question without getting bogged down in immunology by making three simple changes to Alg. (8). First, we trivially modify the EM algorithm to not update mean locations. After this modified EM Algorithm converges we then, in a surrounding loop, remove redundant components with low priors (cf. clonal extinction) and sample new components from the current mixture to add to the mixture in the next iteration (cf. fitness proportional proliferation and mutation). This process is repeated until the outer loop converges; that is, until the repertoire stabilises on its fit to the data (see Alg. 9).

To reduce the degrees of freedom in our analysis, we will also ignore updating each component's covariance or bandwidth. Note that this is not such a compromise as it was in Algs. (5) and (6) as these "fixed regions" are no longer criteria for discrete pairwise separation and removal. Components are free to overlap. This will necessarily reduce their overall fitness by invoking competition in re-

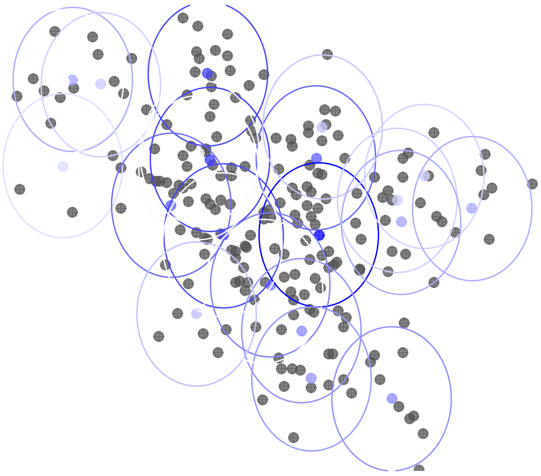


Figure 4.2: Component configuration for Alg. (9) on the *Elements* dataset. Unlike aiNET in Fig. 4.1 components overlap and population levels vary in accord with the underlying prior probabilities; represented here by opacity.

source allocation, but it will also allow the repertoire to properly reflect density in the data. Intuitively, it may be better to compete over a dense region than dominate a sparse region. This intuition is borne out in Fig. 4.2, which shows the configuration of components (i.e. clonotypes) for Alg. (9) on the *Elements* dataset. This configuration should be compared with the aiNET configuration on the same dataset (Fig. 4.1).

One might ask whether the ability of components to overlap reduces the compression ratio of components to data-points. In all our experiments with Alg. (9) the repertoire size never strayed beyond 20-25 components, even though 5 new components were introduced on each iteration for a total of 500 iterations. This suggests that once a stable configuration has been found it becomes increasingly hard for randomly generated components to perturb the repertoire. This suggestion is confirmed by the robust temporal dynamics in Fig. 4.3.

One might also ask how this strategy compares to the EM Algorithm proper. Such a comparison is premature, but it is insightful to consider anyway to motivate further development. In the right hand side of Figure 4.3 we plot the evolution of the likelihoods of observed data (green) and unobserved data (red) drawn from the same underlying model behind the *Elements* dataset. There was no set convergence criteria, but it is clear that from 10 runs with random initial configurations the dynamics do not vary considerably. It is also interesting to

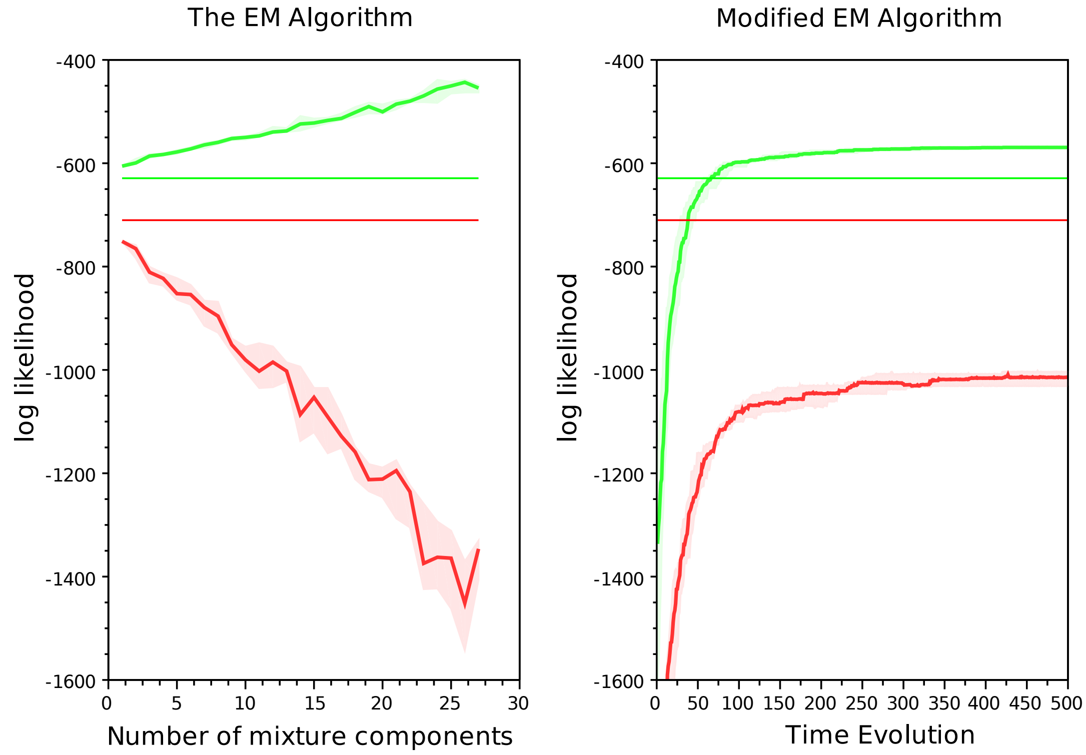


Figure 4.3: Quartiles of observed (green) and unobserved (red) likelihood for the EM and modified EM Algorithm when fit to data generated from the mixture of Gaussians used for the *Elements* dataset. **Left:** The EM Algorithm exhibits characteristic overfitting as the number of components is increased. **Right:** the modified algorithm converges consistently to the equivalent of a 7-component mixture model. The horizontal lines show the same likelihoods under the true generating model. Note that only the y-axes are comparable.

note that at no point does the algorithm overfit to the observed data at some cost to the unobserved performance; but this is most likely explained by the restricted updates making such overfitting impossible. On the left hand side of Figure (4.3) we show the same likelihood measures, but this time for the regular EM Algorithm parametrised with different mixture sizes. Here we see the typical increase in observed data likelihood at the cost of unobserved likelihood as the mixture model's complexity increases and overfits the observed data. The data and y-axis are comparable for these two graphs and it is interesting to note that the modified EM Algorithm performs in-sample roughly the same as a 7-component mixture model (which would be a reasonable choice given the data) although it uses over 20 components and introduces 2,500 components over the course of its execution. Out of sample, the modified algorithm generalises like a 12-component mixture model. That is, it is overfitting above its complementary mixture model in terms of performance on the observed data. It is difficult to say anything general here as performance of the EM Algorithm on unobserved data is not typically reported – its job is to maximise the likelihood, which is maximised by the observed data itself. At the very least, it suggests that there is room for improvement in this basic implementation.

4.3 Conclusion

We have assessed the *status quo* of immune-inspired learning algorithms and found them lacking, both statistically and biologically. A proposed change of abstraction based on probabilistic approximation improves the "metaphor" considerably. This is certainly true from a theoretical perspective; and our initial experiments, though rudimentary, suggest the same might be true empirically.

However, no amount of mathematical elegance can hide that one could derive Alg. (9) without the slightest concern for immunological insight. We are still abstracting the repertoire as points covering shape-space and thus inherit all of the same problems we have bemoaned in earlier chapters. In the remaining chapters, we leave pattern-matching in shape-space behind, in search of degeneracy, constructive representations and systemic responses.

Chapter 5

A Model of Ligand Binding, Competitive Exclusion and Representation Learning

We now move away from models of receptor-ligand interactions in the traditional metric shape-space. In its place, we offer an abstraction that is more biologically plausible, insomuch as it addresses the criticisms we have raised against shape-space and clonal selection as algorithm. By validating our abstraction in terms of computational efficacy, we assert its utility as a basis for both immunological models and applied computational models. This chapter charts the first necessary step in autonomous inference, *representation learning*, on which our thesis of the immune system as a dynamic decision function can build upon.

5.1 Lymphocyte Ecology I

The intuitive notion behind affinity maturation is that a clone "moves" (in shape-space) towards a high-affinity configuration for the antigen inducing proliferation. This is easy to comprehend in the artificial case of one antigen; but in the lymph nodes *many* antigen are being presented at once and what constitutes a high-affinity configuration becomes less clear. For individual clones, there will be a survival trade-off in terms of strongly binding a specific antigen of limited supply, or sufficiently binding many to retain stimulation. Similarly, for the entire repertoire there is a trade-off between maintaining a diversity of clonotypes and allocating resources to specific responses. Just as in natural selection, these issues are not directly resolved by "selection" *per se*, but by the *exclusion* of redundant clones to access limited resources and the subsequent partitioning of resources into *niches.* If one wishes to model stability and diversity phenomena in a population, then one might look to ecology for guidance, rather than immunology. Ecologists have developed a significant body of work around simple, elegant models of interspecies competition [112, 111, 144, 150]. With some notable exceptions [51, 72, 110, 168], what we will call *"lymphocyte ecology"* has been given little attention in the immunology literature; less so in the computational literature, where notions of selection come from evolutionary computing.

5.1.1 The generalised Lotka-Volterra model

The model we will focus on is paradigmatic, originally formulated by Alfred Lotka and Vito Volterra [116, 181], and later developed by many others, notably Levins [111], Roberts [151] and Nowack [145]. It posits an environment with n species where each species has a *carrying capacity*: the maximum population of that species that the environment could support in the absence of any competitors. The independent dynamics of species is initial exponential growth followed by exponential decay towards capacity as the population saturates – the classic sigmoid-shaped curve of the logistic equation. Unlike the logistic equation, reaching capacity is further hindered by competitive effects from other species.

More formally, let ρ_i and k_i represent the population and carrying capacity of the i'th species, respectively. The population dynamics then evolve as follows

$$\frac{d\rho_i}{dt} = \left(\frac{k_i - \sigma_i}{k_i}\right)\rho_i, \quad i = 1 \ldots n \tag{5.1}$$

If $\sigma_i = \rho_i$, then species dynamics are independent of each other and we recover the classic logistic equation. To introduce dependence, and thus competition, we define $\sigma_i = \sum_j K_{ij}\rho_j$, where K_{ij} is the so-called "community matrix" representing the competitive effect of species i on species j. That is, $K_{ij} \geq 0$ and $K_{ii} = 1$ to account for the fact that a species competes with itself. If K is the identity matrix, then $\sigma_i = K_{ii}\rho_i = \rho_i$ and we, again, recover the logistic equation. It is apparent that when $k_i = \sigma_i$ the capacity is equal to the competitive effects and that species reaches equilibrium. If $k_i < \sigma_i$ then the species is out-competed and declines. If $k_i > \sigma_i$ the species grows smoothly towards its, now reduced, capacity. It is straight forward to add additional factors such as growth-decay rates, immigration-emigration terms, and mutually cooperative-antagonistic interactions; but they add little to the immediate exposition.

For the field ecologist, it is often not practical to derive the components of k and K in Eq. (5.1) for the particular ecological community being observed. Nevertheless, a lot of insight and intuition has been gleaned from formally studying artificial ecologies where, for example, the values in K are chosen randomly. Much of the seminal analytical work with this model has been able to elucidate plausible conditions to achieve stability and robustness in ecological dynamics from these synthetic models [80, 128, 129, 175, 153, 96]. We intend to follow this approach in an immunological context. It seem apparent that determining k and K is not going to be any easier for the immunologist[1]. We will only progress if we follow in the ecologist's footsteps and make do with synthetic data. What makes our study different is that we will not use randomness as a model for interactions, but will derive k and K based on explicit receptor and resource representations. However, these representations will not be based on the classical metric shape-space.

5.1.2 A Model of Ligand Binding

To understand our model of ligand binding, we recall the quote from Janeway in Chapter 2. Rather than assume an n-dimensional binding parameter (or pattern matching) space we will explicitly model *epitopes* as compound objects; that is, as *peptides* localised on the surface of the tertiary structure of a protein.

Let us assume n possible such peptides. For example, an immunologically plausible value for n might be 20^9, that is, all possible 9-mer configurations of the 20 amino acids. We will not enforce (or exclude) any further structure on these peptides, which will allow us to abstract from computation and biology. We then define the following

Definition 1. *We abstract the complex surface of a* ***protein*** *as a square symmetric matrix P, where P_{ij} represents the surface correlation of peptides i and j. That is, we do not model the 3-dimensional shape of a protein, but rather provide a statistical description of the surface of this shape. We will further assume that these surface descriptions are additive, in which case individual surfaces can be arbitrarily aggregated to describe compound structures $Q = \sum_k P_{(k)}$.*

Definition 2. *We model a clonotype identifying* ***receptor*** *φ_i as an n-dimensional vector. Each component of φ_i quantifies some ability to bind with the corresponding peptide but, crucially, most components will be assumed zero or negligible. Recall, immunoglobulin does not bind to peptides, but to epitopes, thus binding is a function of multiple peptides being correlated on the protein surface; which we quantified in our previous definition. It will be convenient to set $\|\varphi_i\|_2 = 1$.*

[1] An ecological model would not deny that capacity and competition are fundamental effects driving population dynamics, even if they are difficult to quantify in practice. Compare this with the immunological models of Chapter 2, where such effects are eerily absent.

With the above definition of receptors and ligands, we can now derive the necessary quantities to realise our competitive exclusion model of clonal selection

Definition 3. *Our measure of* ***affinity****, or binding strength, is embodied in the product $\langle\varphi_i|P|\varphi_i\rangle$, which measures the surface correlations on protein P of peptides specific to the receptor φ_i. If P is a projection matrix, i.e. $P^2 = P$, then affinity is the magnitude of φ_i in the subspace of $\mathbb{R}^n$ defined by P.*

Definition 4. *We can now define a clone's* ***capacity*** *as the maximal induction signal available – the sum of affinities to every protein in the environment. Due to additivity, this is simply $k_i = \langle\varphi_i|Q|\varphi_i\rangle = \sum_k \left\langle\varphi_i|P_{(k)}|\varphi_i\right\rangle$. Note that capacity is limited by antigen "supply" as well as receptor-ligand affinity.*

Definition 5. *Finally, we model* ***competition*** *between clones in terms of receptor overlap or receptor-receptor correlation $\langle\varphi_i|\varphi_j\rangle$. Notice that $\langle\varphi_i|\varphi_j\rangle = 1$ when receptors are the same (i.e. of the same clonotype) and $\langle\varphi_i|\varphi_j\rangle = 0$ when there is no overlap. Thus the competitive effect on clone i is an aggregate measure of redundancy and competitor fitness $\sigma_i = \sum_j\langle\varphi_i|\varphi_j\rangle\rho_j$, which includes a clones "competition" with itself, $\langle\varphi_i|\varphi_i\rangle\rho_i$. We collect these correlations in a matrix K that readily satisfies the conditions for Eq. (5.1). Notice that clones should be understood to interact indirectly, through garnering resources that may have otherwise been allocated elsewhere, rather than via receptor-receptor interactions.*

5.1.3 Constructive Approximation

Classically, in biology and computer science, the immune repertoire has been portrayed as points "covering" the shape-space or a population exploring an affinity landscape. Although this has allowed a pragmatic relationship to exist between computer science and immunology, we have argued at length that it does not allow for an *effective* relationship for either. In contrast, we assert the position that, in a quite precise sense, the immune system *constructs* a representation of its environment. That is, that the immune system *approximates* the environment by means of clonotypes and their receptors.

Problem formalisation

The classical approximation problem formulation is to minimise the metric distance between a given vector or function x and its approximation $\tilde{x}$ chosen from some set of elements. Of particular interest here will be additive expansions of basis functions $\varphi_i \in \Phi$ such that

$$x \approx \tilde{x} = \sum_i \alpha_i \varphi_i = \Phi\alpha \tag{5.2}$$

The classic metric of error is the ℓ_2 norm, leading to a least-squares problem

$$\underset{\alpha}{\operatorname{argmin}} \|x - \Phi\alpha\|_2^2 \tag{5.3}$$

which, if the columns of Φ form an orthonormal basis, has the convenient solution $\alpha = \Phi^T x$, that is, $\alpha_i = \langle \varphi_i | x \rangle$. In this case the approximation is exact and easy to compute. But this convenience comes with two undesirable conditions:

1. The constraint of pairwise orthogonality severely limits the form (and amount) of components in the additive expansion [37]. This makes representing some signals extremely convoluted (e.g. representing a sharp, temporally localised wave with periodic functions). This is also a problem when the coefficients of φ_i are to be interpreted (e.g. representing data as a sum of latent factors). In both cases, it is desirable to expand the number and diversity of columns of Φ, resulting in redundant, *overcomplete* representations [5]. However, this results in a non-unique solution to Eq. (5.3).

2. Any least-squares solution to Eq. (5.3) will be *dense*, that is, every basis will contribute to the approximation. In many domains, assuming sparsity in the coefficients is either reasonable or highly desirable. For example, in statistics, one might appeal to parsimony of the model (i.e. *feature selection*); in signal processing, an appropriately chosen basis may induce the representation coefficients to rapidly approach zero, allowing truncation with little perceptible loss in reconstruction (i.e. *lossy compression*).

The ubiquity of these conditions leads to *sparse approximation*

$$\underset{\alpha}{\operatorname{argmin}} \|\alpha\|_{\ell_p} \;\; s.t. \;\; \|x - \Phi\alpha\|_2^2 < \epsilon \tag{5.4}$$

Stated as an optimisation objective, Eq. (5.4) is essentially a regularised variant of Eq. (5.3) that can be used to finesse the over-determined nature of (1) and bias the solution of (2) towards extremal coefficient values. The primary parameter is p, the norm used to constrain α. In principle, the sparsest solution can be quantified using the ℓ_0 pseudo-norm, which counts the non-zero coefficients in α. Unfortunately, its combinatorial nature makes Eq. (5.4) NP-Hard [140].

```
function mp ($x$, $\Phi$)
  $r = x$
  $\alpha = []$
  while $\|r\|_2 > \epsilon$ do
    $i = \text{argmax}_i \langle \varphi_i | r \rangle$
    $\alpha_i = \langle \varphi_i | r \rangle$
    $r = r - \alpha_i \varphi_i$
  end
  return $\alpha$
```

Algorithm 10: Matching Pursuit. Repeated subtraction of the bases most correlated with the residual error. In the classification and regression setting, variations on this algorithm are Least Angle Regression and ℓ_2-Boosting.

State of the art

Briefly, there have been two major thrusts at attacking this problem. Donoho [59] was the first to show that the ℓ_0 and ℓ_1 solutions coincide when $\|\alpha\|_0 < \frac{1+M^{-1}}{2}$, where M is the "coherence" of Φ defined as $\max_{i \neq j} \langle \varphi_i | \varphi_j \rangle$. Using the ℓ_1 norm, it is (somewhat) straight forward to relax this combinatorial optimisation into a quadratic program with linear equality constraints (see e.g. [176, 35]). In the signal processing literature, this method is known as *Basis Pursuit* [36]; in statistical learning, it is called the *Lasso* [172]. Unfortunately, this rigorous approach is prohibitive computationally and scales very poorly, due to the contrived manner in which the problem is recast as a linear program.

The second approach uses heuristic, greedy algorithms to construct a sparse representation *sequentially.* Mallat and Zhang's [120] *Matching Pursuit* algorithm holds a special place in the literature. It is simple, intuitive, and has a rich history within, and outside of, the field [146, 60, 76, 77]. We outline the procedure in Alg. (10): the residual error r is repeatedly stripped of structure correlated with bases until a stopping criterion is satisfied (e.g. number of chosen bases, norm of the residual etc). In regression and classification problems, this approach is known as *Forward Stepwise Regression* and ℓ_2-*Boosting*, respectively. A modern variation on this idea, *Least Angle Regression* [63], avoids overly greedy steps based on $\langle \varphi_i | r_t \rangle$, favouring instead to increase α_i until φ_i is no longer the most correlated with r; at which point a "competing" predictor is introduced into the representation. It is this notion of competition amongst predictors, bases or classifiers that we wish to develop here, albeit without myopic greediness.

5.2 Theoretical Analysis

We now demonstrate formally how the generalised Lotka-Volterra model, together with our ligand binding model, is able to perform approximation of the environment. In order to ease our analysis and make the connection more explicit, we will simplify our model of ligand binding from the matrix-based $\langle\varphi_i|P|\varphi_i\rangle$ to a vectorial $\langle\varphi_i|p\rangle$. This simplification is justified because, in a data-analysis setting, data are typically in vectorial form. To create our matrix "surface representation" in this case, it is natural to use the outer-product $|p\rangle\langle p|$, which correctly satisfies the interpretation of measuring feature correlations. However, $\langle\varphi_i|p\rangle\langle p|\varphi_i\rangle$ visibly reduces to $\langle\varphi_i|p\rangle^2$ and thus, we are simply using the square root of the matrix representation in our simplified analysis. The square root preserves inequalities and thus does not change the optimisation objective[2].

5.2.1 Competition and Approximation

If our basis (or repertoire) matrix Φ were orthonormal, then by definition there would be no competitive effects between clones. The dynamics of ρ smoothly approaches equilibrium where $\rho_i = k_i = \langle\varphi_i|x\rangle$, as would be expected from any orthonormal system. For reasons discussed above, orthonormal bases are not desirable in the applied computational (or biological modelling) setting, but giving up orthogonality forces us to deal with redundancy and dependencies.

The key idea behind our formulation is to exploit a fundamental trade-off that, although originally cast for competing species, can be readily mapped to the regularised optimisation criteria of Eq. (5.4)

- **Maximise capacity.** Growth requires maximising correlation with environmental resources (i.e. capacity). The ecological interpretation is obvious enough; the approximation interpretation is that maximising capacity *minimises* reconstruction error in the approximation. If Φ has an element φ_i that contributes greatly to the approximation, then giving this element as much weight (α_i or ρ_i) as possible improves the approximation.

- **Minimise competition.** Recall that competition is defined in our model as receptor (basis) correlation. For a species of potential high capacity, if many species are competing for similar resources then this capacity will never be reached. However, a species of lesser capacity, that is also under

[2] However, the effect of squaring does encourage more extremal values, which carries additional practical benefits shown later but not relevent to our analysis here.

> less competition, may well prove more successful. In the approximation setting, this translates to redundant, highly correlated clones facing exclusion, driving Φ towards "almost orthogonality". This both increases the sparsity and diversity of a representation. Notice that, in contrast to Donoho's *coherence* (and similar measures in the literature), we need not expect our basis to satisfy "almost orthogonality" *a priori*. Rather, the competition dynamics promote satisfaction in the context of individual approximations, by culling redundant φ_i that fail to garner limited resources.

This intuitive description makes the relationship with the regularised optimisation in Eq. (5.4) clear enough: maximising capacity has the effect of minimising the squared error; minimising competition is effectively similar to minimising the ℓ_p norm. We can more rigorously clarify the approximatory behaviour of the repertoire using our simplified theoretical model, under which the numerators for all species in the dynamical system Eq. (5.1) are written simultaneously as

$$\begin{aligned} k - \sigma &= k - K\rho \qquad (5.5) \\ &= \Phi' x - \Phi'\Phi\rho \\ &= \Phi'(x - \tilde{x}) \end{aligned}$$

which clearly reaches a steady-state when $x = \tilde{x}$ or when the residual error $(x - \tilde{x})$ is in the null-space of Φ'. Notice that the competition vector $\sigma = \Phi'\Phi\rho = \Phi'\tilde{x}$ is implicitly equivalent to clonotype capacity for the approximation $\tilde{x} = \sum \rho_i\varphi_i$. That is, a clonotype is penalised for being more correlated with $\tilde{x}$ than x. Notice also, that Eq. (5.5) is essentially a restatement of the least-squares solution $\rho = (\Phi'\Phi)^{-1}\Phi' x$, stated as $(\Phi'\Phi)\rho = \Phi' x$. Rather than inverting a matrix we are iterating $\Phi'\Phi\rho$. The logistic equation and dynamical system integration further introduces non-linearity and "lag" into Eq. (5.1) and the stable configuration is *not* the least-squares solution. The question is: how does the solution to the dynamical model compare to those found by greedy and global optimisation?

5.2.2 Competition and the Greedy/Global Trade-Off

It should now be clear that we are iteratively solving the approximation problem by integrating a dynamical system that has been crafted to have a sufficiently appropriate steady-state. There is nothing offensively artificial about this crafting; with the right representation, competitive exclusion simply takes over. We now clarify how this strategy relates to greedy iterative methods.

