A Robbins-Monro type learning algorithm for an entropy maximizing version of stochastic Optimality Theory

Markus Fischer

August 4, 2005

Abstract

The object of the present work is the analysis of the convergence behaviour of a learning algorithm for grammars belonging to a special version – the maximum entropy version – of stochastic Optimality Theory. Stochastic Optimality Theory is like its deterministic predecessor, namely Optimality Theory as introduced by Prince and Smolensky, in that both are theories of universal grammar in the sense of generative linguistics.

We give formal definitions of basic notions of stochastic Optimality Theory and introduce the learning problem as it appears in this context. A by now popular version of stochastic Optimality Theory is the one introduced by Boersma, which we briefly discuss. The maximum entropy version of stochastic Optimality Theory is derived in great generality from fundamental principles of information theory. The main part of this work is dedicated to the analysis of a learning algorithm proposed by Jäger (2003) for maximum entropy grammars. We show in which sense and under what conditions the algorithm converges. Connections with well known procedures and classical results are made explicit.

The present work is a slightly modified version of my Master's thesis, which was submitted to the Department of German Language and Linguistics at Humboldt University Berlin in June 2005. The thesis was supervised by Prof. Manfred Krifka and Prof. Gerhard Jäger.

Author's address: Markus Fischer
Reichenberger Str. 166
10999 Berlin
Germany
E-Mail: markus.fischer@alumni.hu-berlin.de

Zusammenfassung

Gegenstand der vorliegenden Arbeit ist die Analyse des Konvergenzverhaltens eines Lernalgorithmus für Grammatiken, die einer speziellen Version, der Maximum-Entropie-Version, stochastischer Optimalitätstheorie angehören. Stochastische Optimalitätstheorie ist wie ihr deterministisches Vorbild, die von Prince und Smolensky eingeführte Optimalitätstheorie, eine Theorie der Universalgrammatik im Sinne der generativen Linguistik.

Grundbegriffe der stochastischen Optimalitätstheorie werden formal gefasst, und das Lernproblem, wie es in diesem Zusammenhang auftritt, wird dargestellt. Eine mittlerweile verbreitete Version stochastischer Optimalitätstheorie ist die von Boersma eingeführte, auf die wir kurz eingehen. Die Maximum-Entropie-Version stochastischer Optimalitätstheorie leiten wir in größerer Allgemeinheit als üblich aus informationstheoretischen Grundprinzipien her. Den Hauptteil der Arbeit nimmt die Untersuchung eines von Jäger (2003) vorgeschlagenen Lernalgorithmus ein. Wir zeigen, in welchem Sinne und unter welchen Bedingungen der Algorithmus konvergiert. Beziehungen zu allgemein bekannten Verfahren und klassischen Resultaten werden dabei erläutert.

Danksagung

Für den Vorschlag des Themas dieser Arbeit und die über den gesamten Zeitraum ihrer Entstehung hin gewährte Betreuung danke ich herzlich Gerhard Jäger. Für langjährige Unterstützung und immer neue Anregung bei meinen Versuchen zur Linguistik möchte ich Manfred Krifka meinen Dank aussprechen. Hilfreiche Gespräche und Einblick in ihre Arbeit gewährten mir die Mitglieder der Projektgruppe P13 "Bidirektionale Optimalitätstheorie" am Zentrum für Allgemeine Sprachwissenschaft (ZAS) Berlin. Ihnen allen danke ich. Schließlich möchte ich Karsten Tabelow für Hinweise zur Physik und Markus Reiss für sein Verständnis Dank sagen.

Contents

Notation and abbreviations

$\mathbf{1}_A$	indicator function of the set A, that is $\mathbf{1}_A$ takes values in $\{0,1\}$ and $\mathbf{1}_A(x) = 1$ iff x is an element of A
$\dot{x}(t)$	derivative of the differentiable function $x\colon [0,\infty) \to \mathbb{R}^N$ at "time" t
$\|v\|$	norm of vector v, usually the Euclidean norm in $\mathbb{R}^N$
e_l	l-th canonical basis vector in $\mathbb{R}^N$
$\mathrm{B}_\rho(x)$	open ball of radius ρ centered at x
$\overline{\mathrm{B}_\rho(x)}$	closed ball of radius ρ centered at x
$\mathcal{B}(S)$	Borel σ-algebra of the topological space S, i. e. the σ-algebra generated by the open sets of S; if S is countable, then $\mathcal{B}(S) = \wp(S)$
$\mathrm{cov}(X,Y)$	covariance of the random variables X and Y
$\mathrm{E}(X)$	expectation of the random variable X w. r. t. a given probability measure
$\mathrm{E}_\mathrm{Q}(X)$	expectation of the random variable X w. r. t. the probability measure Q
GLA	Gradual Learning Algorithm
iff	if and only if
i. i. d.	independently identically distributed (of a family of random variables)
$\ell^1(A)$	vector space of all real-valued absolutely convergent sequences indexed by elements of the countable set A
$\mathrm{M}_+^1(S)$	set of all probability measures on the Borel σ-algebra $\mathcal{B}(S)$ of the topological space S
$\min$	minimum
$\max$	maximum
$\mathbb{N}$	set of all positive natural numbers

$\mathbb{N}_0$	set of all natural numbers including zero
$\mathbb{N}_N$	set of the first N natural numbers starting with one
OT	Optimality Theory
$\mathbb{Q}$	set of all rational numbers