Let $max(x)$ return the index of the maximum component in x, rather than its value. Now observe that in Matching Pursuit (Alg. 10) the index i_1 in the first iteration will be $max(\Phi' x)$, simply because $r_0 = x$. On the second iteration

$$\begin{aligned} i_2 &= max(\Phi' r_1) \\ &= max(\Phi'(x - \langle \varphi_{i_1}|x\rangle \varphi_{i_1})) \\ &= max(\Phi' x - \langle \varphi_{i_1}|x\rangle \Phi' \varphi_{i_1}) \\ &= max(k - \langle \varphi_{i_1}|x\rangle K_{i_1}) \end{aligned}$$

where K_{i_1} refers to the i_1 column of K. In general, we have

$$\begin{aligned} i_{t+1} &= max(k - \sum_{j=1}^{t} \langle \varphi_{i_t}|x\rangle K_{i_t}) \\ &= max(k_t) \end{aligned} \qquad (5.6)$$

where $k_t = k_{t-1} - \langle \varphi_{i_t}|x\rangle K_{i_t}$. What this derivation makes explicit is the implicit role that inter-basis correlation plays in the evolution of the Alg. (10). When a basis φ_{i_t} is selected, those correlated with it suffer a drop in their capacity proportional to their correlation with the signal in the subspace of φ_{i_t}

$$k_{j(t+1)} = k_{j(t)} - \langle \varphi_j|\varphi_{i_t}\rangle \langle \varphi_{i_t}|x\rangle \qquad (5.7)$$

Crucially, notice that we are now dealing solely the same quantities used in Eq. (5.1) – capacity and competition. If we expand Eq. (5.5) as

$$k - K\rho = k - \sum_{j \in \Phi} \rho_j K_j \qquad (5.8)$$

then it becomes clear that, while Alg. (10) greedily sums over the *current* selections weighting by their *maximal* coefficient values (Eq. 5.6), in contrast, competitive exclusion sums over *all* dictionary atoms, weighting by their *current* coefficient values ρ_i. The rest of Eq. (5.1) simply provides an update rule to have $\rho_i \rightarrow k_i$, subject to competitive effects.

So, in contrast to the myopic selective process of Matching Pursuit, Eq. (5.1) uses a more informed eliminatory process, competitive exclusion, while evolving population coefficients. This carries an obvious computational cost above that of greedy approximation, the reality of which will become clear in Sect. 5.4. The only point to be made here is that, if one needs to consider an environment that is changing, rather than a static "signal", then both greedy and global optimisation will require recomputation on a regular basis, whereas the competition dynamics adapt in a timely manner with no additional logic required.

5.2.3 Computational Complexity

Ostensibly, the space complexity of our dynamical system is dominated by $O(n^2)$ for representing an n-dimensional surface environment; and $O(|\Phi|^2)$ for representing the competition matrix K. A saving grace is that, in practice, both of these matrices can be safely assumed to be sparse – that is, most components are zero. The actual space cost will be significantly less given an intelligent sparse matrix implementation, how much so depending on the density of the matrices. In time, k and K can be calculated *a priori* and their cost ameliorated over the entire execution, which is dominated by calculating $\sigma = K\rho$ to be of order $O(|\Phi|^2)$, but again, the sparsity of K will determine the actual cost. This calculation is performed for each time-step of the algorithm until reaching steady-state, which could be a non-trivial multiplicative factor. However, a detail not exploited here is that because our "algorithm" is in fact a dynamical system, the integration till steady-state can (and should) be handled by a standard ODE solver, which will incorporate more sophisticated and highly optimised execution strategies than Euler's method employed here (see e.g. [26]). The potential computational saving here is literally massive. The reason we do not take advantage of this now is the development convenience of controlling the integration method.

The cost of search

For massive Φ, simulating the entire repertoire is not only computationally impractical, but is biologically implausible too. The immune system finesses such physical constraints through a mixture of unbiased random receptor generation in the bone marrow, and biased localised search due to affinity maturation. This translates computationally to allocating a fixed-size repertoire $\tilde{\Phi} \subset \Phi$ where $|\tilde{\Phi}| \ll |\Phi|$ and then exploiting exclusion and random search to develop a sufficiently expressive $\tilde{\Phi}$ over time. This is a natural extension of our model, which is easily integrated into Eq. (5.1) and can draw upon existing research in both AIS and evolutionary dynamics. We will only briefly touch on this in Sect. 5.4.3. Here we present a simple combinatorial analysis of the complexity of such a search for an expressive immune repertoire. The main purpose is to compare this search cost against that of the traditional shape-space. A caveat is that this combinatorial perspective restricts the generality of our notion of receptors as subspaces; but it is a good place to start.

- **Worst-case search**

  Assume that an immunoglobulin can recognise c nearby surface peptides. Thus, we can expect a search space of the order $O\binom{n}{c}$ for the repertoire of the immune system. In the worst case, this scales polynomially in the dimension n, $O(n^c)$. To show this, observe that given $n! < n^c(n-c)!$ it then follows that

  $$\binom{n}{c} = \frac{n!}{c!(n-c)!} \leq \frac{n^c(n-c)!}{c!(n-c)!} = \frac{n^c}{c!} \leq n^c \tag{5.9}$$

  Contrast this with the traditional shape-space, where the search was exponential in the dimension $O(c^n)$, for a different constant c but the same value of n. Note that this worst case bound is only incurred if each peptide is uniformly likely to appear close to another. This does not hold in both the biological and statistical context, because redundancy is rife in meaningful environments. It is this redundancy that makes learning possible [69].

- **Average case search**

  Lacking the inherent parallelism of the biological substrate, generating 10^6 receptors per day in a repertoire of order 10^{12} is not something we can hope to simulate *in silico*. However, it is apparent that the immune system is grossly inefficient insomuch as almost all generated lymphocytes will never bind antigen and will die by apoptosis. Simulating this aspect of the immunology carries no obvious benefit and, unlike the *in vivo* system, we can use information in the surface matrix Q to generate *a priori* competitive immunoglobulin. Treating Q as an adjacency matrix, let us state that each node has, on average, z neighbours. It then follows that to generate a c-sized immunoglobulin requires n choices for the first peptide, and z^{c-1} for the remainder. Thus the complexity of our search is reduced to $O(nz^{c-1})$. It is safe to assume that $z \ll n$. If Q is sparse, z may be very small indeed.

5.3 Dynamic Pursuit for approximation

We dub our competitive exclusion model of approximation *Dynamic Pursuit* in reference to traditional sequential pursuit algorithms and its differentiation as

```
function dp ($x$, $\Phi$)
    $k = \Phi^T x$
    $K = \Phi^T \Phi$
    $\rho = init(x, \Phi)$
    while $\|\rho_{t-1} - \rho_t\|_2 > \epsilon$ do
        for $\varphi_i \in \Phi$ do
            $\rho_i = \rho_i + \Delta \left( \frac{k_i - (K\rho)_i}{k_i} \right) \rho_i$
        end
    end
    return $\rho$
```

Algorithm 11: Dynamic Pursuit. Signal representation is the stable configuration of the clonotype populations, i.e. basis coefficients.

a dynamical system. The pseudo-code is provided in Alg. (11). The only parameters of the algorithm are those used to control the integration by Euler's method: the fixed step-size Δ and the minimum change in population ϵ to detect steady-state. These were set at 0.1 and 0.0001, respectively, as these values gave a reasonable trade-off in numerical accuracy and algorithmic performance.

We also include a number of simple optimisations, not shown in Alg. (11):

- Whenever a sufficient number of clones have been out-competed to extinction, we resize Φ, K and k. This simply avoids redundant calculations that would result in zero values. We perform this operation whenever more than 25% of the population has been excluded since the last resize.

- All species are initialised with the same population magnitude. Ideally, this value would be something intuitive like $|\Phi|^{-1}$, but in practice, a good value depends on the minimum coefficient value that would be accepted as not noise; that is, it depends on properties of the signals being approximated and the basis used for approximation. We finesse this problem during initialisation by scaling the population so that $k_i = \sigma_i$ for the lowest capacity species. That is, the weakest species will be stable in the first iteration, before being out-competed as other species begin to grow. This heuristic simply gets the population to roughly the correct scale. Too small an initial population produces inefficient "burn-in" dynamics of uniform exponential growth until competitive pressure begins to be exerted.

We experimented with various ad-hoc algorithmic modifications that will not be reported here. Empirical improvement was inconclusive and detracted from the elegance of the basic idea. In what follows we will assess our proposed receptor-ligand capacity measure $\langle \varphi_i | X | \varphi_i \rangle$ and the analytical simplification $\langle \varphi_i | x \rangle$.

5.4 Empirical Validation

5.4.1 Protocol

In the following experiments we follow a standard protocol: generate noisy, synthetic signals from a given over-complete "basis" (described below); then approximate each signal using greedy Matching Pursuit, global Basis Pursuit and our proposed competitive exclusion based algorithm, which we dub *Dynamic Pursuit.* For each algorithm, we record the summary statistics (max, min, quartiles, mean and variance) averaged over 100 signals for the metrics

- **Sparsity:** number of non-zeros components $\|\alpha\|_0$.
- **CPU:** time to produce a representation.
- **Reconstruction Error:** $\|x - \Phi\alpha\|_2^2$
- **Synthetic Error:** $\|\beta - \alpha\|_2^2$, described below.

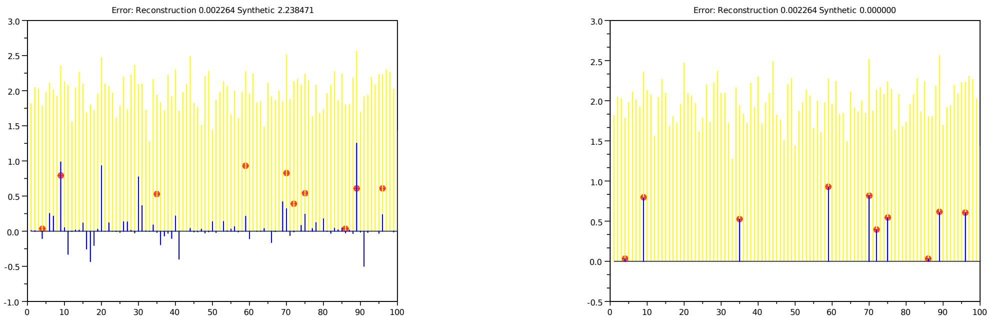
Bases and signal generation

In a typical signal processing application, where these techniques originate, there is a stock collection of (almost) orthonormal bases that can be aggregated to create a non-orthogonal basis, or so-called "dictionary", such as the Fourier basis, Wavelet basis, Haar basis and so on. Thus, non-orthogonality in these cases is rather mild; the result of using more than one orthogonal basis. There are several problems with using these bases in our experiments. First, they do not properly reflect our ligand-binding abstraction. Second, they make extensive use of negative values, in both basis components and coefficients. This can result in either negative populations or negative competition coefficients, both of which are biologically implausible. These issues can be addressed[3] but to minimise ad-hocery we prefer the following procedure

(1) Generate Φ as an $n \times m$ matrix where each entry is set to non-zero with some probability p. Thus, each basis vector has randomly assigned positive entries. Each basis is then normalised so that $\|\varphi_i\|_2 = 1$.

(2) Generate signals x_i as a sparse linear combination of s randomly chosen basis vectors in Φ, with randomly chosen coefficients $\beta_i \in [0, max]$. Each signal is then corrupted with Gaussian noise to add realism.

[3] For example, by doubling the size of Φ to include $-\Phi$ then if φ_i has a negative coefficient, $-\varphi_i$ will have a positive coefficient. To deal with negative competition coefficients in K one might simply set all $K_{ij} < 0$ to 0. But such tricks confound and complicate our study.



(a) Matching Pursuit

(b) Basis Pursuit

Figure 5.1: Illustration of differences in synthetic error for Matching Pursuit and Basis Pursuit, which both achieve the same reconstruction error for this problem. Each bar on the x-axis represents a basis. Light (yellow) background bars represent each bases correlation with a signal. Heavy (blue) foreground bars represent coefficients α selected for approximation. Dots (red) represent actual coefficients β used in the signal generation process.

Note that the generation of Φ in the first step is quite arbitrary, but should have little effect on algorithm performance when approximating signals that have been generated from the same basis used for the approximation procedure. The second step is standard protocol, regardless of how Φ is produced.

Reconstruction and synthetic errors

Squared reconstruction error is the *de facto* metric in these types of experiments. However, reconstruction error is only a proxy measure *implying* that the algorithm has found a good representation. When using synthetic signals it is possible to measure the actual error in representation, that is, the error in selected coefficients and their magnitude. We refer to this as *Synthetic Error*: $\|\beta - \alpha\|_2^2$ where β denotes the coefficients used to generate the synthetic datum. In contrast to pure approximation, this metric can be important when the bases have application-specific meaning and their coefficients are to be interpreted, which is often the case in practice. This issue is illustrated further in Fig. 5.1, where two algorithms, both achieving the same reconstruction error, have produced representations with quite different synthetic errors, one clearly superior to the other.

5.4.2 Comparison with state of the art

In Figure 5.2 we graph the performance of the algorithms in approximating 100-dimensional signals, each sparsely generated from 10 bases selected at random from a 1000 element basis. We include two variants of our algorithm: *dp* represents the simplified model used in our theoretical analysis; *dp*2 uses the outer-product matrix representation of signals $|x\rangle\langle x|$. It can be seen that the former is more accurate, but slower and denser. The latter is faster and sparser. Notice that the difference between variants in terms of synthetic error is negligible. The density of representation in *dp* can be explained by retaining many low coefficient (population) bases (clonotypes). In contrast, the extremising effect of squaring on capacity makes it harder for low population clones to survive. This is beneficial, because these low population clones are modelling the Gaussian noise, not the underlying structure of the signal.

In terms of CPU efficiency, integrating dynamics (*dp*) will always be computationally more demanding than greedy approximation (*mp*), though the difference is not as large as one might expect; even when using the inefficient Euler's method of integration. More importantly, it is significantly more efficient than performing optimisation by linear programming. Basis Pursuit's CPU time was over 200 seconds and is well outside the bounds of this graph. It is also interesting to note that in all other respects *dp* performs similarly to *bp*, but at a fraction of the computational effort in both time and space.

Matching Pursuit achieves the lowest reconstruction error, but it does so at a significant cost to synthetic error and sparsity (recall Fig. 5.1(a)). Matching Pursuit, Basis Pursuit and our own algorithm without quadratic capacity all significantly *under*-estimate the true sparsity (i.e. 10), employing between 60 and 90 bases. In contrast, the quadratic capacity version of our algorithm is able to drive the sparsity down to around 20. The cost here is a notable increase in reconstruction error which, as we explained above, is in part caused by failing to model the Gaussian noise. The slight variance in synthetic error seems acceptable.

5.4.3 Comparison with a constrained repertoire

We now compare the basic dynamic pursuit implementation, which integrates the entire repertoire, to one that maintains a fixed-size repertoire $\tilde{\Phi} \subset \Phi$. The goal is to ensure that our results do not depend on the rather unrealistic assumption that the entire repertoire is available and initialised with uniform population. Implementing the affinity maturation aspect of clonal selection would require exploring and justifying somewhat arbitrary decisions on mutation and local search strategies. Instead we simply extend Alg. (11) so that clones in $\tilde{\Phi}$ with negligible population (i.e. coefficients) are replaced by random samples from Φ. We explore two sampling strategies

1. **With Replacement.** Samples are selected uniformly at random from the columns of Φ. The same clonotype may be sampled more than once.
2. **Without Replacement.** Samples are selected uniformly at random from the columns of Φ but the same clonotype may not be resampled.

In Figure 5.3 we demonstrate the effect of sampling on the same metrics and signals used in Sect. 5.4.2 when using a fixed-size $|\tilde{\Phi}| = 100$ for $|\Phi| = \{100, 1000, 10000\}$. It is apparent that any negative effects are contained to sampling with replacement (*with*), which in the case of $|\Phi| = 10000$ takes longer to reach steady-state and suffers notable variance in *sparsity*. Sampling with replacement is computationally easier to achieve, but would seem less biologically plausible. The general robustness of sampling can be explained because Φ is *redundant*. The algorithm has no preference for which φ_i makes it into a representation, other than it be fit enough to compete; the population dynamics then adjusts the representation as necessary. One would expect any additional local search via "mutation" to improve on the results presented here.

5.5 Conclusion

We have introduced a more biologically plausible abstraction of ligand binding and shown how this abstraction, coupled with ecological competition dynamics, is effective: as a qualitative model of the clonal selection; as a quantifiable interpretation of constructive representation learning; and as an applied method for solving sparse approximation problems. The fundamental conceptual shift was to understand the immune repertoire as performing approximation of the environment via an additive expansion of basis (receptors) and their coefficients (clonotype population). Once properly formulated, the competition dynamics cannot "help", so to speak, but to perform such an approximation. This competition dynamic addresses many of the criticisms raised against existing clonal selection algorithms in earlier chapters: where the notion of clonotype is absent; emphasis is on *ad hoc* selection rather than principled exclusion; interactions are defined based on implausible immune network metaphors; and the model of receptor-ligand interactions is based on the traditional metric shape-space. In the next chapter, we extend this basic lymphocyte ecology towards immunological models of self/non-self discrimination.

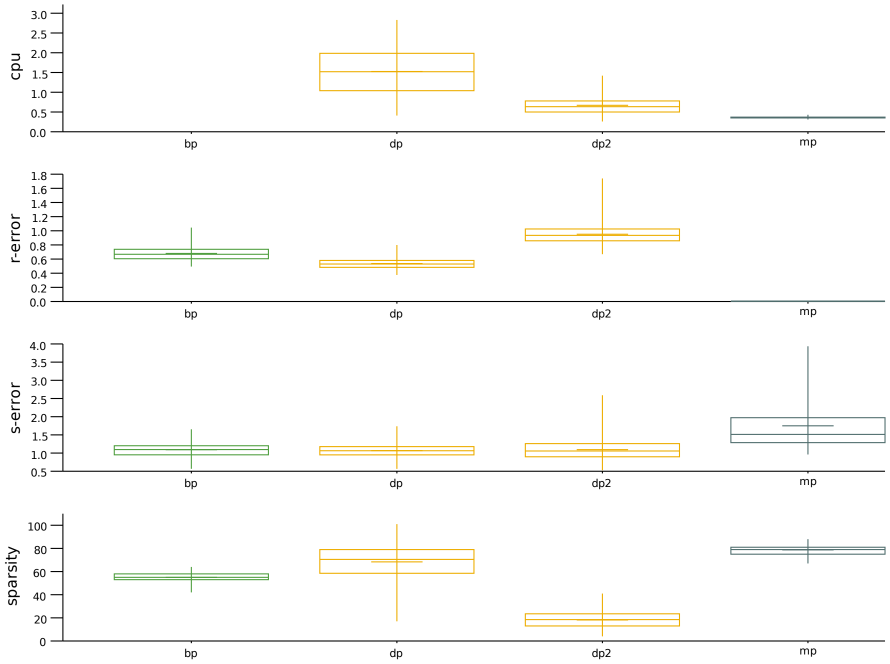


Figure 5.2: Approximation results for greedy Matching Pursuit (*mp*), global Basis Pursuit (*bp*) and two variants of our competitive exclusion algorithm (*dp* and *dp*2) that represent our simplified theoretical model and actual proposed model, respectively. See text for discussion.

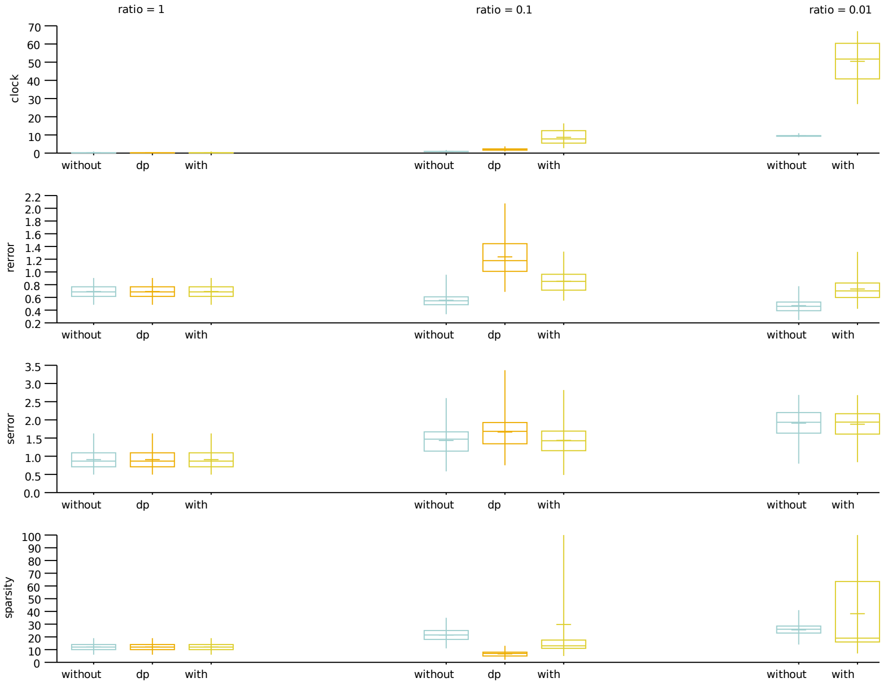


Figure 5.3: Degradation in *dp* performance when it is constrained to use a fixed-size repertoire that is only a subset of the basis used to generate the data. As basis coefficients become negligible they are replaced with random samples from the full basis, both *with* and *without* replacement. The *ratio* refers to the fixed size of the repertoire (100) divided by the full size of Φ.

Chapter 6

A Systemic Two-signal Model of Decision Making

We may find a more reasonable analogy between language and the immune system, by regarding a given antibody not as a word; but as a sentence or phrase.
N. K. Jerne

The meaning of a word is its use in the language.
Wittgenstein

Having developed the necessary methods to describe representation learning in the immune system as an approximation problem, we now extend this approximation framework to encompass *decision making*. The essential difference is that we are now approximating an unknown function of our signal $\theta = f(x)$, rather than approximating x itself. Our derivation for the immunological form of $f(\cdot)$ may offer some insight into the role and sufficiency of immunological components.

6.1 Lymphocyte Ecology II

Although the ecological dynamics of the previous chapter are plausible with respect to clonal selection, clonal selection does not in itself explain self/non-self discrimination – indeed, it cannot because it makes no semantic distinction between different antigen. Of the remaining theories in Chapter 2, there is one particular dichotomy we hope to offer some relief from. With the exception of Janeway and Matzinger's theories, none of the models have clear semantics for what would make up a self or non-self response. However, both Janeway and

Matzinger's theories rely on the notion that an evolutionarily ancient, germline encoded immune system is responsible for self/non-self discrimination and the "switch" for the adaptive response. Although Carneiro et al. assert the importance of holistic dynamics over a reductionist switch, the dynamics of the cross-regulation model are once again ambiguous to the semantics of the self/non-self distinction. In principle, we are seeking *both* proper semantics and systemic "switch free" effects. In practice, the numerical methods underlying statistical inference can provide us with that.

6.1.1 Revisiting the alpha and omega

Recall, in Chapter 3 we derived a dual-like relationship between the nearest neighbour and linear regression models of learning. We will now reverse that derivation, introduce irreversible non-linearity and move away from locality-based shape-space models of immune decision making.

Recall also, the form of the linear classifier. Given a column matrix X of observations and an accompanying vector θ of class labels $\theta_i \in [-1, 1]$, we have

$$\begin{aligned} f &= \operatorname{argmin} \|\theta - X'f\|_2^2 \\ &= (X^+)'\theta \\ &= (XX')_k^{-1}X\theta \\ &= G_k^{-1}\tilde{f} \end{aligned} \tag{6.1}$$

where we have collected the terms $G = (XX') = \sum_i |x_i\rangle\langle x_i|$ and $\tilde{f} = X\theta = \sum_i \theta_i x_i$ for presentational convenience. Now, given an unobserved $\hat{x}$ the linear decision function predicts

$$\begin{aligned} \hat{\theta} &= f(\hat{x}) \\ &= \langle \hat{x}|f\rangle \\ &= \left\langle \hat{x}|G_k^{-1}|\tilde{f}\right\rangle \end{aligned} \tag{6.2}$$

with the decision boundary lying orthogonal to f, i.e. where $\langle \hat{x}|f\rangle = 0$. From a numerical computing perspective, the major transformational effect and computational cost is the inversion of G, which we turn to now.

The spectral theorem

The famous Spectral Theorem [12] states that G can be decomposed[1] as a *superposition* of basis vectors

$$G = \Phi\Lambda\Phi' = \sum_i \lambda_i |\varphi_i\rangle\langle\varphi_i| \tag{6.3}$$

where the basis vectors φ_i are the *eigenvectors* of G and the λ_i coefficients are their accompanying *eigenvalues*. The significance of the eigen-decomposition is that it is the only factorisation that diagonalises Λ, in effect, decoupling all of the factors. This occurs because, in addition to being an orthonormal basis, the eigenvectors are the *invariant subspaces* of the vector space spanned by G – that is, multiplication $G\varphi_i$ does not change the direction φ_i points in. This somewhat magical property is exploited throughout applied mathematics for a multitude of reasons. The relevant reason here is that factorising G greatly simplifies mathematical operations on G such that for some functions $f(G) = \Phi f(\Lambda)\Phi'$ and Λ behaves algebraically similar to a scalar, because it is a diagonal matrix.

One such function is inversion, that is $G_k^{-1} = \Phi\Lambda^{-1}\Phi'$. Inversion of a diagonal matrix (unlike any other type of matrix) is simply the scalar inversion of its diagonal components, thus

$$G_k^{-1} = \Phi\Lambda^{-1}\Phi' = \sum_i^k \frac{1}{\lambda_i}|\varphi_i\rangle\langle\varphi_i| \tag{6.4}$$

Now, substituting Eq. (6.4) into Eq. (6.2) we can derive a mathematically equivalent, but semantically quite different, interpretation of the linear classifier based on its numerical, rather than algebraic, solution

$$\begin{aligned}
\hat{\theta} &= \left\langle \hat{x}|G_k^{-1}|\tilde{f}\right\rangle \\
&= \left\langle \hat{x}|\sum_i \frac{1}{\lambda_i}|\varphi_i\rangle\langle\varphi_i||\tilde{f}\right\rangle \\
&= \sum_i \frac{1}{\lambda_i}\langle\hat{x}|\varphi_i\rangle\left\langle\varphi_i|\tilde{f}\right\rangle \\
&= \sum_i \frac{\langle\hat{x}|\varphi_i\rangle\left\langle\varphi_i|\tilde{f}\right\rangle}{\langle\varphi_i|G|\varphi_i\rangle}
\end{aligned} \tag{6.5}$$

[1] In more detail, G must satisfy certain conditions to be diagonalisable. In our case these conditions are immediately satisfied by virtue of G being the product of a positive matrix X with its transpose X', thus rendering G both non-negative $G_{i,j} \geq 0$ and symmetric $G_{i,j} = G_{j,i}$.

where $\lambda_i = \langle \varphi_i | G | \varphi_i \rangle$. Equation (6.5) shows us that the classification decision of the linear classifier is the integration of *three key measures of information about each basis vector*; each of which is a coefficient of correlation or approximatory capacity. Notice that $\langle \varphi_i | G | \varphi_i \rangle$ appears in the denominator, thus eigenvectors with large eigenvalues – the most important in terms of reconstructing X and G – are weighted *less* in the classification decision. Statisticians call this inverse relationship between representational power and discriminatory power *precision.*

6.1.2 A model of systemic response

The conceptual leap from a statistical model of linear functional relationships to a non-linear dynamical model of an immune response now rests upon two very simple ideas:

1. **A change of basis.** For the same reasons discussed in the previous chapter, an orthonormal basis is only desirable from a platonic mathematical point of view. For solving computational problems in applied mathematics, a redundant overcomplete "basis" can be highly desirable. Thus, we reinterpret the φ_i in Eq. (6.5) as arbitrary basis vectors or functions. Recall, *only* the eigenvectors satisfy Eq. (6.4) but, using an argument similar to that for justifying nearest-neighbour decisions (Chapter 3), we assert that this technical omission can be ignored if the benefits of expanding Φ offset any costs in inaccurately approximating G^{-1}. For classification, this is entirely plausible as approximation errors do not necessarily imply classification errors – e.g. only the sign of the decision needs to be correct.

2. **Resolve dependencies.** The consequence of no-longer using an orthonormal Φ is that bases become dependent. We solve this using exactly the same technique employed in the last chapter. With a slight abuse of notation, let each $\langle \cdot \rangle$ in Eq. (6.5) be *upper-bounded* by the value of the inner-product. That is, this upper-bound is the *capacity* of φ_i. The actual value $\langle \cdot \rangle$ takes will be the equilibrium population emerging from competition dynamics.

Using the notation $\varphi_i(x)$ to represent the steady-state population of species i in approximating the signal x, we now rewrite Eq. (6.5) as

$$\begin{aligned} \hat{\theta} &= f(\hat{x}) \\ &= \sum_{\varphi_i \in \Phi} \frac{\varphi_i(\hat{x}) \varphi_i(\tilde{f})}{\varphi_i(G)} \end{aligned} \tag{6.6}$$

which, as in Eq. (6.5), states that the immune response is the integration across the repertoire of three key pieces of information about the fitness of each clonotype in competing for (i.e. approximating) three different environmental resources.

Definition 6. *We refer to $\hat{x}$ as the **Target** as it is the object that the response is being directed against. Notice that it is an arbitrary compound object – not an antigen or epitope – best described, again, by a surface representation. Generalising the vector $\hat{x}$ to $|\hat{x}\rangle\langle\hat{x}|$ or some other matrix $\hat{X}$ is easily done.*

Definition 7. *As in Chapter 5, we refer to $G = \sum_i |x_i\rangle\langle x_i|$ as the **Environment**, that is, the sum of protein surface descriptions. Notice that Eq. (6.5) and (6.6) make no explicit reference to individual observation vectors x_i. A more general matrix or graph-based surface description may be introduced.*

Definition 8. *We refer to $\tilde{f}$ as the **Context** because the components of $\tilde{f} = X\theta = \sum_i \theta_i x_i$ represent the bias of each dimension towards a positive or negative response. Again, there is no explicit reference in Eq. (6.5) and (6.6) to observations x_i and their labels θ_i. That is, the "learning from examples" protocol is only implied. To avoid negative quantities, it will be convenient to expand $\tilde{f}$ as*

$$\tilde{f} = \left(\sum_{\theta_i>0} x_i\right) - \left(\sum_{\theta_j<0} x_j\right) = \tilde{f}^{+} - \tilde{f}^{-} \tag{6.7}$$

though we will use $\tilde{f}$ notation when the distinction is not important.

Given these definitions, then under Eq. 6.6 the response $\hat{\theta}$ is the integration of individual clonotype responses; where each response is a function of competitiveness in garnering resources from the *environment*, the *target* and the *context*. Following the immunological models of Chapter 2 we will refer to both $\varphi_i(\hat{x})$ and $\varphi_i(G)$ as *signal one* because this is, quite precisely, what they represent: a clonotypes ability to garner binding sites on surfaces of the target and in the general antigenic environment, respectively. The context $\varphi_i(\tilde{f})$ requires some additional justification, which we provide now.

6.1.3 A *two-signal* systemic response

A fundamental property of the two-signal models in Chapter 2 is that they all involve feedback based on either

- The presence or absence of activated T-Helper cells
- The ratio of T-Effector to T-Regulatory cells

- The presence of Danger or PAMP "signals" during antigen presentation

Notice that, regardless of the proposed mechanism, each of these second signals is based on (fragments of) peptides – the unit of interaction for T-cells and antigen presenting cells. Thus, *signal two* in either of its forms is directed at peptides, not epitopes, which is ultimately what antibody bind to. Now, because our representational abstraction makes the distinction between peptides and epitopes, we are able to represent this different granularity of feedback.

In the definition of *context* above, we have already shown that the statistical role of $\tilde{f}$ is to encode the correlation of X and θ: each $\tilde{f}_i$ represents the bias of the i'th component towards one response or the other. In our systemic model, each $\tilde{f}_i$ represents a particular peptide, so *context* is functionally equivalent to all of the above immunological descriptions. Because of the structural similarity of this aspect of the immunological models, we have some freedom in how to interpret $\tilde{f}$. Each $\tilde{f}_i$ may represent a T-cell clonotype that favours one response over another, e.g. activated or anergised, effector or regulator etc. Similarly, each $\tilde{f}_i$ may represent the sum of antigen presenting cell profiles with the presence or absence of co-stimulation derived from danger signals, pathogen associated molecular patterns, and so on. Regardless of the interpretation, assuming these effects are additive the net effect will be the same – $\tilde{f}$ – and the ability of induced B-cells to garner *signal two* is encoded in $\varphi_i(\tilde{f})$. There are some caveats:

1. We are assuming that B-cells only receive a second signal from T-cells that bind to the same peptides that make up a B-cell receptor epitope. This is of course a simplification of the biology: B-cells interact with T-cells by presenting peptides derived from matter endocytosed during binding [164]. This would presumably include cognate peptides, but not exclusively.

2. We are ignoring that T-cell receptors are themselves highly degenerate. This *"one T-cell per peptide"* model is again a simplification of the underlying biology. More elaborate T-cell representations could readily be incorporated into this basic model, but we will not do so here.

3. For completeness, by defining $\theta \in [-1, +1]$ we are assuming the presence of pro- and anti-response forces. This matches the theoretical immunology but it would also be possible to consider $\theta \in [0, 1]$.

To a first approximation, the biological simplifications would seem acceptable. Given that we are not aware of *any* model in the literature that includes a multi-level response with T-cell/B-cell ligand distinction, this basic abstraction may still yield insights that immunological sophistication can be built upon.

6.2 The Statistical Immune Response

The fundamental concept behind our formulation of systemic decision making was to recast it as another form of approximation. Equation (6.6) has allowed us to abstract quite a lot from the immunology, but we had to trade off a closed form mathematical solution to the decision-making problem for an approximation with no real guarantees about solution quality. Here we try to strengthen Eq. (6.6) by revisiting the boosting methodology of Chapter 3. We first motivate this with a statistical analysis that may also offer immunological insight into the roles of the innate "reductionist switch" and the adaptive "holistic dynamics" of Chapter 2.

6.2.1 The role of Danger as "switch"

It follows quite directly from Eq. (6.6) that decisions at the peptide level may not be, in a statistical sense, sufficiently discriminatory. Consider some bio-chemical compound structure, of unknown "self/nonself-ness", statistically described as a vector x. That is, x_i quantifies the amount of i'th peptide that could be scavenged and appear on the surface of antigen presenting cells. For any given *context*, a peptide-specific response to this structure would be the integration of the bias of peptides in this structure towards pro- or anti-response, e.g.

$$\hat{\theta} = f(x) = \sum_i g(x_i) g(\tilde{f}_i) \approx \sum_i x_i \tilde{f}_i \tag{6.8}$$

where $g(\cdot)$ represents the complex extra-cellular and intra-cellular processes that result in peptides being phagocytosed and presenting on cell surfaces. Assuming this is roughly the same for all peptides, the approximation in Eq. (6.8) seems reasonable[2]. Statistically speaking, this produces a very particular kind of linear decision boundary (Fig. 6.1) with the following properties

- The classes *self* and *non-self* should be separable orthogonally to the mean observation. That is, the decision boundary lies orthogonal to the difference vector $\tilde{f} = \tilde{f}^{+} - \tilde{f}^{-}$ and self/non-self discrimination requires that classes lie on opposite sides of this boundary. This only occurs if a pathogen's peptides occur more than average in the context of danger *and* less than average in the context of non-danger, or vice-versa.

[2] In fact, this is not the same for all peptides (see e.g. [50]) but there is no semantic distinction in the differences of peptide egression rates, so our approximation still seems valid.

- Peptides that occur more than (or less than) average in the context danger *and* non-danger are not discriminatory unless the decision boundary can be translated from the origin. This is equivalent to a decision boundary where $\langle f|\cdot\rangle = \gamma$ for some threshold γ.

- The classes *self* and *non-self* should have similar variance characteristics. This is simply because there is no way to adjust for variance using only $\tilde{f}$.

These are fairly mundane statistical criteria. More subtle issues arise in determining how the immune system might meet them. For example, the assumption that the response must cross a threshold γ seems innocuous enough. But how the immune system could *learn* that threshold is not clear. This is a function of the distributions of *self* and *non-self* and thus the immunological mechanisms of Chapter 2 are moot on this point. The assumption of equal class variance is, of course, questionable; but a more pressing issue is that Eq. (6.8) implicitly requires the class distribution of *self* and *non-self* to be equal – which is certainly false. The issue here is the effects of class skew, which effects the decision slightly differently whether $\tilde{f} = \tilde{f}^{+} - \tilde{f}^{-}$ is interpreted as the difference in class-conditional *means*, or the difference in class-conditional *sums*, of the observations. Skew in the class distribution will bias class-conditional sums towards the majority class and $\hat{\theta}$ towards a majority class constant decision. In contrast, the class conditional means may be skewed in the other direction, because the minority class will have a smaller denominator.

Note that these are not just technical arguments: any transformation $\tilde{f} \to f^{*}$ requires additional information – such as peptide-peptide correlations – that may not be plausibly available to T-cells, antigen presenting cells, or any other *peptide-specific* component.

6.2.2 The role of systemic response

The systemic response model in Eq. (6.6) incorporates the danger-based "switch" element, but does not allow it to fully dictate the response. We now consider what our statistical perspective suggests the adaptive immune system contributes beyond the simpler and evolutionarily ancient innate system.

Sub-optimal decisions in more expressive spaces

The most obvious improvement is representational. Rather than represent biochemical compound structures of unknown "self/nonself-ness" in *peptide-space*, the *receptor-space* representation is (*i*) a projection into Φ space, where $|\Phi|$ may

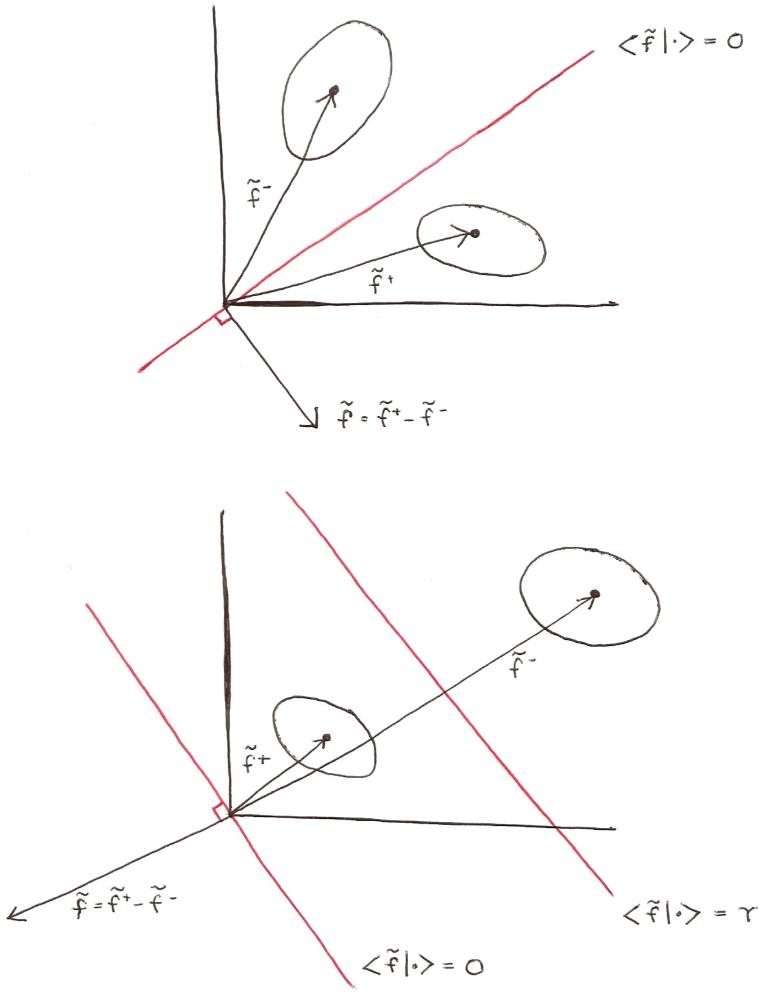

$\langle \tilde{f} | \cdot \rangle = 0$
$\tilde{f}^-$
$\tilde{f}^+$
$\tilde{f} = \tilde{f}^+ - \tilde{f}^-$
$\tilde{f}^-$
$\tilde{f}^+$
$\tilde{f} = \tilde{f}^+ - \tilde{f}^-$
$\langle \tilde{f} | \cdot \rangle = \gamma$
$\langle \tilde{f} | \cdot \rangle = 0$

Figure 6.1: The geometry of linear decision functions. The decision boundary (red) produced by a response such as Eq. (6.8) only separates classes that occur in a very particular configuration (top). Other configurations (bottom) can be finessed by the introduction of a threshold γ, but the optimal value for this threshold must be learnt from the data and depends on class variance.

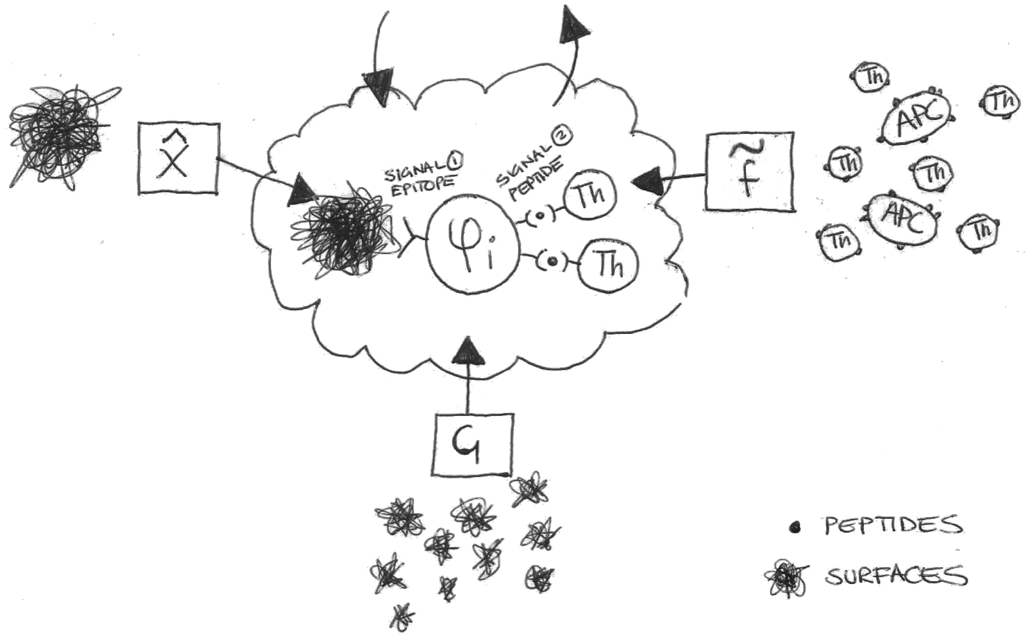


Figure 6.2: A conceptual diagram illustrating how the mathematical quantities in Eq. (6.6) and immunological components of Chapter 2 relate to each other. B-cell clonotypes φ_i compete over epitope binding sites G and $\hat{x}$ and a second peptide-specific second signal derived from T-cell/APC interactions $\tilde{f}$. As in the statistical setting, $\tilde{f}$ encodes how individual peptides (i.e features) bias the response towards a particular decision. However, the systemic response is epitope driven. This provides a larger, synthetic feature space where the competition dynamics provide non-linearity and subsequent feature reduction by exclusion. The mathematical and biological realisations of $\tilde{f}$, G and $\hat{x}$ are quite flexible, with the exception that the former is a peptide-specific signal and the latter are epitope-specific signals based on peptide surface correlations.

be very large indeed, and (*ii*) the subsequent reduction in the representation as a result of competitive exclusion dynamics. Thus, the systemic response both expands the dimensionality by feature *generation* from the original representation; and performs a reduction in dimensionality by feature *selection* on the new representation. The competition dynamics make this transformation non-linear and "adaptive" in the sense that a representation is approximant specific[3]. Alone, this change of representation could, in principle, be sufficient to allow even a sub-optimal linear decision in receptor-space better linear decisions in peptide-space[4].

The interpretation of context: precision and tolerance

Even if one accepts the mapping between statistical and immunological components, it remains to assert why the systemic response would take the mathematical form of Eq. (6.6). The answer is straight forward and best understood by rearranging the equation

$$\hat{\theta} = \sum_{\varphi_i \in \Phi} \frac{\varphi_i(\tilde{f})}{\varphi_i(G)} \varphi_i(\hat{x}) = \sum_{\varphi_i \in \Phi} \alpha_i \varphi_i(\hat{x}) \tag{6.9}$$

That is, a clone's response is proportional to its competitiveness for the target. This proportion α_i is the distribution of *signal two* amongst those members that have achieved *signal one.* Although a simplification of the underlying biology, this is an intuitive statement that follows logically from the two-signal models. Understanding how this ratio can be realised gives some insight into the statistical and immunological effects of Eq. (6.6). Notice that a receptor's contribution to the response is no longer in absolute value. An ostensibly small difference in $\varphi_i(\tilde{f}^+) - \varphi_i(\tilde{f}^-)$ will be weighted more if $\varphi_i(G)$ is also small. Likewise, ostensibly large differences will be weighted less if $\varphi_i(G)$ is also very large. Further, ostensibly low differences with large $\varphi_i(G)$ will be penalised: proliferate epitopes with little discriminatory power contribute less to the response, even if the difference in class-bias is relatively large in absolute terms. The final permutation is where $\varphi_i(\tilde{f}^+) - \varphi_i(\tilde{f}^-)$ is large and $\varphi_i(G)$ is small, suggesting a clonotype that has much more available *signal two* than *signal one.* This may represent clonotypes that are highly effective latecomers to a response, or perhaps low population resting memory clones. Such a contribution to the response would be weighted more. Figure 6.3 illustrates the presence of this effect in a simulated response.

[3] This is the mathematical, rather than biological, meaning of the word "adaptive".

[4] The efficacy of this would depend on how clonotype capacity was defined as a function of epitope representation. In practice, the algebraic formulation introduced here is formally convenient but unlikely to be *that* effective, but it is a good starting point for developing immunological and mathematical sophistication.

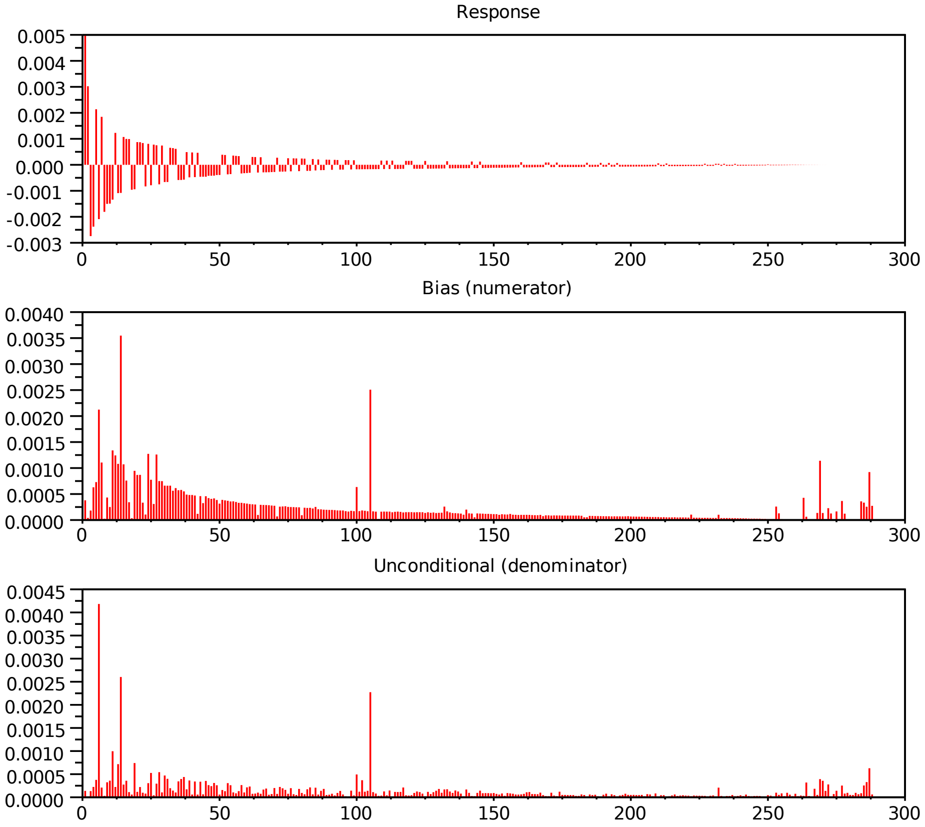


Figure 6.3: An empirical *target* response from a repertoire of 300 clones. Clones are ranked by magnitude of response in each graph. **Top:** The response of each clone. **Middle:** The absolute bias of each clone: abs $\left(\varphi_i(\tilde{f^+}) - \varphi_i(\tilde{f^-})\right)$. **Bottom:** The unconditional fitness of each clone $\varphi_i(G)$. The effect of this appearing in the denominator of Eq. (6.6) is that high magnitude responses are not necessarily the product of clones with the greatest concentration or bias.

It is insightful to briefly consider the other possible rearrangement of Eq. (6.6)

$$\hat{\theta} = \sum_{\varphi_i \in \Phi} \frac{\varphi_i(\hat{x})}{\varphi_i(G)} \varphi_i(\tilde{f}) = \sum_{\varphi_i \in \Phi} \alpha_i \varphi_i(\tilde{f}) \tag{6.10}$$

Regardless of a clonotypes ability to garner *signal two* its contribution to a response can be radically reweighted depending on the ratio of *target* to *environment* competitiveness for *signal one*. Again, high *environment* competitiveness leads to a lower weighted response: this is the inverse relationship between representational power and discriminatory power alluded to earlier. Equation (6.10) suggests that *high target* competitiveness and *low environment* competitiveness maximises the response weight of a clonotype. That is, highly specific responses to less abundant epitopes.

A corollary to both of these interpretations is that, because $\varphi_i(G)$ appears in the denominator, the sheer abundance of self-epitopes provides its own *tolerising* effect. Even if for some self-epitopes, the context wrongly suggests $p(\text{non-self}) > p(\text{self})$, which might be expected to occur transiently during a response, the abundance of $\varphi_i(G)$ epitopes will keep the weight of associated clonotypes responses low. This is by no means an explanation of tolerance, but of all the assumptions that could be made about self-epitopes, that they are abundant would seem the least questionable. An immune system that *habituates*, in a sense, to the continual presentation of self, even when presented in the context of danger, offers a less assumptive alternative to Matzinger's explanation of why danger does not have the side-effect of overt autoimmunity (e.g. [127]).

The strength of any particular clonotype response is a trade off between pathogen specificity, epitope abundance and discriminatory bias. The fundamental mechanism that induces this trade off is the distribution of *signal two* amongst cells achieving *signal one*. This is biologically plausible, immunologically functional and, though not statistically optimal, is statistically sound. This leads us to the question that opened this section: is this enough to ensure "correct" responses?

6.2.3 "Boosting" the immune system

Recall, that both Matching Pursuit and Boosting algorithms iteratively fit "basis functions" to a residual vector (See Alg. 12 for a side-by-side comparison). For Matching Pursuit, the residual is the reconstruction error in representing the observation x. For ℓ_2 Boosting, the residual is the loss in representing the decision surface θ. Recall also, that the "trick" that allowed us to demonstrate the ap-

```
$r = x$
$\alpha = []$
while $\|r\|_2 > \epsilon$ do
    $i = \text{argmax}_i \langle \varphi_i | r \rangle$
    $\alpha_i = \langle \varphi_i | r \rangle$
    $r = r - \alpha_i \varphi_i$
end
```

```
$\mathbb{R} = \theta$
$F = \emptyset$
for $t = 1 \dots T$ do
    $f_t = \text{argmin}_{f \in \mathcal{F}} \|\mathbb{R} - f(X;\mathbb{R})\|_2^2$
    $F = F + f_t$
    $\mathbb{R} = \mathbb{R} - f_t(X)$
end
```

Algorithm 12: A comparison of the algorithms Matching Pursuit (left) from Chapter 5 and ℓ_2 Boosting (right) from Chapter 3. The underlying connection between strategies is made particularly clear if one recalls that $\text{argmin}\, \|\mathbb{R} - f_t\|_2^2 = \text{argmin}\, \|\mathbb{R}\|_2^2 + \|f_t\|_2^2 - 2\,\langle \mathbb{R} | f_t \rangle \approx \text{argmax}\ \langle \mathbb{R} | f_t \rangle$

proximatory capacity of competitive exclusion relied on using, what we will call, an *implicit residual*

$$k - K\rho = \Phi x - \Phi\Phi'\rho = \Phi\left(x - \Phi\rho\right) = \Phi\left(x - \tilde{x}\right) \tag{6.11}$$

In principle, this same trick can be applied where one would rather minimise the residual $\theta - \hat{\theta}$. The most explicit method of casting boosting in the competitive exclusion framework would be redefining capacity and competition as

$$\frac{d\rho_i}{dt} = \left(\frac{\left\langle \theta | \hat{\theta}_{\varphi_i} \right\rangle - \sum_j \left\langle \hat{\theta}_{\varphi_i} | \hat{\theta}_{\varphi_j} \right\rangle \rho_j}{\left\langle \theta | \hat{\theta}_{\varphi_i} \right\rangle} \right) \rho_i \tag{6.12}$$

where $\hat{\theta}_{\varphi_i}$ are the decisions of an arbitrary weak learner φ_i. Thus, diversity is induced by having learners with similar decision vectors suffer competition, and accuracy is improved by favouring high capacity learners most correlated with the ground-truth θ. To the best of our knowledge, Eq. (6.12) is an entirely novel take on Friedman et al's gradient boosting [79, 21, 122], where the stagewise gradient descent is replaced by competitive exclusion. Notice also that the inner-products can easily be substituted with arbitrary *utility* (rather than *loss*) functions.

Unfortunately, it does not seem biologically plausible to assume clonotypes retain a record of previous performance on each ligand and have direct access to the ground-truth θ. We need something more subtle. Taking our lead from ℓ_2 boosting, an implicit residual formulation might attempt to minimise $X(\theta - \hat{\theta})$, which is the steady-state of the following competition dynamics:

$$\begin{aligned} k - K\rho &= \tilde{f} - G\rho \qquad (6.13) \\ &= X\left(\theta - X'\rho\right) \end{aligned}$$

that is, $\rho \approx G^{-1}\tilde{f}$ and $\hat{\theta} = X'\rho$. But the population now consists of n species, where n is the number of features. Equation (6.13) represents a competition dynamic among peptides or peptide-specific components such as T-cells. It is certainly plausible that T-cells undergo competitive exclusion, though why competition amongst T-cells (or any other peptide-specific component) should be quantified by peptide surface correlations in G is not obvious[5].

A more pressing statistical issue is that the steady-state of Eq. (6.13) is a linear decision boundary in peptide-space. But we have already solved the non-linear representation learning problem. Replacing the matrix X with the matrix X_Φ of repertoire representations $X_\Phi = [\Phi(x_1), \ldots, \Phi(x_m)]$ gives

$$\begin{aligned} k - K\rho &= \tilde{f}_\Phi - G_\Phi \rho \qquad (6.14) \\ &= X_\Phi \left(\theta - X_\Phi \rho\right) \end{aligned}$$

which is still linear in the features – but the number of features is now $|\Phi|$ and these features are not linear transformations $\langle \varphi_i | x \rangle$, but the steady-state of competition dynamics. Undoing the non-linearity for the moment, observe that

$$f_\Phi = \sum_i \theta_i \Phi\left(x_i\right) \approx \sum_i \theta_i \left(\Phi' x_i\right) = \Phi' \left(\sum \theta_i x_i\right) = \Phi' \tilde{f} \qquad (6.15)$$

is the *capacity* vector k in $\varphi(\tilde{f})$ of Eq. (6.6), and similarly

$$\left(G_\Phi\right)_{ij} = \left(\Phi(X)\Phi(X)'\right)_{ij} \approx \left(\left(\Phi' X\right)\left(\Phi' X\right)'\right)_{ij} = \left(\Phi' G \Phi\right)_{ij} = \langle \varphi_i | G | \varphi_j \rangle \qquad (6.16)$$

is a minor variant of the *competition* matrix K shared by all $\varphi(\cdot)$ in Eq. (6.6)[6]. Minor variations aside, the point is that in the linear setting this implicit formulation of boosting is competitive exclusion over *context*

$$\begin{aligned} k_i - (K\rho)_i &= (\tilde{f}_\Phi)_i - (G_\Phi \rho)_i \qquad (6.17) \\ &= \left\langle \varphi_i | \tilde{f} \right\rangle - \sum \langle \varphi_i | G | \varphi_j \rangle \rho_j \end{aligned}$$

[5] Competition for APC binding sites, perhaps? Assuming that surface correlated peptides are likely to be presented as fragments on the same APC, this might be plausible.

[6] In fact, this may be a superior measure of clonotype competition than $\langle \varphi_i | \varphi_j \rangle$. Notice $\langle a | G | b \rangle = \sum_i \sum_j a_i b_j G_{ij}$. It follows that a and b may be considered orthogonal with respect to G, $\langle a | G | b \rangle = 0$, even if $\langle a | b \rangle > 0$. Orthogonality in this context is called "conjugacy" and this better abstracts that the φ_i only interact indirectly. If there is no resource to compete over, e.g. $G_{ij} = 0$, then overlapping receptors are not competing, even if $\langle \varphi_i | \varphi_j \rangle > 0$.

which is already incorporated into Eq. (6.6). Thus we may be reassured that $\varphi_i(\tilde{f})$ should be driving the response towards a correct one, even though θ is not provided explicitly and $\hat{\theta}_{\varphi_i}$ are not retained. The compromise, compared to Eq. (6.12), is that we lose generality in the definition of weak learners and utility. Insight regarding the non-linear setting eludes us at this time. In the meantime, Sections 6.5.1 and 6.5.2 offer additional empirical evidence for our claims.

6.3 Dynamic Pursuit for decision making

As we have just emphasised, during decision making pathogen specificity, epitope abundance and discriminatory bias are each mediated through competitive exclusion. In a maximal simplification of the biology, we simulate each of these competitions independently until reaching steady-state. That is, we focus only on the steady-state values, not how these quantities interact as they evolve. Ostensibly, we must run at least four simulations of Alg. (11) to compute $\varphi_i(G)$, $\varphi_i(\tilde{f}^+)$, $\varphi_i(\tilde{f}^-)$ and $\varphi_i(x)$. Depending on the experimental setup, the first three may be computed once and ameliorated over multiple $\varphi_i(x)$ – e.g. see Alg. (13).

In the empirical validation that follows, Alg. (13) was used for the batch learning experiments of Sect. 6.5.1. However, in order to demonstrate the scalability and adaptation of Dynamic Pursuit for the continuous learning experiments in Sect. 6.5.2, it was necessary to implement our own sparse linear algebra data-structures and routines. This gave us opportunity to make the following improvement to Alg. (13). The independence of each exclusion process means that it is possible to compute all of the steady-states in one *non-terminating* simulation of the original Dynamic Pursuit algorithm (Alg. 11). The trick to making this work is simply to perform the necessary linear algebra routines on "scalars" that are in fact 4-tuples $\{f^+, f^-, G, x\}$. These scalars are then used to represent clonotype population ρ_i, clonotype capacity k_i and surface correlations in x, G, $\tilde{f}^+$ and $\tilde{f}^-$, now folded into a single sparse matrix. As the simulation runs the user is free to perturb the *environment* and *context* (i.e. provide "feedback") as well as perturb the *target* (i.e. elicit a prediction). A response is always available on demand, but the nature of the response will change as the population adapts to perturbations. A decision stabilises with the population. This is not only a lot more biologically satisfactory than the rigid logical procedure in Alg. (13), but is computationally more efficient as there is only one sparse surface matrix and the community matrix K is shared across each competitive process.

```
function dpl($x$ ; $\Phi$, $G$, $\tilde{f}^+$, $\tilde{f}^-$)
  if environment or context has changed then
    // update populations to reflect change
    $\rho^G = dp(G, \Phi, \rho^G)$
    $\rho^+ = dp(\tilde{f}^+, \Phi, \rho^+)$
    $\rho^- = dp(\tilde{f}^-, \Phi, \rho^-)$
  end
  // this simulation run is decision specific
  $\rho^x = dp(x, \Phi)$
  // integrate systemic response
  $\hat{\theta} = \sum_i \left(\frac{\rho_i^+ - \rho_i^-}{\rho_i^G}\right) \rho_i^x$
  return $\hat{\theta}$
```

Algorithm 13: Dynamic Pursuit for decision making, which makes use of the original Dynamic Pursuit (Alg. 11) as a subroutine to compute the steady-state of clonotypes competing over x, G, $\tilde{f}^+$, and $\tilde{f}^-$. Note that the presentation here is designed for clarity, see the text for a more elegant implementation.

6.4 Justification of Surrogate Data

Before empirical analysis, we briefly discuss the nature of our data. Sufficient biological data to assess our model's precision (in Levin's sense of the word) is not currently available. Although epitope prediction [85, 68, 149] is an active field, with some data available, at the time of writing this work has mostly been limited to *contiguous*, so-called "linear", epitopes (although things are improving [169, 102]). Such epitopes do not reflect our ligand-binding model and only account for a negligible minority of epitopes *in vivo* [85]. As a surrogate we will use textual data extracted from natural language documents. That this is appropriate is by no means obvious, so we now defend this decision.

In 1949, George Kingsley Zipf published an ambitious position that human behaviour could be understood by a single *Principle of Least Effort*, much akin to the principle of least action from physics [192]. Through a mixture of brilliant rhetoric and ingenious empirical demonstration, Zipf built his argument's foundation on a statistical study of natural language texts, culminating in several empirical laws. The most famous of these, which bears his name, is that the occurrences of words are inversely proportional to their *rank* in *frequency*

$$f_i \cdot r_i = k \tag{6.18}$$

where k is a constant. Thus, a rank-frequency graph exhibits an exponential decay; linear with logarithmic axis (see Fig. 6.4(a)). Conversely, the amount of words in a corpus is approximately the sum of a harmonic series $F = k \sum_r \frac{1}{r}$

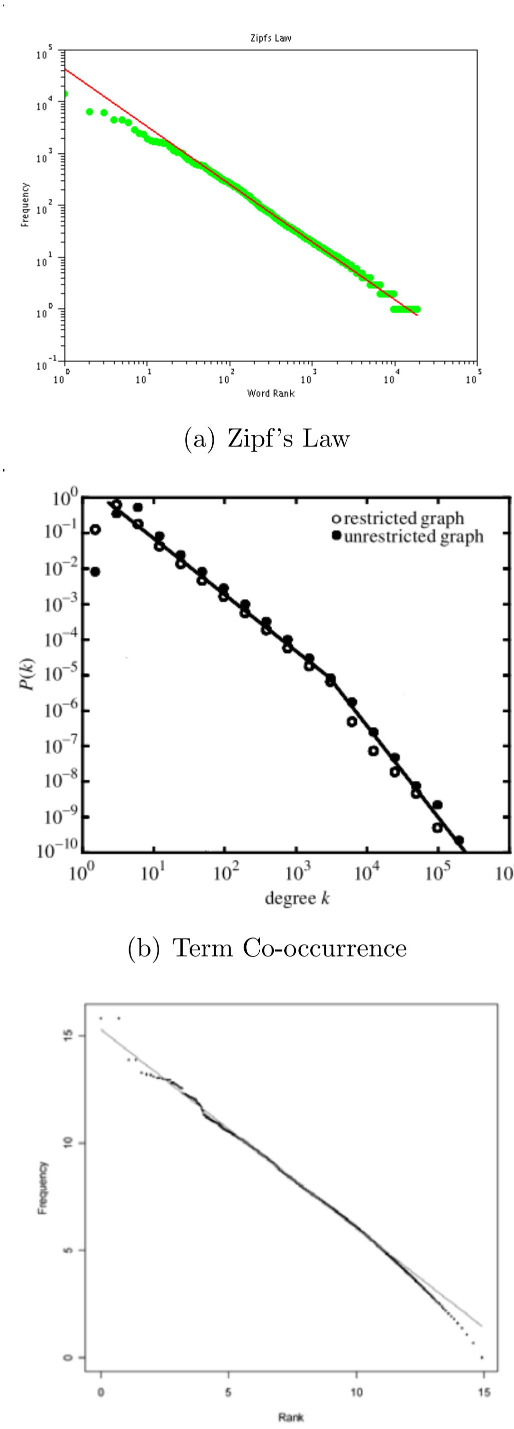
– with few words being used significantly more than average, but many words significantly less than. For Zipf, this law was the result of two competing "forces" of economy in communication: the *force of unification*, resulting from the desire of the speaker to economise to a single word repeated with 100% frequency to represent all meanings; and the *force of diversification*, resulting from the desire of the listener to economise to a 1:1 mapping between word occurrence and meaning. A compromise between these extremes results in words being repeated, and reused, as is consistent with natural languages. Zipf's harmonic series can be generalised to decaying curves of the form

$$f_i = kr_i^{-p} \tag{6.19}$$

where for Zipf, $p = 1$. More recently, Sole et al. [67] have observed that the topologies of word co-occurrence networks exhibit this characteristic decay also (Fig 6.4(b)). It turns out that such "power laws" are proliferate in physics, biology and economics [142, 82]. These ideas have recently enjoyed a cross-disciplinary renaissance due to the work of Barabási on scale-free[7] network topologies [9]. Though the validity of some of these observed "laws" are considered dubious [38], we are less interested in whether a *bona fide* power-law exists, than in emphasising the long-tailed distribution that does accurately describe all of these domains.

Notably, such power laws are present throughout the environment of the immune system: protein-protein interactions, protein functions, metabolic pathway connectivity and, most crucially from our perspective, the occurrence of n-mer base sequences in DNA [117], which ultimately transcribe into amino-acid sequences and thus peptides, and the occurrence of peptides in digested proteins [113]. Thus, textual data exhibits the appropriate statistical properties. Less quantitatively, in his 1987 Noble lecture Jerne drew an analogy between linguistics and immunology (recall the epigraph that opens this chapter), observing *"Every amino acid sequence is a polypeptide chain, but not every sequence will produce a well-folded stable protein . . . some grammatical rules would seem to be required"*. More recently, immunologists have explored spam-detection as a suitable surrogate for assessing immunological functionality [2]. Thus, both quantitatively and qualitatively, there is some justification for our approach.

[7] A genuine power-law exhibits no natural scale, that is, the curve is the same shape at any magnification. In physics at least, such scale-free configurations tend to occur at the boundary of critical phase transitions, thus indicating interesting phenomena are occurring.



(a) Zipf's Law

(b) Term Co-occurrence

(c) Peptide occurrence in proteins

Figure 6.4: The statistical properties of language revolve around the same distributions that pervade biology. **(a)** Zipf's empirical law of word occurrences demonstrated on the text of Herman Melville's *Moby Dick*; **(b)** The degree distribution of word co-occurrences, taken from [67]. **(c)** The distribution of peptide occurences in digested proteins [113].

6.5 Empirical validation

It is important to note that we do not think it reasonable to expect Alg. (13) to perform well on arbitrary classification tasks. It is specifically suited to exploit the statistical properties of sparse, high-dimensional problems. This is not so bad, given that these are the mainstream problems of modern statistical inference. We justified above that this may well describe the immunological context too.

6.5.1 Batch learning

Our first set of experiments use a subset of the UCI *newsgroups* dataset to produce a task of discriminating `comp.graphics` from `alt.atheism` postings. Recall that in Chapter 4 the clonal selection based algorithm AIRS was shown to perform no better than random guessing on this dataset. Our goal here is to assert the efficacy of our systemic response model; but we will not be overly concerned with optimising performance metrics.

Protocol

We compare the performance of our algorithm against the k-nearest neighbour classifier ($k = 7$ was empirically best) and the linear classifier. These algorithms represent the theoretical extremes between which our approach lies. The following experimental protocol is standard for text classification (see e.g. [121])

- **Preprocessing.** The newsgroups data comes in raw SMTP e-mail format. We perform a basic preprocessing of the text that involves removing SMTP related data and punctuation. It is convenient to also stem words to their common prefix and remove functional stop-words, such as "the", leaving a vocabulary of 5000 words. Each document is stored as a sparse vector x with dimensionality $n = 5000$ and is normalised $\|x\|_2 = 1$ so that word frequency is relative to document size. We do not perform any other common text preprocessing based on global analysis of the corpora, such as inverse-document-frequency term weighting.

- **Cross-Validation.** The collection of document vectors is shuffled randomly and split into $c = 10$ disjoint subsets, or "folds". This allows us to perform c replications of the learning experiment. In each replication $i = 1 \rightarrow c$, the i'th fold is retained and the algorithms allowed to observe the other $c - 1$ folds (i.e. "training"). The i'th fold provides unobserved data to assess the algorithms ability to generalise (i.e. "testing").

Documents are distributed evenly between classes. For reasons discussed later, we cap the total number of documents at 1000. For each algorithm we record the summary statistics for the following metrics

- **Classification performance:** the *accuracy*, *sensitivity*, *specificity* and *precision* of each algorithm. See Sect. 3.1.2 for a discussion.
- **Computational performance:** the computational time in seconds (per observation) to both train the algorithm and to produce decisions.

Unlike the experiments of Chapter 5, Dynamic Pursuit does not have an *a priori* provided repertoire (basis) with which to construct representations. This is a potentially open-ended area for empirical research. We use the following procedure, not because it is sophisticated, but because it minimises confounding factors in assessing our proposed model; maintains some sense of biological plausibility; and provides empirical support for our analysis in Sect. 5.2.3:

- We decide on an *a priori* fixed size repertoire $|\Phi| = 10,000$. Each φ_i is allocated $c = 3$ non-zeros components[8] that are generated by taking a uniformly random c-step walk on the matrix/graph $G = \sum |x_i\rangle\langle x_i|$. Thus each φ_i can reasonably be expected to attain some capacity $\langle \varphi_i | G | \varphi_i \rangle$.

Note that $10,000 \approx 5 \times 10^{-7}$ of the possible $\binom{5000}{3}$ receptor space. We perform no additional search for new receptors during algorithm execution. We also do not *a priori* assess receptors with respect to producing good results or any other metric of quality. *These are very severe restrictions*, that might be considered unreasonable to impose on a classification algorithm. We stress again that our goal is demonstrating the efficacy of competitive exclusion between randomly generated receptors in producing a coherent systemic response.

Computational performance analysis

We plot our results in Fig. (6.5). First, notice the performance extremes represented by k-nearest neighbour (*knn*) and least squares (*lsq*). The former is impossible to beat in *training CPU*, because *knn* involves no training whatsoever; the latter is impossible to beat in *test CPU*, because the *lsq* decision is simply a dot-product calculation. The train/test optimality for *knn* and *lsq*, respectively, comes with a necessary computational cost: each must invest all of its

[8] The value 3 is unfortunately a "magic number". Lacking any compelling reason to choose one value over another we chose this because it is well-known to be the average number of keywords in a search engine query. This is not very scientific, but it is pleasing as a modern twist on Jerne's epigraph for this chapter: an antibody is a *query*, not a sentence or phrase.

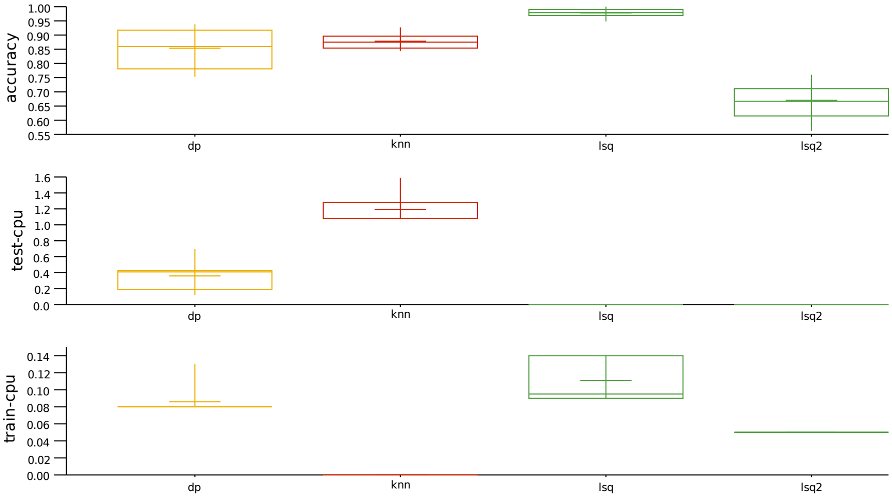


Figure 6.5: High-level classification performance metrics comparing competitive exclusion with randomly generated receptors against two statistical algorithms.

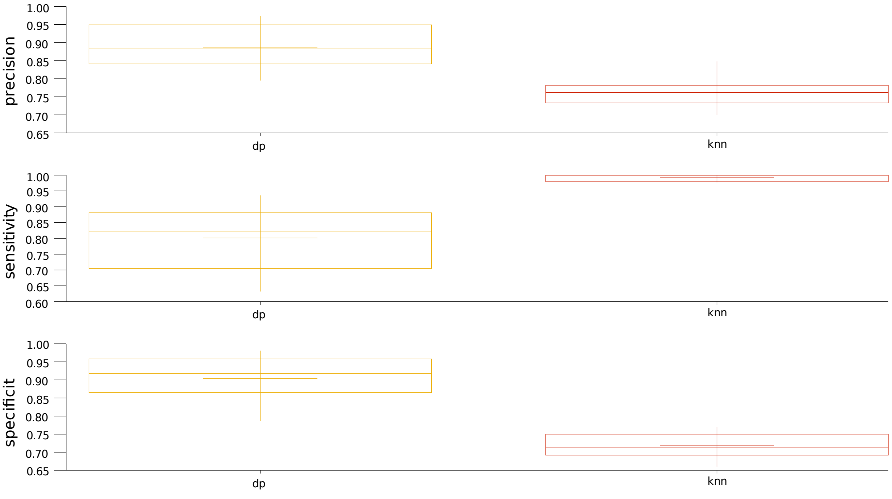


Figure 6.6: Elucidation on the learning behaviour of competitive exclusion against that of k-nearest neighbour, which achieves similar accuracy.

computational effort in the complementary learning phase. Thus we see that the *test CPU* time of *knn* is penalised by calculating nearest-neighbours for n training documents. Similarly, the *train CPU* of *lsq* is penalised in performing either of the matrix inversions in Eq. (3.9) – both very large matrices. For completeness, we include two computational variants of the least squares method. The first, (*lsq*) is the method described in this thesis. The second (*lsq*2) uses Matlab's built-in inversion operator "\" which exploits a variety of highly optimised and complex matrix factorisation techniques. Figure 6.5 makes it clear that the saving in computational effort carries significant cost in the numerical quality of the results. We will not consider *lsq*2 a valid competitor, but include it because it represents a potential "folk wisdom" argument against our analysis: that the matrix inversion is an algebraic artifact and need not be done in practice.

It is with respect to this trade-off in computational performance that Dynamic Pursuit (*dp*) is particularly interesting. Ostensibly, *dp test CPU* performance is substantially faster than *knn*; its *train CPU* performance is marginally better than *lsq*. But this is only true at this particular snapshot of 1000 documents. What this graph does not show is that as the number of documents grows, both *knn* and *lsq* CPU performance grows much faster, of the order $O(n^2)$ and $O(n^3)$, respectively. Including *lsq* in this experiment is why the total number of documents was capped at 1000. In contrast, the *test CPU* of *dp* depends on calculating the steady-state $\varphi_i(x)$ and then integrating the systemic response from surviving clonotypes. This computation is dominated by repertoire size, not the number of observations or their dimensionality. Similarly, the *train CPU* of *dp* depends on calculating the steady-states $\varphi_i(G)$ and $\varphi_i(\tilde{f})$. This again is independent of the quantity of observations, beyond the trivial construction of G and $\tilde{f}$. But the dimensionality of observations *will* become a dependency if one plans to search through $\binom{n}{c} < n^c$ possible receptor configurations. We have used the strategy suggested in Sect. 5.2.3 to make this search somewhat more reasonable. We now show that exhaustive search is unnecessary for performant inference.

Classification performance analysis

Ostensibly, the classification *accuracy* of our algorithm is, perhaps reasonable, but not overly compelling. Mean and median performance is comparable to *knn*, albeit noticeably more variable. The additional variability in *dp* accuracy is the result of random receptor generation during training. This variability seems unavoidable without some form of compensation – e.g. by employing a substantial repertoire size that ensures "good" receptors are always included with high probability; or optimising the generation process to produce consistently "good" receptors. The latter is a fairly standard pre-processing procedure in statistical inference; the former may be more descriptive of the immune system's strategy. But even though sampling $10^{-5}\%$ of the repertoire, uniformly at random, incurs a cost in variance, this variance in accuracy is still between $80-90\%$. This strongly suggests that the algorithm is still *learning*, even under such severe conditions. We quantify just how the algorithm is learning in Figure 6.6. It is apparent that although *knn* has high *sensitivity* (probability of correctly predicting "positive" given a positive observation) its *precision* (probability of correctly predicting positive given a positive prediction) is in fact quite low. That is, it is biased towards positive predictions. In contrast, *dp* is less sensitive but more precise. That is, it is less likely to predict positive, but if it does, it is more likely to be correct. The *dp* algorithm has higher *specificity* (probability of correctly predicting "negative" given a negative observation) but is slightly less specific than *knn* is sensitive. In terms of unconditional discriminatory power, these differences average out to qualitatively similar *accuracy*. Though it is clear from Fig. 6.6 that both algorithms are in fact learning in quite different ways.

In Fig. 6.5 we can also observe the theoretical effects of dimensionality on the trade-off between classifier complexity and accuracy (see Chapter 3). Non-linear decision boundaries have not provided sufficient advantage to *knn* to improve performance over a linear decision boundary; whereas the linear model is more accurate and robust (lsq), albeit subject to numerical instability ($lsq2$). A crucial detail not explicit in Fig. 6.6, but apparent in figure 6.15 later, is that although we generate 10,000 receptors during initialisation, only 50-1500 survive the competitive exclusion process during training. That is, the representation constructed by Dynamic Pursuit is 1%-30% the dimensionality of the original observations and 0.5%-15% of the receptor space – with little practical difference in classification performance from *knn*. It is perhaps remarkable that so few, low-dimensional random projections are capable of retaining sufficient representational and discriminatory power.

6.5.2 Continuous learning

It would seem evident that an "immune-inspired" algorithm should not proceed through a batch training (i.e. model fitting) stage and subsequent deployment stage. Like the immune system, the algorithm should learn continuously. This is probably in part why the non-parametric, nearest-neighbour approach has traditionally been so popular in AIS – there is no model to be fit. Dynamic Pursuit is designed to be well suited for continuous learning. As shown in the preceding analysis, its scaling properties do not depend on the number of observations. Although our model is somewhat parametric, by unrolling the matrix inversion in Eq. (6.5) we leave it in a state where the model fitting and decision making can be interleaved. Lastly, the quantities dynamic pursuit relies on, G and $\tilde{f}$, can readily be updated incrementally

$$\begin{aligned} G_t &= G_{t-1} + |x_t\rangle\langle x_t| \\ \tilde{f}_t &= \tilde{f}_{t-1} + \theta_t x_t \end{aligned}$$

Notice also that G can be updated independently from $\tilde{f}$. That is, we can incorporate more data into our model of the environment G than we have explicit feedback on $\tilde{f}$. From the inferential perspective, this would take us into the domain of *semi-supervised learning* [191] – using both labelled and unlabelled observations – but here we focus on the incremental aspect of our model only.

Information filtering

We continue in the domain of text classification, but relax the somewhat artificial batch learning methodology. Instead, the problem formulation is that documents now arrive sequentially in a "stream" and must be classified on-line, in a timely manner. Classic examples are spam filters or aggregated news filters: each must learn to produce decisions from limited previous exposure to both the preferences of the user (the unknown function) and the sources that documents are arriving from (the unknown population). Typically, the class distribution of "good" and "bad" documents is also heavily skewed [10, 86].

Nanas et al. [139, 138] have been strong advocates of this learning domain as the most appropriate for immune-inspired algorithms. We are inclined to agree: statistically and methodologically it seems closer to the biology. In developing *Nootropia*, Nanas et al. proposed an empirical framework for assessing the ability of filtering algorithms to adapt over time to user's changing interests [136, 137]. As a step towards producing comparable results, we work with the same dataset as Nanas et al. However, our protocol (discussed next) differs from theirs in several respects: (*i*) we do not perform global analysis of the data as a preprocessing step; (*ii*) we insist that each algorithm produce a hard classification, rather than a ranking from most to least *relevant*; and (*iii*) we allow algorithms to make use of negative as well as positive observations. In all, we consider these differences to make for a more stringent comparative assessment.

Protocol

Nanas et al's preferred dataset is a subset of the publicly available[9] collection of Reuters news-wire articles collected over 1987. Many articles have been manually classified into one or more news-related topics (e.g. earnings, acquisitions, etc.) and it is these topics that we will take to indicate *relevance*. Articles with no classification and classes with less than 100 relevant articles were removed from the original dataset, leaving 6753 articles. The distribution of topics is shown in Table 6.1. Our text preprocessing methods were the same as for the previous experiments, resulting in a vocabulary of 20,121 words (i.e. dimensions).

Due to the sheer magnitude and rate of change of streaming data, traditional batch learning algorithms such as least squares and prototype-based algorithms such as k-nearest neighbour are not feasible to deploy. For comparative analysis, we use as a baseline *Rocchio's Algorithm* [152]. This algorithm is well established in the information filtering and retrieval literature and is the benchmark algorithm used by Nanas et al. Its popularity for continuous learning stems from its minimal space and time complexity: it retains only a single "profile" vector per class that represents the mean document in that class. Thus it is very efficient to train and produce decisions. We include two variants of Rocchio's algorithm from the literature. The first is the classical Rocchio's algorithm

$$\hat{\theta} = f(\hat{x}) = \langle \hat{x} | f \rangle = \left\langle \hat{x} | \frac{1}{N^+} \sum_{\theta_i = +} x_i - \frac{1}{N^-} \sum_{\theta_j = -} x_j \right\rangle \tag{6.20}$$

Equation (6.20) is the same decision function as our Danger model, Eq. (6.8). Such comparison allows us to empirically assert earlier theoretical claims. The second variant [155] is used by Nanas et al. where class profiles are updated as

$$\begin{aligned} f^+_{t+1} &= \delta f^+_t + \beta x_t \quad \text{if} \quad \theta_t = + \\ f^-_{t+1} &= \delta f^-_t + \beta x_t \quad \text{if} \quad \theta_t = - \end{aligned}$$

In contrast to Eq. (6.20) this update allows the influence of documents to decay over time. We use the same parameters reported by Nanas et al. $\delta = 0.95$ and $\beta = 0.25$. Preliminary experimentation (not reported) showed that we can substantially improve the performance of this algorithm by predicting

$$\hat{\theta} = f(\hat{x}) = \frac{\langle x | f^+ \rangle}{\|x\|_2 \|f^+\|_2} - \frac{\langle x | f^- \rangle}{\|x\|_2 \|f^-\|_2}$$

[9] Reuters-21578 Distribution 1.0 `http://www.research.att.com/~lewis`.

The reason for this improvement is two-fold. Firstly, by normalising profile vectors we achieve a much stronger conditionalisation on classes than using a scaling factor such as $\frac{1}{N}$. This is because each feature is weighted relative to the other features in that class, rather than simply its mean value. Secondly, the difference between class-specific dot products, rather than the dot-product with the class difference vector $\langle \hat{x} | f^+ - f^- \rangle$, better quantifies which class $\hat{x}$ is closest to because it is less effected by skew class distributions.

To assess each algorithm, the following procedure was repeated for each topic, with that topic being considered *relevant* and all others *irrelevant*:

- Each article was processed in chronological order. Algorithms were allowed to passively observe $\{x, \theta\}$ pairs until 10 *relevant* articles had been observed.
- After this initialisation stage, for each article taken from the stream the algorithms were requested to make a prediction $\hat{\theta}$ of that article's relevance to the current topic, given only x.
- After prediction, the algorithms received feedback θ on the article's relevance which they may use to update their parameters.

The accumulation of *true positives*, *false positives* and so on were recorded over the entire stream. In addition to the standard metrics, we include our own metric *discrimination*, defined as

$$\left(\frac{TP}{TP + FN} \right) + \left(\frac{TN}{TN + FP} \right) - 1 \tag{6.21}$$

that is, the *true positive rate* plus the *true negative rate*, minus one. This metric lies in the interval $[-1, +1]$ representing 100% incorrect and 100% correct, respectively. How the metric differs from e.g. *accuracy* is that majority class constant decisions, random guessing and any other strategy that produces results with the same distribution as classes are all assigned 0 discrimination. Thus, the range $[0, 1]$ of this metric quantifies improvement beyond trivial unlearnt decisions.

Implementation notes

For continuous learning, it is not possible for Dynamic Pursuit to generate receptors *a priori*, as it had been in the batch setting. Instead, we implement a minor variation on the receptor generation strategy employed earlier:

- We do not *a priori* fix the repertoire size. Again, each φ_i is allocated $c = 3$ non-zero components that are generated by taking a uniformly random

topic	$p(topic)$	topic	$p(topic)$
acq	0.31	coffee	0.02
corn	0.03	crude	0.08
earn	0.27	gold	0.01
grain	0.08	interest	0.05
livestock	0.01	money-fx	0.09
money-supply	0.01	nat-gas	0.01
oilseed	0.02	ship	0.04
soybean	0.01	veg-oil	0.01
wheat	0.04		

Table 6.1: The distribution of topics in the Reuters news-wire article stream. Notice that the topics *acq* (acquisitions) and *earn* (earnings) dominate the stream but all topics have a highly skewed distribution of *relevance* and *irrelevance*.

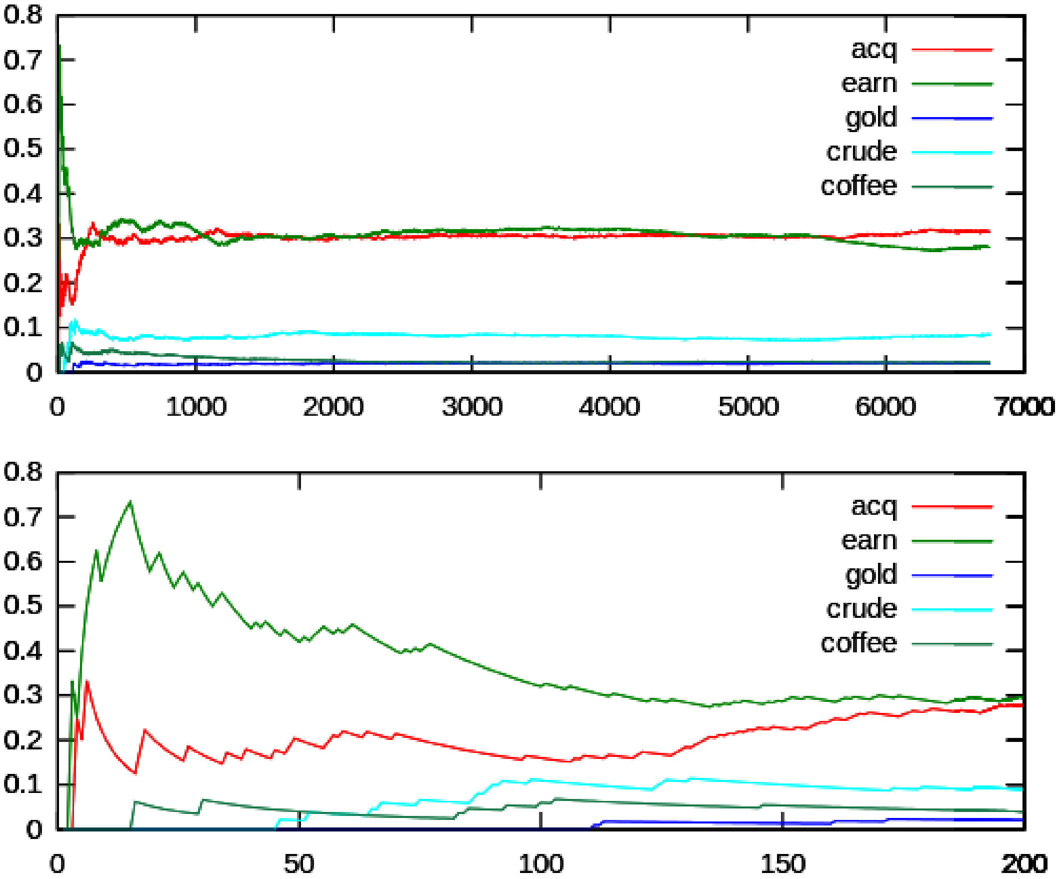


Figure 6.7: **Top:** The true probability of relevance as a function of time for the subset of topics our analysis focuses on. **Bottom:** Close-up of the first 200 articles where the topic *earn* rapidly dominates the stream's content.

c-step walk. However, instead of taking a random walk on G during initialisation, we take a random walk on the surface description of each article $|x\rangle\langle x|$ as it is processed. Thus, there are always φ_i that can be reasonably expected to attain some *target* capacity $\langle\varphi_i|x\rangle\langle x|\varphi_i\rangle$. Again, no *a priori* assessment of receptor quality was carried out.

The number of receptors generated per document was set to 100. Experimentation with values up to 1000 produced no obvious improvement. The average number of unique words per article was 56.9 ± 42.7 with a maximum number of 390 and minimum 20. Thus, 100 receptors covers between 0.08 and 0.00001 possible receptors with an average of 0.003, for $\binom{20}{3}$, $\binom{390}{3}$ and $\binom{56}{3}$ respectively. With only 56 unique words 100 × 3 receptors can afford to cover each word 6 times on average. Integration into the repertoire depends on how these receptors generalise to G and $\tilde{f}$, either exploiting uncrowded "niches" or having sufficient capacity to overcome an initially low concentration.

Experimental analysis

Due to the similarity of results for classes with similar probabilities of relevance, we focus our analysis on the classes *acquisitions*, *earnings*, *crude oil*, *coffee* and *gold*. Respectively, these represent $p(\theta = +)$ ranging across 0.31, 0.27, 0.08, 0.02 and 0.01. We include *crude oil* in particular as it has been observed [100] that these articles have a restricted vocabulary with words that are highly indicative of document class. In contrast, topics such as *acquisitions* have a much broader vocabulary. The reason for including both *acquisitions* and *earnings*, which have similar probability of relevance, is that the latter produced results that were anomalous with respect to the trend in all other datasets.

Figures 6.8-6.12 plot the progression of metrics *accuracy*, *sensitivity*, *specificity* and *discrimination* for each algorithm as the stream is processed. There are some general observation that can be made from these results

- For very low probability $p(\theta = +) < 0.1$ topics, the classic Rocchio algorithm *Rocchio (mean)* – which is also our Danger model in Eq. 6.8 – performs poorly. This is because the mean profile vectors are still subject to class skew effects, resulting in this case in a tendency to predict *relevant*. This effect is most notable in *coffee* and *crude*, where the true negative rate of *Rocchio (mean)* plummets and never recovers. This validates our claims in Sect. 6.2.1. In the immune system proper, $p(nonself) \ll 0.1$.

- For the same $p(\theta = +) < 0.1$ topics, the improved *Rocchio (cosine)* and *Dynamic Pursuit* perform comparably. For *gold*, *Rocchio (cosine)* is more accurate, and marginally more discriminatory, due to improved specificity. For *coffee* and *crude*, Dynamic Pursuit is marginally more accurate – but *less* discriminatory, due to a preference for the majority class at some cost to *sensitivity*. *Crude* represent the largest margin between *sensitivity* for both algorithms, which may be due to *Rocchio (cosine)* being able to make more use of the restricted vocabulary than Dynamic Pursuit's random repertoire.

- For the *acquisitions* topic, the difference in performance between algorithms is negligible, practically. *Rocchio (mean)* is less plagued by class skew. *Rocchio (cosine)* and *Dynamic Pursuit* look like mirror images of each other, the former favouring *sensitivity* the latter *specificity*.

- On the *earnings* topic, *Dynamic Pursuit's* performance is radically altered: rapidly approaching around 99% true positive rate and 85% false positive rate. That is, in contrast to all previous topics, Dynamic Pursuit now strongly favours predicting the minority class *relevant*, almost to the exclusion of ever predicting *irrelevant*. This anomalous behaviour can be explained as the consequence of both the rapid domination of the *earnings* topic in the early stages of stream processing (Fig. 6.7) coupled with the breadth of vocabulary in that topic due to its agnosticism with industry sectors. This results in Dynamic Pursuit producing a repertoire that quickly becomes biased towards *earnings* and is very difficult for new clonotypes to infiltrate as they tend to be redundant. The *Rocchio* algorithms are more robust to this effect because their features are never competing.

In Tables 6.2 and 6.3 we quantify these observations more precisely by providing a statistical analysis of each algorithm's mean performance across topics. For completeness, we include a statistical hypothesis test of the significance in *accuracy* differences between Dynamic Pursuit and the Rocchio-based algorithms. The statistically (and practically) significant difference from *Rocchio (mean)* substantiates our earlier claims about the contribution of systemic response beyond peptide-specific Danger signals. The lack of statistically significant difference with *Rocchio (cosine)* is also interesting. As shown in Fig. 6.15, the average repertoire size for this dataset was just over 300 clones – on a dataset with 20,121 dimensions. That is, the representation learnt by Dynamic Pursuit is only 1.4% the dimensionality of the representation used by the Rocchio-based algorithms and only 0.05% of the $\approx 600,000$ generated receptors. For the most part, performance is maintained, by maximising capacity while minimising competition.

Algorithm	Accuracy
Dynamic Pursuit	0.75 ± 0.272
Rocchio (mean)	0.59 ± 0.262
Rocchio (cosine)	0.83 ± 0.243

Table 6.2: Mean performance across topics.

Statistical significance of performance differences			
H_0	Difference	95% Confidence interval	Reject H_0
Rocchio (mean)	$+0.16 \pm 0.065$	$+0.032, +0.287$	yes
Rocchio (cosine)	-0.08 ± 0.044	$-0.167, +0.007$	no

Table 6.3: Statistical significance of *accuracy* differences with Rocchio-based baselines. *Dynamic Pursuit* can claim a statistically significant difference with *Rocchio (mean)*, which lends weight to our theoretical discussion of the benefits of systemic behaviour over peptide-specific responses. The lack of significant difference with *Rocchio (cosine)* is discussed in the text.

However, it is apparent that *Rocchio (cosine)* still has the advantage as algorithm. The development work necessary to deploy Dynamic Pursuit as an information filtering algorithm is of practical importance, but not part of this thesis.

Robustness

In Figure 6.13 we plot 10 runs of Dynamic Pursuit on the *acquisitions* topic. It is apparent that the particular repertoire that is realised during a run has an effect on the trade-off between *sensitivity* (true positive rate) and *specificity* (true negative rate). However, the effect on overall *accuracy* is negligible. Likewise, in Fig. 6.15 we plot the variation in population size and clonotype turnover over the same 10 runs. Recall, we do not fix the repertoire size. The stability of the repertoire size can be explained because clonotype viability is a function of antigen supply and demand and the competition dynamics ensure that redundancy is driven out. Recall also, that 100 clones were randomly generated with each new article and we did not *a priori* assess receptors for representation or discriminatory power. The rapid decline of survival for these new clones reflects that the competition dynamics only integrates those receptors that add value to the repertoire: either by preference for resources not currently competed over or sufficient capacity to overcome an initially low clonotype concentration with respect to established competitors.

Adaptation

Although not part of our thesis, demonstrating the capacity for adaptation (in a biological sense) is certainly a natural goal for biologically inspired work. What follows is more a record of future issues to be addressed.

To model a user with changing notions of relevance, we repeat the above experimental protocol but now consider both topics *gold* and *coffee* relevant. However, rather than simply label both as $\theta = +$ we assign positive labels probabilistically. The probability of one, or the other, topic receiving a positive label is linearly interpolated across the entire stream from $p(\theta = +|gold) = 1 \rightarrow p(\theta = +|gold) = 0$ and $p(\theta = +|coffee) = 1 - p(\theta = +|gold)$. Thus, the stream starts off with *gold* being the relevant class and gradually gives way to *coffee*, simulating a decrease (respectively, increase) of interest in a topic over time. In Figure 6.14 we plot our initial attempts at assessing the adaptivity of Dynamic Pursuit. The results are to be expected: only *Rocchio (cosine)* is capable of any form of adaptation due to the decay term in Eq. (6.21). Clearly, a simple decay of parameters is quite a weak interpretation of "adaptation". Both *Dynamic Pursuit* and *Rocchio (mean)* have similar flaws insomuch as the quantities they depend on are only ever added to. This means they cannot adapt until sufficient data has been observed to overcome the original class distributions. *Dynamic Pursuit* does manage to retain accuracy, simply because its *specificity* is robust to this change, but this preference for the majority class results in a drop in *discrimination*.

From the immunological perspective, the problem is clear enough. Although we make use of the systemic response to predict observation classes, the response does not in itself effect the environment. For example, in an immune response the production of antibodies signals the eventual destruction of pathogen, which changes the environment and, in turn, feeds back into the concentration of Danger signals, pro- and anti-inflammatory cytokines, and so on. None of these effects are present in our current model, or those in Chapter 2, but clearly are important.

6.6 Conclusion

We have extended the approximatory behaviour of competitive exclusion to the decision making setting. We discussed the qualitative likeness to the arrangement of components and mechanisms in the immune response and justified this arrangement theoretically in terms of both its relationship with the numerical methods underlying least squares and as an *implicit* variation on boosting. We then quantitatively demonstrated the efficacy of competitive exclusion in both batch and online inference and prediction. Competitive exclusion performed comparatively to well-established algorithms, but was additionally able to significantly reduce the representational complexity of observations. This is entirely consistent with the regularised optimisation criteria in Eq. (5.4) – Dynamic Pursuit retains accuracy, but prefers simplicity, by maximising capacity and minimising competition.

Taken together, this confirms that competitive exclusion is a robust mechanism for turning randomly generated receptors into representative and discriminatory detectors. However, a more sophisticated receptor generation process would be required to best make use of the repertoire's ability to expand and contract the representation. Integrating receptor mutation and the subsequent affinity maturation would be an obvious next step. Further, Eq. (5.1) and (6.6) need to be elaborated to better capture "adaptation", in the biological sense.

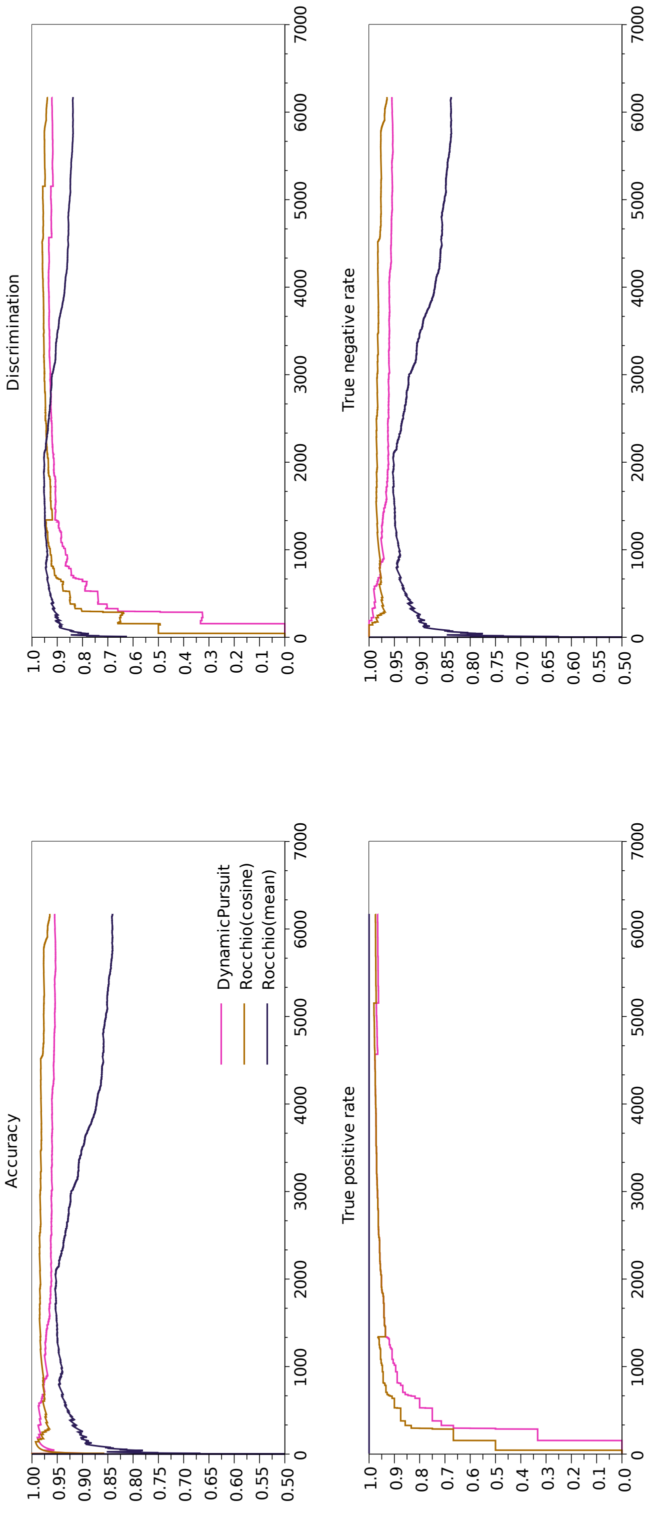


Figure 6.8: Filtering *gold* related articles where $p(\theta = gold) = 0.01$.

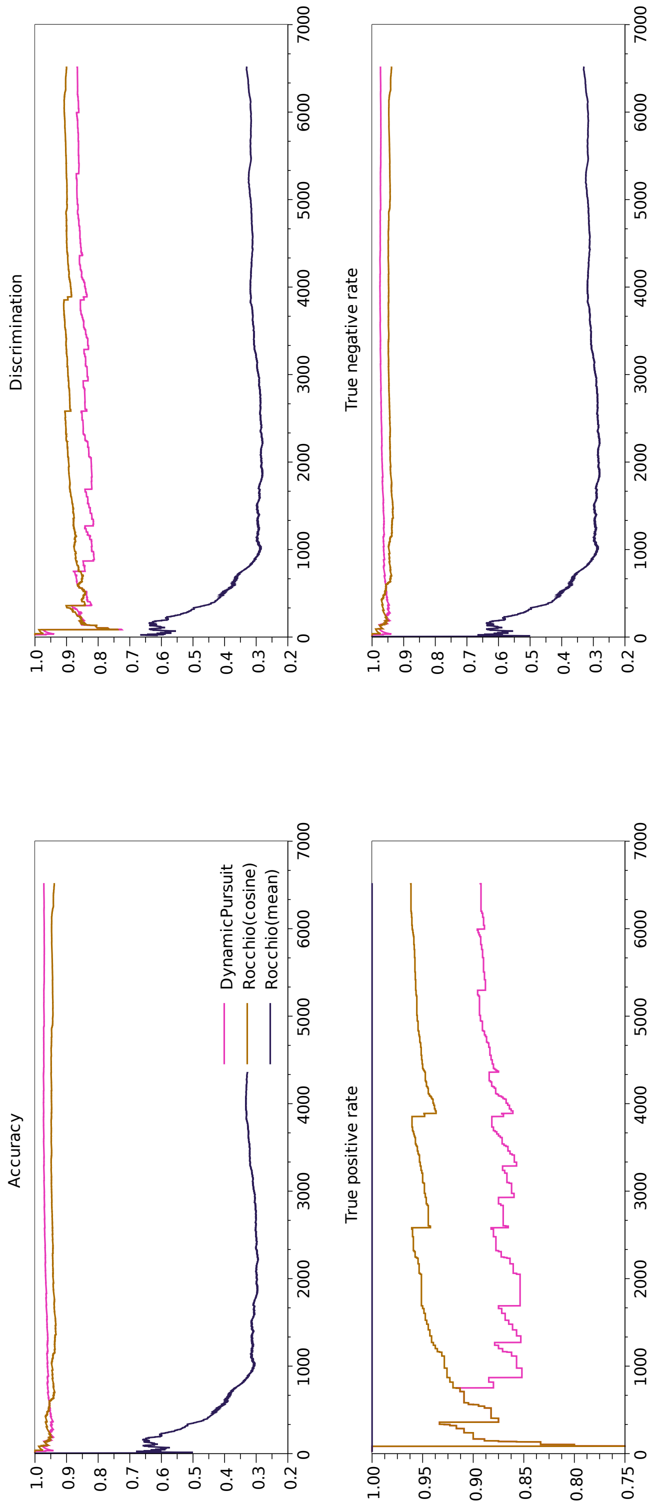


Figure 6.9: Filtering *coffee* related articles where $p(\theta = \mathit{coffee}) = 0.02$.

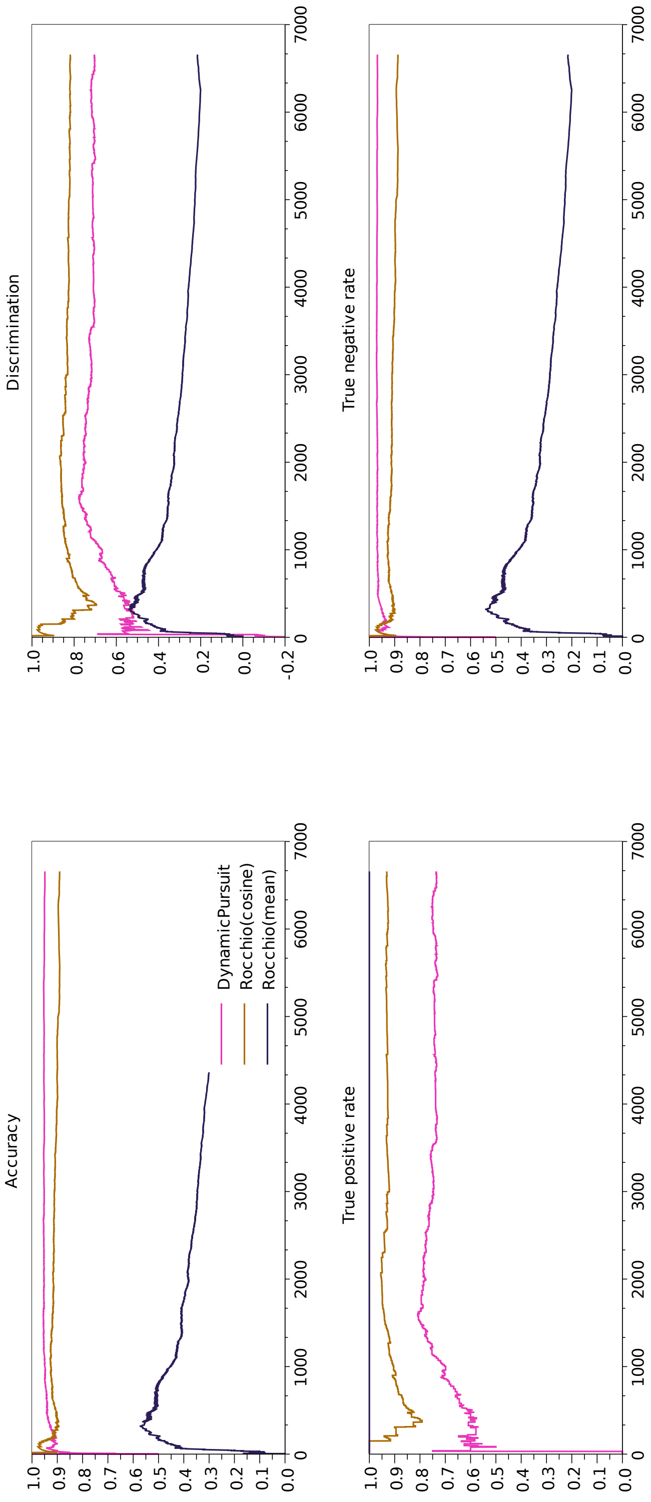


Figure 6.10: Filtering *crude oil* related articles where $p(\theta = crude) = 0.08$.

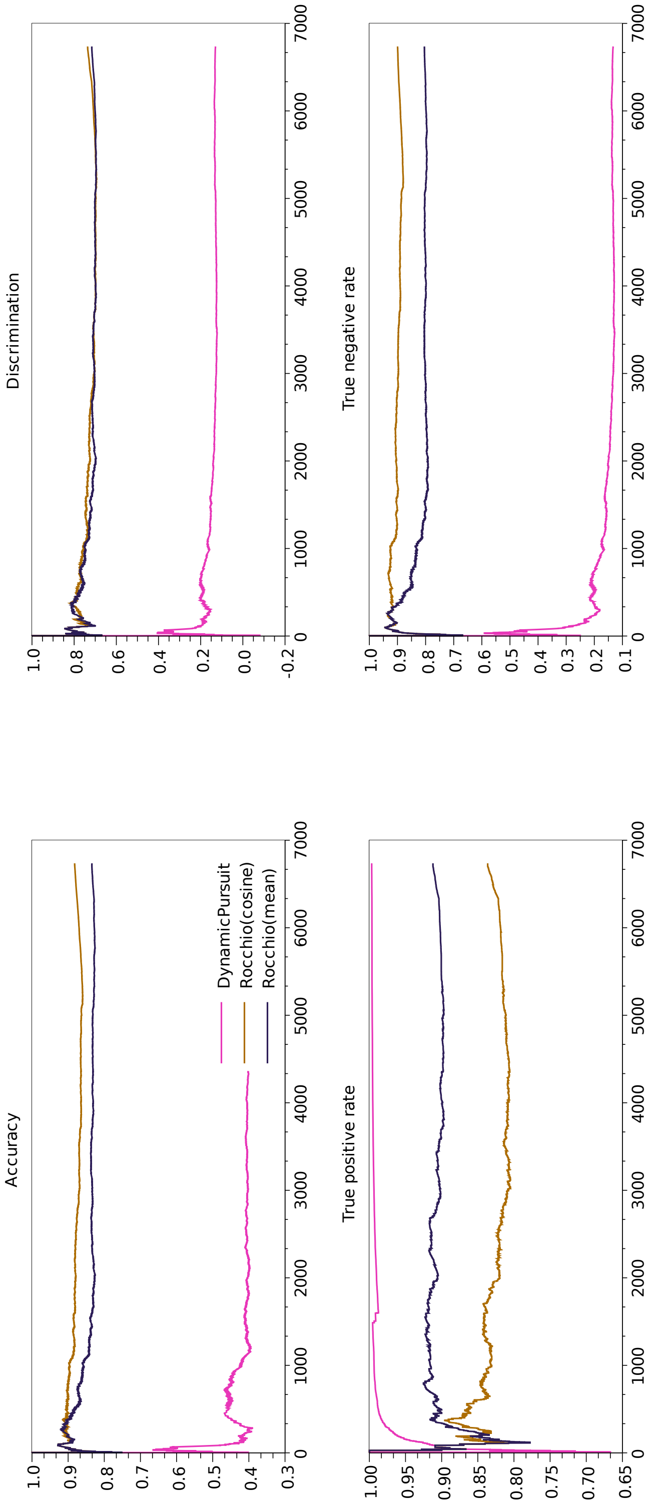


Figure 6.11: Filtering *earnings* related articles where $p(\theta = earn) = 0.27$. Note that these results are anomalous. See text for discussion.

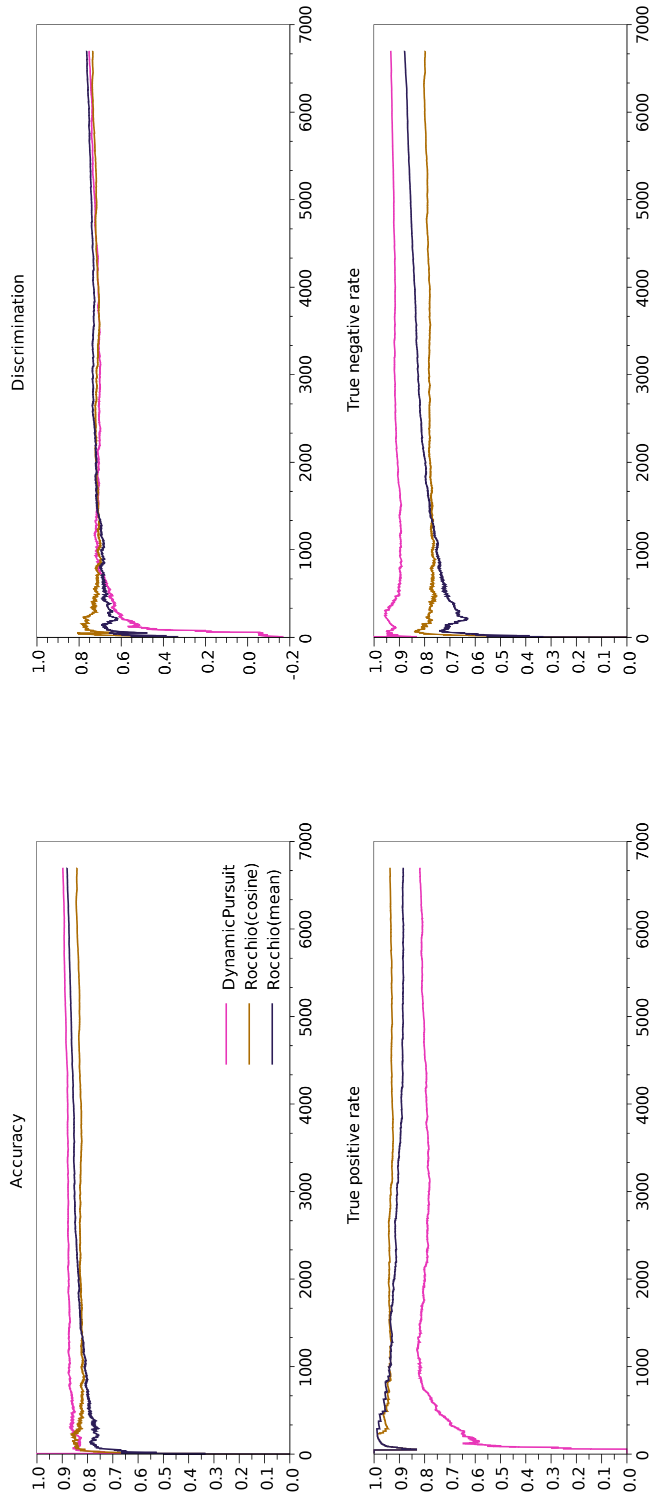


Figure 6.12: Filtering *acquisitions* related articles where $p(\theta = acq) = 0.31$.

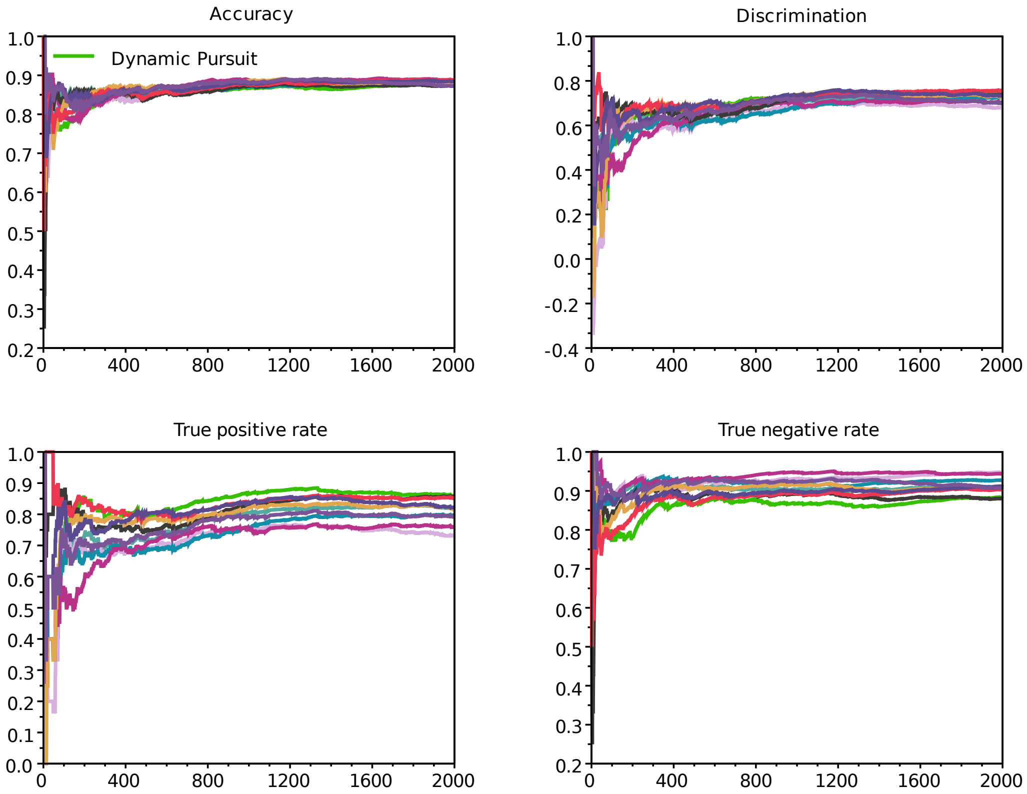


Figure 6.13: Replications of Dynamic Pursuit on the *acquisitions* topic. Although individual runs may have different *sensitivity* and *specificity* due to the particular repertoire, the overall effect on *accuracy* and *discrimination* is negligible.

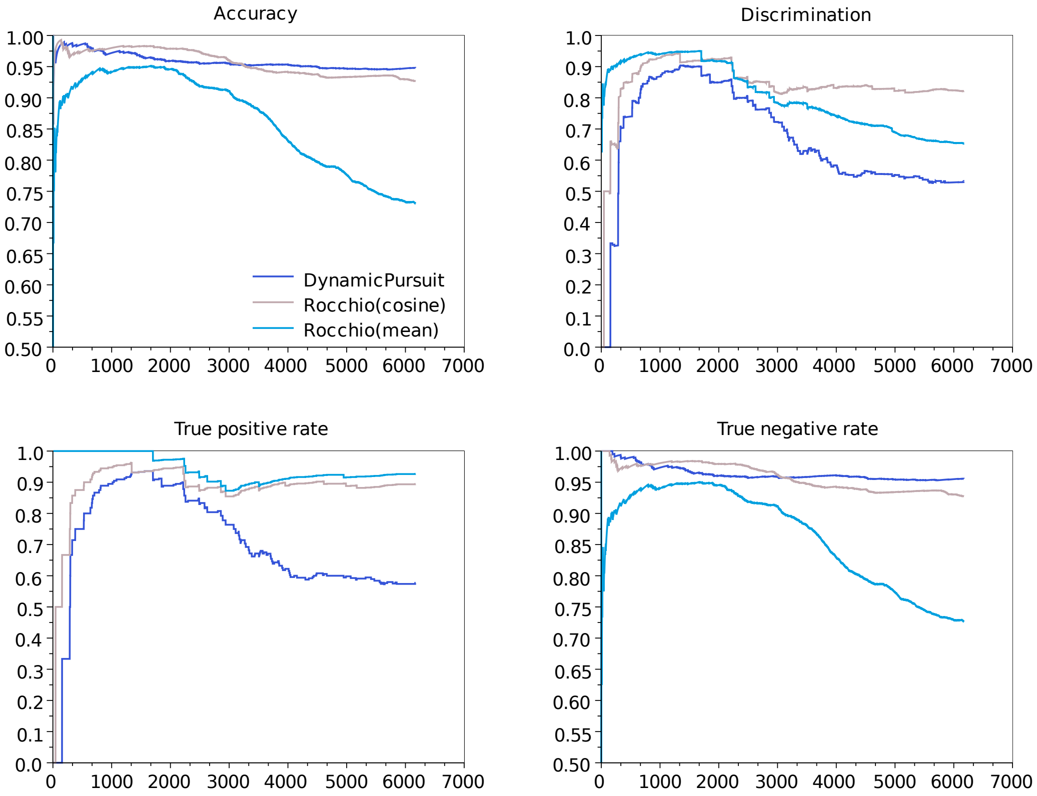


Figure 6.14: Performance of algorithms on the adaptation dataset where $p(\theta = +)$ changes linearly between the classes *coffee* and *gold* as the stream progresses. Only *Rocchio (cosine)* exhibits adaptation due to the decay term δ in Eq. (6.21).

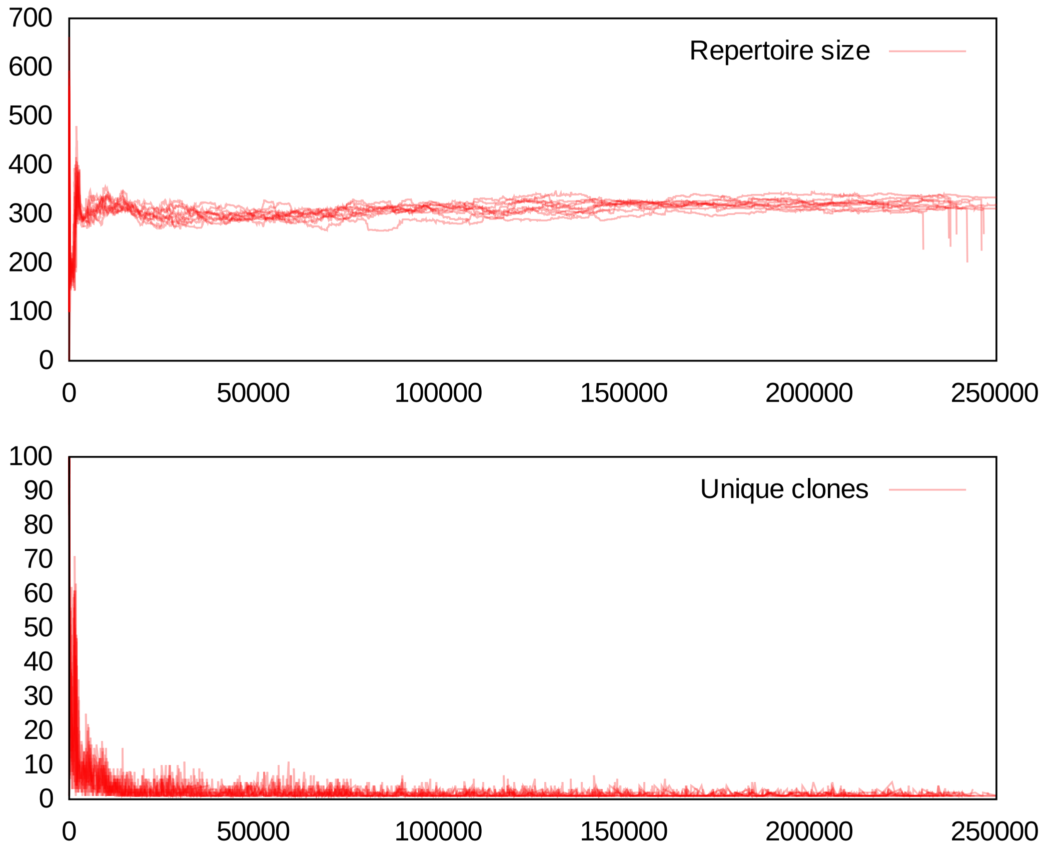


Figure 6.15: Evolution of the population size and number of surviving newly generated clones over 10 replication runs of Dynamic Pursuit on the *acq* topic. **Top:** Although no fixed repertoire size is enforced, the size of the repertoire is a function of antigen supply/demand and is consistent across replications. **Bottom:** The number of newly introduced clones that survive exclusion is a decreasing function of time, as the repertoire becomes less redundant.

Chapter 7

Conclusion

Fundamental progress has to do with
the reinterpretation of basic ideas.
Alfred North Whitehead

All models are wrong; but some are useful.
George E. P. Box

Let us is briefly review the thesis. In Chapters 2-4 we established foundational problems with the representational abstractions and decision-making mechanisms employed by immunologists and computer scientists in interpreting the immune response. In Chapters 5 and 6 we formulated an alternative interpretation grounded in the numerical methods of approximation, simulation and statistical inference. Recall, our research questions from Chapter 1.

- Can knowledge of the requirements for statistical decision making be applied to develop a plausible model of processes in the immune system?
- If so, does such a perspective offer novel insight that can be exploited by immunologists, computer scientists or statisticians?

With respect to the former, we submit a possibly contentious, *yes*. Our model is able to express the same components and relations as the immunological models in Chapter 2, but goes further by theoretically predicting and empirically demonstrating the response behaviour when simulated with thousands of components and a changing antigenic environment. As a modelling strategy, we have attempted to determine sufficient immunological detail that statistically exhibits the correct behaviour, rather than reverse engineer the correct behaviour from a mountain of experimental observations of unknown significance. This is not an approach we expect to be readily accepted by immunologists, but when facing a phenomenon as exquisitely complex as the immune system, one has to ask: how much of that complexity is necessary to support and explain immunity and tolerance? This is a very different question than that faced by, for example, the designer of a targeted drug treatment; but still an important one. With respect to the latter, we now reflect on our contributions and omissions.

7.1 Contributions

Our thesis was that a statistical perspective offered insight and abstraction to immunologist and computer scientists alike. For clarity, we separate our discussion of how we have asserted this thesis into that based on immunological ideas and that based on computer science and statistics.

7.1.1 An immunological perspective

Lymphocyte ecology

It is a curious omission of the self/non-self models of Chapter 2 that none of them depend on antigen supply and demand. The assumption that such effects are negligible in establishing what is necessary or sufficient to produce a coherent response would seems to be highly questionable. What our approach loses in using surrogate data to quantify such interactions is, arguably, more than made up for by the demonstration that redundant competitive interactions are sufficient to produce qualitatively appropriate systemic behaviour. It is not clear from the literature, given that clonal selection was *"an attempt to apply the concepts of population genetics to the mesenchymal cells within the body"* [25], why the ecological view of the immune repertoire was not pursued further than it has been (see e.g. [51, 72, 110, 168]). The key manoeuvre that allowed us to build upon this basic principle was recognising that, once properly formulated, the generalised Lotka-Volterra steady-state provides a solution to the sparse approximation problem. Such a technical move is certainly not beyond the mathematical sophistication of the typical theoretical immunologist. But making such a connection possibly does benefit from the statistical perspective we have advocated.

The form of the generalised Lotka-Volterra model employed in this thesis is minimal. It contains the key dynamics but omits additional factors such as clonotype decay rates and immigration-emigration terms due to mutations. Introducing such factors would certainly improve the biological plausibility of our model, but they would also complicate the dynamics and subsequent analysis. The basic dynamical behaviour established here provides intuition that can be built upon.

Systemic dynamics and self/non-self semantics

Another omission from the self/non-self models in Chapter 2 was the apparent mutual exclusion between unambiguous response semantics and accounting for the evolutionary role of the adaptive immune system: the dichotomy between the "reductionist switch" and "holistic emergence" views. Our systemic response model offers one possible approach to resolving this dichotomy: the innate response may well provide feedback on peptide-class correlations; but this feedback may not be sufficient to produce non-trivial decision boundaries without further accounting for peptide-peptide correlations and the inverse relationship between representational capacity and discriminatory capacity. The statistical view makes these numerical aspects clear. It remains to be seen how relevant these aspects are biologically – but they are easier to digest and test than philosophical debate.

The main failure of our systemic model is that it is still a little too algebraic. The competition dynamics for $\varphi_i(x)$, $\varphi_i(G)$ and $\varphi_i(\tilde{f})$ occur independently of each other, which is acceptable if one is only interested the final steady-state values, but it would seem to be more plausible to have a compartmentalised model where e.g. a *naive* clone compartment competing over signal one flows into an *induced* clone compartment competing over signal two; and so on. How much the behaviour of such a model would deviate from that presented here is unclear, but such development could better assert validity as a *bona fide* biological model.

Shape, degeneracy and redundancy

In addition to the issues with the self/non-self discrimination models, the isotropic recognition volumes in shape-space were found to be unable to express receptor-ligand degeneracy and beneficial redundancy. Further, they were subject to a breakdown in intuitions about the properties of n-dimensional spaces that directly affects both the plausible size of the repertoire and the specificity of receptor-ligand binding. Under our formulation, ligands are not atomic entities: binding is a function of correlation in physical space, not distance in shape space. By recasting receptors and ligands as sub-spaces, rather than points covering shape-space, we were able to provide one possible formalisation of degeneracy: antigen will intersect with many, but not all, subspaces sensed by the φ_i; and conversely, each φ_i will intersect with many distinct antigen – as demanded by poly-clonality and poly-recognition, respectively. Similarly, the benefits of redundancy become apparent in this formalism as those same benefits afforded by overcomplete representations in the approximation setting. Crucially, we have not entirely abandoned the powerful conceptual tool of geometric thinking, only changed spaces.

The particular subspace implementation used in this thesis is more proof-of-concept than a mathematically sophisticated abstraction. In particular, the choice of 3 non-zero uniformly weighted components in each φ_i was arbitrary and simplistic. A more plausible formulation would account for the fact that if φ_i can bind with epitopes that include e.g. peptide "ARNDC" then it may also attribute some lesser affinity to those that include "ARNDG". These rules could satisfy known physico-chemical properties or more abstract binding relationships such as those used in the traditional shape-space. Similarly, the use of vector outer-products $|x\rangle\langle x|$ to represent surface correlation is limited and biologically naive. However, the strength of the abstraction is its generality. How the surface representation is best instantiated and then carved up is an open-ended question, that would benefit from deeper immunological and mathematical insight.

Constructive representations and co-respondence

It is a very elegant aspect of the applied statistics that this work draws upon, that the same approximation strategy has been applied to representation learning at one level, and decision making at another. By connecting this distinct research with the approximatory behaviour of competitive exclusion, we were able to gather a lot of seemingly diverse ideas into one very simple idea. We submit that this simple idea provides a quantifiable, formally malleable definition of the influential, but essentially rhetorical, ideas of immunologists such as Cohen and Varela. Thus, the immune system's self-constructed internal representation, and the integration of diverse, limited and contradictory components into a coherent systemic response, become two different perspectives of the same underlying phenomenon – the repertoire approximating its environment.

7.1.2 A computational perspective

The statistical approach our thesis advocates was invaluable in allowing us to make a clean transition between biology and computer science. This led to a contribution that does not depend on its biological inspiration to assert novelty.

Competitive exclusion as numerical method

In addition to being an elegant interpretation of the immunology, we have shown competitive exclusion to be a very successful algorithmic strategy. The primary benefit competitive exclusion has over existing iterative sparse approximation and boosting algorithms is that it relaxes the myopic, greedy nature of these algorithms. Competition automatically resolves redundancies and dependencies without significant *a priori* effort in the design of base components or the underlying algorithmic logic. Indeed, the beauty of this approach is that, once properly formulated, the dynamics cannot help but "do the right thing", even if the environment being approximated or the components of the approximation are continually changing. Formulated as a dynamical system it is inherently adaptive; but in a way that may be relied upon as an algorithm because its systemic behaviour is not entirely unpredictable. Ecological interpretation aside, this is a potentially elegant numerical method in its own right.

Like all numerical methods, competitive exclusion has some nuisance parameters that are difficult to provide a one-size-fits-all value for. In particular, there is no obvious value for a clonotype's initial population size. This problem is particularly acute in the case where one also wants to determine a minimum population size where a clonotype can be considered extinct and removed. Valid parameter values depend on the properties of both the signals and bases of the approximation problem – the acceptable value of coefficients that should not be considered noise. There is an additional trade-off between choosing an ϵ sufficiently small to determine stability, but sufficiently large to avoid glacial convergence times. Our experience is that approximation error converges much earlier than population levels; allowing one to choose quite a coarse ϵ without significantly degrading performance. From an algorithmic perspective, we are less concerned with a numerically accurate integration than, for example, a mathematical modeller might be, although some level of accuracy is necessary to ensure the dynamics are not corrupted by numerical instabilities. In practice, these issues are quite easy to identify and resolve by simply observing the dynamic evolution of the population, but a more principled approach would seem preferable.

Immune-inspired computing

We have been quite critical of immune-inspired computing throughout this thesis, bemoaning its lack of both statistical and biological insight. Of course, it is easier to criticise than to construct. We hope that the constructive aspects of this thesis, if not adopted or furthered by others, will at least inspire others to think outside of the stagnant "GA without crossover" paradigm that has dominated the field for almost 20 years. To be clear, this is not a criticism of Forrest and Perelson's seminal work. Quite the contrary, we would hope to see the field of artificial immune systems return to the interface between immunology and computation, where these ideas were born. Both immunology and computer science have changed dramatically since 1993. So should the interface between them.

That is not to say that previous work in immune-inspired computing is redundant. The most glaring omission from the work presented here is the reluctance to integrate *mutation*, and the subsequent affinity maturation, into the competitive exclusion dynamics of clonal selection. We have little reason to doubt that mutation would improve the efficacy as well as plausibility of our model. This is quite straight-forward to formalise (see e.g. [144, 173]) but is not very practical in terms of running simulations and algorithms. One alternative is to adopt the methods employed in evolutionary algorithms and more traditional artificial immune systems. This gives us a rich literature to draw from in implementing mutation, but leaves us with little theoretical formalism to fall back on. The statistical sampling perspective exploited in Chapter 4 may offer some compromise between these extremes, but requires a probabilistic reinterpretation of our algebraic model. This is a natural progression to make if one accepts the statistical perspective advocated here.

This is also not to say that our peers are not also addressing the imbalance in the biological and statistical efficacy of immune-inspired computing. Of particular interest at the time of writing are Owens' model of receptor binding and intracellular signalling [65] and Abi-Haidar and Rocha's implementation of Carneiro's cross-regulation model [1]. Both are based on biologically detailed models and have been shown to exhibit promising results when applied in the statistical inference domain. However, these models are isolated and not obviously understood as different parts of the same system. We have argued that the strength of our particular approach is that it allows us to clearly isolate the important factors in decision making: how they are organised and interact, and how they relate to immunological components and interactions. It seems possible that research such as Owens' and Abi-Haidar's could offer more biologically detailed and statistically stronger instantiations of ligand binding, inter-cellular learning and the dynamics of context, than the abstraction presented here. Such development would seem to be a step towards an *artificial immune system* that is in the same spirit as what the term conjures, at least in this author's mind.

7.2 Future Work

Disregarding incremental improvements and addressing highlighted omissions, we see several opportunities for development of the work introduced in this thesis that seem valid for both real and artificial immune systems:

- Integration of mutation and affinity maturation.
- Dynamics that incorporate how dependencies between environmental, target and contextual signals effect the evolution of clonotype concentrations.
- An explicit incorporation of antibody production, antigen clearance and their effect on the environmental and contextual signals.
- Development of T-Cells and antigen presenting cells beyond a simple scalar quantity and the inclusion of T-Cell receptor degeneracy.

With regard to our reluctance to implement mutation and the possible value of the probabilistic sampling framework developed in Chapter 4, it is well-known that the replicator equation (Eq. 4.7) and the Lotka-Volterra model (Eq. 5.1) have a formal connection (see e.g. [144]). Generalising the algebraic formulation in chapters 5 and 6 under the probabilistic setting of chapter 4 would seem the most profitable means toward developing a coherent theoretical foundation for the interface between immunology and computer science.

7.3 Concluding Remarks

From the right perspective, mechanistic descriptions of inferential behaviour are the essential stuff of immunological modelling and statistical numerical methods. What the latter provides the former is clarity on what are sufficient components and interactions. What the former provides the latter is insight into how inferential behaviour can be made autonomous. We assert that this statistical perspective *is* the interface between immunology and computer science.

No-one would knowingly attribute cognitive capacity to a predictive statistical model, though many critical decisions might be made on the basis of their predictions. Attributing such capacity to the immune system would be a similar mistake. Inference and prediction are not cognition: they can be entirely explained in terms of mechanistic reactions coupled to a higher-level process that sets up the conditions for these reactions. Each of the algorithms proposed in this thesis continuously iterate a highly simplified version of the mathematical and statistical modelling strategy: *generate components; fit them to the environment; and infer the state of nature as best as one can.* A cognate modeller may seek explanation and meaning, but predictive power is sufficient to confer survival advantage. That is what we have hypothesised the natural immune system to be capable of. That is what we have demonstrated existing artificial immune systems to be incapable of. That is what we have shown to be possible and effective, with only a basic ecological principle and a repertoire of redundant low quality components. That is our thesis.

Appendix A

Mathematical Background

The reader is directed to the textbooks [8, 154] for a deeper treatment of the material presented here. We first recall the definition of a vector space. A real[1] vector space $\mathcal{V}$ is a set endowed with two operations

1. Element addition: $u + v \in \mathcal{V} \quad \forall u, v \in \mathcal{V}$

2. Scalar multiplication: $\alpha v \in \mathcal{V} \quad \forall v \in \mathcal{V}, \alpha \in \mathbb{R}$

Addition is commutative and associative. $\mathcal{V}$ contains an additive identity 0 such that $u + 0 = u$ and an additive inverse: $\exists w \in \mathcal{V}$ such that $u + w = 0$. Essentially, addition works like normal. Scalar multiplication is distributive and also has an identity, 1. In general, our vector spaces will tend to be finite-dimensional Euclidean space $\mathbb{R}^n$, though this is easy to generalise if need be. We now recall some basic features of vector spaces:

1. A **subspace** of a vector space, is a subset $\mathcal{U} \subset \mathcal{V}$ of elements such that the conditions for a vector space still apply. For example, the subspaces of $\mathbb{R}^2$ are $\{0\}$, $\mathbb{R}^2$, and the set of all lines in $\mathbb{R}^2$ passing through the origin.

2. The set of all linear combinations $\alpha_1 v_1 + \ldots + \alpha_m v_m$ of m vectors $v_i \in \mathcal{V}$ is called the **span** of those vectors. It is a subspace of $\mathcal{V}$.

3. Any linearly independent set of vectors than span $\mathcal{V}$ are a **basis** for $\mathcal{V}$. Any $v \in \mathcal{V}$ can be uniquely represented as a linear combination of basis vectors.

To avoid digression we omit the definition of “linearly independent” which we will supplant later. The notion of *basis* (pl. *bases*) is pivotal in much of

[1] We will only be concerned with vector spaces over the real numbers, but the definition is much the same for complex numbers.

this thesis, so we elaborate briefly to clarify the idea for the unfamiliar reader. Consider an arbitrary vector in $\mathbb{R}^3$

$$v = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = x \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{A.1}$$

Here we have represented the vector v (uniquely) as a linear combination of the *standard* basis vectors. This apparent contrivance belies the fact that there is nothing special about the standard basis: other vectors may provide a more convenient basis. What is crucial to appreciate is that when the basis changes, so do the values of the coefficients x, y and z. It is still the same vector, but *represented* differently.

Inner product spaces and orthonormal bases

It is useful to extend our basic vector space with the notions of magnitude, distance and angle. We do this with an **inner product**

$$\langle v|u\rangle = \sum_{i=1}^{n} v_i \cdot u_i$$

Technically, this is the Euclidean inner product which, again, can be generalised. The inner product allows us to define vector length, or **norm**

$$\|v\|_2 = \langle v|v\rangle^{\frac{1}{2}}$$

which is simply an n-dimensional generalisation of Pythagoras' Theorem. We define the **Euclidean distance** between vectors u and v as $\|u - v\|_2$, that is, the norm of the difference vector. The **Triangle Inequality** $\|u + v\|_2 \leq \|u\|_2 + \|v\|_2$ holds for all u and v, making the Euclidean distance a *metric*. We can further define the angle between vectors u and v via

$$\cos(\theta_{u,v}) = \frac{\langle u|v\rangle}{\|u\|_2 \cdot \|v\|_2} = \left\langle \frac{u}{\|u\|_2} | \frac{v}{\|v\|_2} \right\rangle = \langle \tilde{u}|\tilde{v}\rangle$$

where the right hand side emphasises that this is simply the inner product of **normalised** vectors, $\|\tilde{u}\|_2 = \|\tilde{v}\|_2 = 1$. Note that normalisation changes length, but not direction. The fact that θ is well-defined follows from the **Cauchy-Schwarz Inequality** $|\langle u|v\rangle| \leq \|u\|_2 \cdot \|v\|_2$ and noting that $\langle \tilde{u}|\tilde{u}\rangle = 1$, $\langle \tilde{u}|\widetilde{-u}\rangle = -1$ and $\langle \tilde{u}|\tilde{v}\rangle = 0$ when u and v are **orthogonal**. In the 2-dimensional plane defined by u and v this is the same as cos(0), cos(180) and cos(90), respectively.

If a collection of normalised vectors $\tilde{u_i}$ are pairwise orthogonal, they are **orthonormal**. Orthogonality is sufficient for the $\tilde{u_i}$ to be a basis for the vector space they span. A desirable property of such bases is that the inner product recovers the coefficients of an arbitrary vector v represented in that basis, say $\tilde{v}$

$$\tilde{v} = \sum_{i=1}^{n} \langle u_i | v \rangle \, u_i$$

It is apparent that Eq. (A.1) is a special case of this formula. A significant part of this thesis is based around what can be done when it is undesirable to use an orthonormal basis to represent vectors and the above convenience is lost.

Low rank approximation and inversion

The **eigen decomposition** is a fundamental operation in applied and theoretical mathematics. Any square symmetric $m \times m$ matrix M can be decomposed into a sum of matrices formed from the **eigen vectors** φ_i and their **eigen values** λ_i,

$$M = \Phi \Lambda \Phi' = \sum_{i=1}^{m} \lambda_i |\varphi_i\rangle\langle\varphi_i|,$$

The eigen decomposition is the only decomposition that diagonalises M (i.e. Λ is a diagonal matrix), decoupling all of the φ_i. The eigen vectors form an orthonormal basis for the subspace spanned by M. One consequence of this is that, assuming the eigen vectors are ordered in decreasing magnitude of eigen value, the optimal rank $k < m$ approximation to M is

$$\tilde{M}_k = \Phi_k \Lambda_k \Phi_k' = \sum_{i=1}^{k} \lambda_i |\varphi_i\rangle\langle\varphi_i| \approx M,$$

where optimality is defined in terms of squared error $\|M - \tilde{M}_k\|_2$. Another consequence is that

$$M^{-1} = \Phi \Lambda^{-1} \Phi' = \sum_{i=1}^{m} \frac{1}{\lambda_i} |\varphi_i\rangle\langle\varphi_i|,$$

and thus, the eigen-decomposition is a method of inverting a square symmetric matrix when $\lambda_i > 0 \ \forall i$. That is, M is invertible when $\lambda_i > 0 \ \forall i$. That this is an inverse follows from $\mathbb{I} = M^{-1}M = \Phi\Lambda^{-1}\Phi'\Phi\Lambda\Phi' = \Phi\Phi' = \Phi\Phi^{-1}$. If some $\lambda_i = 0$ we can, as with approximation, consider a rank $k < m$ inversion using only (a ranked subset of) the eigen vectors with non-zero eigen values

$$\tilde{M}_k^{-1} = \Phi_k \Lambda_k^{-1} \Phi_k' = \sum_{i=1}^{k} \frac{1}{\lambda_i} |\varphi_i\rangle\langle\varphi_i| \approx M^{-1}$$

The generalisation of the eigen decomposition to an arbitrary $n \times m$ matrix X is the **Singular Value Decomposition**

$$X = USV' = \sum_{i=1}^{k} s_i |u_i\rangle\langle v_i'|$$

where $k \leq \min(m, n)$ is the **rank** of X, U are the eigenvectors of the square symmetric matrix XX', V are the eigenvectors of the square symmetric matrix $X'X$ and $S = \mathrm{sqrt}(\Lambda)$, where Λ is a diagonal matrix of eigen values shared between both decompositions. Similar to the eigen decomposition,

$$X^{-1} = VS^{-1}U' = \sum_{i=1}^{k} \frac{1}{s_i} |v_i\rangle\langle u_i'| \tag{A.2}$$

From Eq. (A.2) we can derive the identities, $U' = S^{-1}V'X'$ and $V = X'US^{-1}$. Plugging these back into Eq. (A.2), we note that

$$\begin{aligned} X^{-1} &= VS^{-1}U' \\ &= (X'US^{-1})S^{-1}U' \\ &= VS^{-1}(S^{-1}V'X') \end{aligned}$$

and thus, by associativity of multiplication, $X^{-1} = (V\Lambda_k^{-1}V')X' = X'(U\Lambda_k^{-1}U')$. Recalling our earlier presentation of the eigen decomposition and the fact that U and V are the eigen vectors of XX' and $X'X$, we conclude that $(X^{-1})' = (XX')_k^{-1}X = X(X'X)_k^{-1}$, where k denotes a low rank inversion, as defined above. This is the well-known **Moore-Penrose pseudo inverse** X^+, although the presentation given is non-classical. In Eq. (3.7) and (3.8), we make use of this row/column-space duality in order to elucidate a connection between the least squares and k-nearest neighbour solutions to the statistical inference problem. In Eq. (6.6), the expansion of $(X^+)'\theta = (U\Lambda_k U')X\theta$ is central to the statistical justification of our systemic model of the immune response.

Bibliography

[1] A. Abi-Haidar and L. Rocha. Biomedical article classification using an agent-based model of T-cell cross-regulation. In *Artificial Immune Systems*, pages 237–249. Springer, 2010.

[2] Alaa Abi-haidar and Luis M Rocha. Adaptive Spam Detection Inspired by the Immune System. *Artificial Life*, 2008.

[3] C.C. Aggarwal, A. Hinneburg, and D.A. Keim. On the surprising behavior of distance metrics in high dimensional spaces. *Lecture Notes in Computer Science*, 1973:420 – 434, 2000.

[4] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. *Lecture Notes in Computer Science*, 1973:420–??, 2001.

[5] M Aharon, M Elad, and a Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear Algebra and its Applications*, 416(1):48–67, 2006.

[6] Colin Anderson and Polly Matzinger. Danger: the view from the bottom of the cliff. *Seminars in Immunology*, 12:231–238, 2000.

[7] Paul Andrews and Jon Timmis. Adaptable Lymphocytes for Artificial Immune Systems. In *Artificial Immune Systems*, pages 376–386. Springer, 2008.

[8] S.J. Axler. *Linear algebra done right.* Springer Verlag, 1997.

[9] A.L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[10] Nicholas J. Belkin and W. Bruce Croft. Information filtering and information retrieval: two sides of the same coin? *Commun. ACM*, 35(12):29–38, December 1992.

[11] R Bellman. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, 1961.

[12] Richard Bellman. *Introduction to Matrix Analysis.* SIAM Classics, 1997.

[13] H. Bersini. Immune network and adaptive control. In *Toward a practice of autonomous systems, Proceedings of the First European Conference on Artificial Life*, pages 217–226, 1992.

[14] H. Bersini. Reinforcement and recruitment learning for adaptive process control. In *Proc. Int. Fuzzy Association Conference (IFAC/IFIP/IMACS) on Artificial Intelligence in Real Time Control*, pages 331–337, 1992.

[15] H. Bersini. *Artificial Immune Systems and Their Applications*, book chapter/section The Endoge. Springer-Verlag, 1999.

[16] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When Is "Nearest Neighbor" Meaningful? *Lecture Notes in Computer Science*, 1540:217–235, 1999.

[17] G. B. Bezerra, T. V. Barra, L. N. de Castro, and F. J. Von Zuben. Adaptive radius immune algorithm for data clustering. In *4th internation conference on artificial immune systems*, pages 290–303. Springer, 2005.

[18] J A Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4, 1998.

[19] Anders Bjorkstrom. Ridge regression and inverse problems, 2001.

[20] Léon Bottou. Stochastic Learning. In Olivier Bousquet and Ulrike von Luxburg, editors, *Advanced Lectures on Machine Learning*, Lecture Notes in Artificial Intelligence, LNAI˜3176, pages 146–168. Springer Verlag, Berlin, 2004.

[21] Leo Breiman. Prediction Games and Arcing Algorithms. *Neural Comp.*, 11(7):1493–1517, October 1999.

[22] P. A. Bretscher and M. Cohn. A theory of self-nonself discrimination: paralysis and induction involve the recognition of one and two determinants on an antigen, respectively. *Science*, (169):1042–1049, 1970.

[23] P. Bühlmann and B. Yu. Boosting with the L2 Loss: Regression and Classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

[24] Andreas Buja and Werner Stuetzle. Response to Mease and Wyner, Evidence Contrary to the Statistical View of Boosting, JMLR 9:131–156, 2008. *Journal of Machine Learning Research*, 9, 2008.

[25] F. M. Burnet. A modification of Jerne's theory of antibody production using the concept of clonal selection. *The Australian Journal of Science*, 3:67–69, 1957.

[26] S. L. Cambell, J. P. Chancelier, and R. Nikoukhah. *Modelling and simulation in Scilab/Scicos*. Springer, 2006.

[27] J Carneiro, T Paixao, D Milutinovic, J Sousa, K Leon, R Gardner, and J Faro. Immunological self-tolerance: Lessons from mathematical modeling. *Journal of Computational and Applied Mathematics*, 184(1):77–100, 2005.

[28] Jorge Carneiro. *Towards a Comprehensive View of the Immune System*. Ph.D thesis, 1997.

[29] Jorge Carneiro, Antonio Coutinho, Jose Faro, and John Stewart. A Model of the Immune Network with B-T Cell Co-operation. I - Prototypical Structures and Dynamics. *Journal of Theoretical Biology*, 182:513–529, 1996.

[30] Jorge Carneiro, Antonio Coutinho, and John Stewart. A Model of the Immune Network with B-T Cell Co-operation. II - The Simulation of Ontogenisis. *Journal of Theoretical Biology*, 182:531–547, 1996.

[31] Jorge Carneiro, Kalet Leon, Iris Caramalho, Carline van den Dool, Rui Gardner, Vanessa Oliveira, Marie-Louise Bergman, Nuno Sepúlveda, Tiago Paixão, Jose Faro, and Jocelyne Demengeot. When three is not a crowd: a Crossregulation model of the dynamics and repertoire selection of regulatory CD4+ T cells. *Immunological reviews*, 216:48–68, April 2007.

[32] Jorge Carneiro and John Stewart. Rethinking Shape Space: Evidence from simulated docking suggests that steric shape complementarity is not limiting for antibody-antigen recognition and idiotypic interactions. *J.Theor.Biol*, 169:391–402, 1994.

[33] F. Castiglione, S. Motta, and G. Nicosia. Pattern Recognition by primary and secondary response of an Artificial Immune System. *Theory in Biosciences*, 2(120):93–106, 2001.

[34] G. Celeux, D. Chauveau, and J. Diebolt. On stochastic versions of the EM algorithm, 1995.

[35] K.M. Cheman. *Optimization Techniques for Solving Basis Pursuit Problems.* Masters, 2006.

[36] Scott Shaobing Chen, David L. Donoho, and Michael a. Saunders. Atomic Decomposition by Basis Pursuit. *SIAM Review*, 43(1):129, 2001.

[37] Ole Christensen. *Frames and Bases: An Introductory Course (Applied and Numerical Harmonic Analysis).* Birkhauser, illustrate edition, July 2008.

[38] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4), 2009.

[39] I R Cohen. Immune system computation and the immunological homunculus. In O Niestrasz Et al., editor, *MoDELS 2006*, pages 499–512, 2006.

[40] IR Cohen. The cognitive paradigm and the immunological homunculus. *Immunology Today*, 13(12):490, 1992.

[41] I.R. Cohen. The cognitive principle challenges clonal selection. *Immunology Today*, 13:441–441, 1992.

[42] Irun R Cohen. *Tending Adam's Garden: Evolving the Cognitive Immune Self.* Academic Press, 2004.

[43] Irun R Cohen. Real and artificial immune systems: computing the state of the body. *Group*, 7(July):569–574, 2007.

[44] Irun R. Cohen and Lee A. Segel. *Design Principles for the Immune System and Other Distributed Autonomous Systems.* Oxford University Press Inc, 2001.

[45] Antonio Coutinho and Werner Haas. In vivo models of dominant tolerance: where do we stand today? *TRENDS in immunology*, 22(7):350–351, 2001.

[46] V. Cutello, N. Krasnogor, G. Nicosia, and M. Pavone. Immune Algorithm Versus Differential Evolution: A Comparative Case Study Using High Dimensional Function Optimization. *Adaptive and Natural Computing Algorithms*, pages 93–101, 2007.

[47] V. Cutello and G. Nicosia. An Immunological Approach to Combinatorial Optimization Problems. In *Advances in Artificial Intelligence*, pages 361–370. Springer, 2002.

[48] V. Cutello, G. Nicosia, and M. Pavone. Real coded clonal selection algorithm for unconstrained global optimization using a hybrid inversely proportional hypermutation operator. In *Proceedings of the 2006 ACM symposium on Applied computing*, page 954. ACM, 2006.

[49] Vincenzo Cutello, Giuseppe Nicosia, Mario Romeo, and Pietro S. Oliveto. On the Convergence of Immune Algorithms. *2007 IEEE Symposium on Foundations of Computational Intelligence*, pages 409–415, April 2007.

[50] Neil Dalchai, Andrew Phillips, Leonard D. Goldstein, Mark Howarth, Luca Cardelli, Tim Elliot, and Joern M. Werner. A kinetic model for predicting MHC class I presentation of competing peptides. *PNAS (in press)*, 2010.

[51] Rob J. De Boer and Alan S. Perelson. T Cell Repertoires and Competitive Exclusion. *Journal of Theoretical Biology*, 169:375–390, 1994.

[52] L N de Castro and J Timmis. An artificial immune network for multimodal function optimization. In *Proceedings of the 2002 congress on evolutionary computation, CEC*, volume 2, pages 12–17, 2002.

[53] L N de Castro and F J Von Zuben. aiNet: An Artificial Immune Network for Data Analysis. In A Abbass, R A Sarker, and C S Newton, editors, *Data Mining: A Heuristic Approach*, pages 231–259. Idea Group Publishing, 2001.

[54] L N De Castro, F J Von Zuben, and Others. Learning and optimization using the clonal selection principle. *IEEE Transactions on Evolutionary Computation*, 6(3):239–251, 2002.

[55] Leandro N De Castro and Jonathan Timmis. *Artificial Immune Systems: A New Computational Intelligence Approach.* Springer Verlag London, 2002.

[56] A.P. Dempster, N.M. Laird, D.B. Rubin, and Others. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[57] Vincent Detours, Hugues Bersini, John Stewart, and Francisco Varela. Development of an Idiotypic Network in Shape Space. *Journal of Theoretical Biology*, 170(4):401–414, October 1994.

[58] David L Donoho. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. *Statistics*, pages 1–33, 2000.

[59] David L Donoho. For Most Large Underdetermined Systems of Linear Equations the Minimal 1 -norm Solution is also the Sparsest Solution. *Statistics*, 40698:1–28, 2004.

[60] D.L. Donoho, I. Drori, Y. Tsaig, and J.L. Starck. Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit, 2006.

[61] G. M. Edelman and J. A. Gally. Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24):13763–13768, 2001.

[62] Gerald M Edelman and Joseph A Gally. Degeneracy and complexity in biological systems. *Most*, 98(24):13763–13768, 2001.

[63] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–451, 2004.

[64] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

[65] D. Estimation. T Cell Receptor Signalling Inspired Kernel Density Estimation and Anomaly Detection. In *Artificial Immune Systems: 8th International Conference, ICARIS 2009, York, UK, August 9-12, 2009, Proceedings*, page 122. Springer-Verlag New York Inc, 2009.

[66] J. D. Farmer, N. H. Packard, and A. S. Perelson. The Immune System, Adaptation and Machine Learning. *Physica*, 22:187–204, 1986.

[67] R. Ferrer, I. Cancho, and R. Sole. The small-world of human language. In *Proceedings of the Royal Society B: Biological Science*, 2001.

[68] Darren R. Flower. Towards in silico prediction of immunogenic epitopes. *Trends in Immunology*, 24(12), 2003.

[69] Sally Floyd and Manfred Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, December 1995.

[70] S. Forrest, B. Javornik, R.E. Smith, and A.S. Perelson. Using genetic algorithms to explore pattern recognition in the immune system. *Evolutionary computation*, 1(3):191–211, 1993.

[71] A.A. Freitas and J. Timmis. Revisiting the Foundations of Artificial Immune Systems for Data Mining. *Evolutionary Computation, IEEE Transactions on*, 11(4):521–540, 2007.

[72] Antonio A. Freitas and Benedita Rocha. Population Biology of Lymphocytes: The Flight for Survival. *Annual Review of Immunology*, 18:83–111, 2000.

[73] Yoav Freund and Robert E. Schapire. A Decision theoretic Generalisation of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[74] Yoav Freund, Robert E Schapire, and Murray Hill. Game Theory, On-line Prediction and Boosting. *Learning*, 1996.

[75] J Friedman, T Hastie, and R Tibshirani. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

[76] J H Friedman and J W Tukey. A Projection Pursuit Algorithm for Exploratory Data Analysis. *Computers, IEEE Transactions on*, C-23(9):881–890, 1974.

[77] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

[78] Jerome H Friedman. Recent Advances in Predictive (Machine) Learning. In *PHYSTAT2003*, 2003.

[79] J.H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, (2):1189–1232, 2001.

[80] Mark R. Gardner and Ross W. Ashby. Connectance of Large Dynamic (Cybernetic) Systems: Critical Values for Stability. *Nature*, 228, 1970.

[81] Neil Gershenfeld. *The naturwe of mathematical modelling*. Cambridge University Press, 1999.

[82] T Gisiger. Scale invariance in biology: coincidence or footprint of a universal mechanism? *Biological reviews of the Cambridge Philosophical Society*, 76(2):161–209, May 2001.

[83] D. Goodman, L. Boggess, and A. Watkins. Artificial Immune System classification of multiple-class problems, 2002.

[84] D.E. Goodman, L. Boggess, and A. Watkins. An investigation into the source of power for AIRS, an artificial immune classification system. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 3, pages 1678–1683 vol.3, 2003.

[85] Jason A Greenbaum, Pernille Haste Andersen, Martin Blythe, Huynh-hoa Bui, Raul E Cachau, James Crowe, Matthew Davies, A S Kolaskar, Ole Lund, Sherrie Morrison, Brendan Mumey, Yanay Ofran, Jean-luc Pellequer, Clemencia Pinilla, Julia V Ponomarenko, G P S Raghava, Marc H V Van Regenmortel, Erwin L Roggen, Alessandro Sette, Avner Schlessinger, Johannes Sollner, Martin Zand, and Bjoern Peters. Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *Journal of Molecular Recognition*, 2007.

[86] U. Hanani, B. Shapira, and P. Shoval. Information filtering: Overview of issues, research and systems. *User Modeling and User-Adapted Interaction*, 11(3):203–259, 2001.

[87] Emma Hart and Jonathan Timmis. Application Areas of AIS: The Past, the Present and the Future. 2005.

[88] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[89] U Hershberg, S Solomon, and I R Cohen. What is the basis of the immune system's specificity? In V.Capasso, editor, *Mathematical Modelling & Computing in Biology and Medicine*, pages 377–384, 2003.

[90] A. Hinneburg, C.C. Aggarwal, and D.A. Keim. What is the nearest neighbor in high dimensional spaces. *Proceedings of the 26th International Conference on Very Large Data Bases*, pages 506–515, 2000.

[91] Moira Howes. Self, intentionality and immunological explanation. *Seminars in Immunology*, 12:249–256, 2000.

[92] Edith I. R. Cohen. Real and artificial immune systems: computing the state of the body. *Nature Reviews Immunology*, 7:569–574, 2007.

[93] Charles A. Janeway and Ruslan Medzhitov. Innate Immune Recognition. *Annual Reviews Immunology*, 20:197–216, 2002.

[94] Charles A Janeway, Paul Travers, Mark Walport, and Mark Schlomchik. *Immunobiology: the immune system in health and disease*. Garland, 2001.

[95] W. Jank. Stochastic variants of EM: Monte Carlo, quasi-Monte Carlo and more. In *Proceedings of the American Statistical Association*, 2005.

[96] Vincent A. A. Jansen and Giorgos D. Kokkoris. Complexity and stability revisited. *Ecology Letters*, 6:498–502, 2003.

[97] E. T. Jaynes. *Probability theory: the logic of science*. Cambridge University Press, 2003.

[98] N. K. Jerne. Towards a network theory of the immune system. *Annals d'immunologie*, 1974.

[99] Niels K. Jerne. The Generative Grammer of the Immune System. *Nobel Lecture*, 1984.

[100] T. Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Machine Learning International Workshop*, pages 143–151, 1997.

[101] Michael J Kearns and Umesh V Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.

[102] Urmila Kulkarni-Kale, Shriram Bhosle, and A. S. Kolaskar. CEP: a conformational epitope prediction server. *Nucleic Acids Research*, 33, 2005.

[103] KJ Lafferty and AJ Cunningham. A new analysis of allogeneic interactions. *Immunology and Cell Biology*, 1975.

[104] Nicole Le Douarin, Catherine Corbel, Antonio Bandeira, Veronique Thomas-Vaslin, Yves Modigliani, Antonio Coutinho, and Josselyne Salaun. Evidence for a thymus-dependent form of tolerance that is not based on elimination or anergy of reactive T cells. *Immunological Reviews*, 149, 1996.

[105] P. Leder. The genetics of antibody diversity. *Scientific American*, 246:102–115, 1982.

[106] Heung Kyu Lee and Akiko Iwasaki. Innate control of adaptive immunity: Dendritic cells and beyond. *Seminars in Immunology*, 19(1):48–55, February 2007.

[107] Heung Kyu Lee and Akiko Iwasaki. Innate control of adaptive immunity: dendritic cells and beyond. *Seminars in Immunology*, 19:48–55, 2007.

[108] K Léon. *A quantitative analysis of dominant tolerance*. Phd, University of Porto, 2002.

[109] K Leon, J Carneiro, R Perez, E Montero, and A Lage. Natural and Induced Tolerance in an Immune Network Model. *Journal of Theoretical Biology*, 193:519–534, 1998.

[110] Kalet Leon, Rolando Perez, Agustin Lage, and Jorge Carneiro. Modelling T-Cell-Mediated Suppression Dependent on Interactions in Multicellular Conjugates. *Journal of Theoretical Biology*, 207:231–254, 2000.

[111] R Levins. *Evolution in changing environments*. Princeton University Press, 1968.

[112] Richard Levins. The strategy of model building in population biology. *American Scientist*, 54(4), 1966.

[113] Dajeong Lim, Hee-Seok Oh, and Heebal Kim. Theoretical peptide mass distribution in the non-redundant protein database of the NCBI. *Genomics & Informatics*, 4(2):65–70, 2006.

[114] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

[115] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261, February 1994.

[116] A.J. Lotka. Analytical note on certain rhythmic relations in organic systems. *Proceedings of the National Academy of Sciences of the United States of America*, 6(7):410, 1920.

[117] Nicholas M Luscombe, Jiang Qian, Zhaolei Zhang, Ted Johnson, and Mark Gerstein. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome biology*, 3(8), July 2002.

[118] Roman Werner Lutz and Markus Kalisch. Robustified L2 boosting. *Computational Statistics*, 52:3331–3341, 2008.

[119] J. MacQueen. Some methods for classification and analysis of multivariate observations. volume 1, pages 281–297, 1967.

[120] S.G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[121] Christopher D. Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing.* MIT Press, 2003.

[122] Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent in function space. In *Proc. NIPS*, volume 12, pages 512–518. Citeseer, 1999.

[123] Humbert R. Maturana and Francisco J. Varela. *Autopoiesis and Cognition: The Realization of the Living.* Kluwer Academic Publishers, 1979.

[124] P. Matzinger. Tolerance, danger, and the extended family. *Annual review of immunology*, 12(1):991–1045, 1994.

[125] Polly Matzinger. An innate sense of danger. *Seminars in Immunology*, 10:399–415, 1998.

[126] Polly Matzinger. The danger model in its historic context. *Scandinavian Journal of Immunology*, 54:4–9, 2001.

[127] Polly Matzinger. The Danger Model: A Renewed Sense of Self. *Science*, 296(April):301–305, 2002.

[128] Robert M. May. Will a large complex system be stable? *Nature*, 238, 1972.

[129] Robert M. May and F. George Oster. Bifurcations and Dynamic Complexity in Simple Ecological Models. *The American Naturalist*, 110(974):573–599, 1976.

[130] Christopher H McEwan and Emma Hart. On Clonal Selection. *Theoretical Computer Science*, (In press, Accepted Manuscript), 2010.

[131] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions.* Wiley-Interscience, 1 edition, November 1996.

[132] David Mease and Abraham Wyner. Evidence Contrary to the Statistical View of Boosting. *J. Mach. Learn. Res.*, 9:131–156, 2008.

[133] Ruslan Medzhitov and Charles A Janeway. Innate Immunity: The Virtues of a Nonclonal System of Recognition. *Cell*, 91:295–298, 1997.

[134] M Mendao, J Timmis, P S Andrews, and M Davies. The Immune System in Pieces: Computational Lessons from Degeneracyin the Immune System. In *Foundations of Computational Intelligence (FOCI 2007)*, 2007.

[135] M. Mendao, J. Timmis, P.S. Andrews, and M. Davies. The Immune System in Pieces: Computational Lessons from Degeneracy in the Immune System. In *Foundations of Computational Intelligence, 2007. FOCI 2007. IEEE Symposium on*, pages 394–400, 2007.

[136] N Nanas and A De Roeck. Autopoiesis, the immune system, and adaptive information filtering. *Natural Computing*, 2009.

[137] N. Nanas, M. Vavalis, and L. Kellis. Immune learning in a dynamic Information Environment. In P. Andrews, J. Timmis, E. Hart, U. Aickelin, A. Hone, and A. Tyrrell, editors, *Artificial immune systems*, pages 192–205. Springer, 2009.

[138] Nikolaos Nanas, Anne De Roeck, and Manolis Vavalis. What Happened to Content-Based Information Filtering? *Advances in Information Retrieval*, pages 1–13, 2010.

[139] Nikoloas Nanas, Victoria S. Uren, and Anne de Roeck. Nootropia: A User Profiling Model Based on a Self-Organising Term Network. In *Artificial Immune Systems*, 2004.

[140] B. K. Natarajan. Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing*, 24(2):227, 1995.

[141] J. Newborough and S. Stepney. A generic framework for population-based algorithms, implemented on multiple FPGAs. In *Artificial Immune Systems*, pages 43–55. Springer, 2005.

[142] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351, 2006.

[143] Jerzy Neyman and Egon Pearson. On The Problem Of The Most Efficient Tests Of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, pages 289–337, 1933.

[144] Martin A. Nowak. *Evolutionary Dynamics: Exploring the Equations of Life.* Belknap Press of Harvard University Press, September 2006.

[145] K Page and M Nowak. Unifying Evolutionary Dynamics. *Journal of Theoretical Biology*, 219(1):93–98, November 2002.

[146] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40–44 vol.1, 1993.

[147] A S Perelson and G Oster. Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self non-self discrimination. *Journal of Theoretical Biology*, 81:645–670, 1979.

[148] Alan S Perelson and Gerard Weisbuch. Immunology for Physicists. *Review of Modern Physics*, 69, 1997.

[149] B. Peters, J. Sidney, P. Bourne, H-H. Bui, and S. Buus. No Title. *PLoS Biology*, 3(3), 2005.

[150] Simone Pigolotti, Cristobal Lopez, and Emilio Hernandez-Garcia. Species clustering in competitive Lotka-Volterra models. *Physics Review Letters*, 98(25), 2007.

[151] A. Roberts. The stability of a feasible random ecosystem. *Nature*, 251:607–608, 1974.

[152] J.J. Rocchio. *Relevance feedback in information retrieval*, pages 313—-323. 1971.

[153] Ian D. Rozdilsky and Lewi Stone. Complexity can enhance stability in competitive systems. *Ecology Letters*, 4:397–400, 2001.

[154] L.A. Sadun. *Applied linear algebra: the decoupling principle.* Amer Mathematical Society, 2008.

[155] G Salton and C Buckley. Improving retrieval performance by relevance feedback. *Journal of the American society for information science*, 41(4):288–297, 1990.

[156] Robert E. Schapire. The Strength of Weak Learnability. *Machine Learning*, 5:197–227, 1990.

[157] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *Computational Learning Theory*. Springer-Verlag, 2001.

[158] Bernhard Schölkopf, Alexander J. Smola, and Klaus R. Muller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem, 1996.

[159] Cosma Rohilla Shalizi. Dynamics of Bayesian Updating with Dependent Data and Misspecified Models. *Electronic Journal of Statistics, In press*, 2009.

[160] John Shaw-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2006.

[161] Arthur M. Silverstein. The clonal selection theory: what it really is and why modern challenges are misplaced. *Nature Immunology*, 3:793–796, 2002.

[162] Marina Skurichina and Robert P. W. Duin. Bagging, Boosting and the Random Subspace Method for Linear Classifiers. *Pattern Analysis & Applications*, 5(2):121–135, June 2002.

[163] Derek J. Smith, Stephanie Forrest, Ron R. Hightower, and Alan S. Perelson. Deriving shape-space parameters from Immunological Data. *Journal of Theoretical Biology*, 189(2):141–150, 1997.

[164] Lauren Sompayrac. *How the immune system works*. Blackwell Publishing, 2003.

[165] John Stewart and Antonio Coutinho. The Affirmation of Self: A New Perspective on the Immune System. *Artificial Life*, 10:261–276, 2004.

[166] T. Stibor and J. Timmis. An Investigation on the Compression Quality of aiNet. In *Foundations of Computational Intelligence, 2007. FOCI 2007. IEEE Symposium on*, pages 495–502, 2007.

[167] Thomas Stibor. *On the Appropriateness of Negative Selection for Anomaly Detection and Network Intrusion Detection*. Phd, Technische Universität Darmstadt, 2006.

[168] E.R. Stirk, C. Molina-Paris, and H.A. Van Den Berg. Stochastic niche structure and diversity maintenance in the T cell repertoire. *Journal of Theoretical Biology*, 255:237–249, 2008.

[169] Jing Sun, Di Wu, Tianlei Xu, Xiaojing Wang, Xiaolian Xu, Lin Tao, Y. X. Li, and Z. W. Cao. SEPPA: a computational server for spatial epitope prediction of protein antigens. *Nucleic Acids Research*, 37(2), 2009.

[170] Alfred Tauber. Moving beyond the immune self? *Seminars in Immunology*, 12:241–248, 2000.

[171] Alfred Tauber. *The biological notion of self and non-self.* 2006.

[172] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[173] J Timmis, A Hone, T Stibor, and E Clark. Theoretical advances in artificial immune systems. *Theoretical Computer Science*, 403(1):11–32, August 2008.

[174] Giulio Tononi, Olaf Sporns, and Gerald M. Edelman. Measures of degeneracy and redundancy in biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 96(6):3257–3262, March 1999.

[175] Ken Trogonning and Alan Roberts. Complex systems which evolve towards homeostasis. *Nature*, 281, 1979.

[176] Joel A Tropp. Just relax: convex programming methods for subset selection and sparse approximation, 2004.

[177] J.W. Tukey. *Exploratory data analysis.* Addison-Wesley, 1977.

[178] F. Varela, A. Coutinho, B. Dupire, and N. M. Vaz. *Theoretical Immunology, vol. II*, book chapter/section Cognitive. Addison-Wesley, 1988.

[179] Francisco J. Varela and Antonio Coutinho. Second generation immune networks. *Immunology Today*, 12(5):159–166, 1991.

[180] R. Violato, A. Azzolini, and F. J. Von Zuben. Antibodies with adaptive radius as prototypes of high-dimensional datasets. In *9th International conference on artificial immune systems*, pages 158–170, 2010.

[181] V. Volterra. Variations and fluctuations of the number of individuals in animal species living together. *ICES Journal of Marine Science*, 3(1):3, 1928.

[182] G. Wahba. *Spline Models for Observational Data.* SIAM, 1990.

[183] Andrew Watkins, Jon Timmis, and Lois Boggess. Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm. *Genetic Programming and Evolvable Machines*, 5(3):291–317, 2004.

[184] Andrew Watkins, Jon Timmis, and Lois Boggess. Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm. *Genetic Programming and Evolvable Machines*, 5(3):291–317, September 2004.

[185] Andrew B. Watkins. *AIRS: A resource limited artificial immune classifier.* Thesis (master's), 2001.

[186] Andrew B. Watkins. *Exploiting Immunological Metaphors in the Development of Serial, Parallel, and Distributed Learning Algorithms.* Thesis (phd), 2005.

[187] J. Whitacre and A. Bender. Degeneracy: a link between evolvability, robustness and complexity in biological systems, 2009.

[188] Darcy B Wilson, Dianne H Wilson, Kim Schroder, Clemencia Pinilla, Sylvie Blondelle, Richard A Houghten, and Christopher C Garcia. Specificity and degeneracy of T cells. *Molecular immunology*, 40(14-15):1047–1055, February 2004.

[189] K.W. Wucherpfennig, P.M. Allen, F. Celada, I.R. Cohen, R. De Boer, K.C. Garcia, B. Goldstein, R. Greenspan, D. Hafler, P. Hodgkin, and Others. Polyspecificity of T cell and B cell receptor recognition. *Seminars in immunology*, 19(4):216–224, 2007.

[190] G. A. Young and R. L. Smith. *Essentials of Statistical Inference.* Cambridge University Press, 2005.

[191] Xiaojin Zhu and Andrew B. Goldberg. Introduction to Semi-Supervised Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130, January 2009.

[192] GK Zipf. *Human behavior and the principle of least effort: An introduction to human ecology.* Addison-Wesley Press, 1949.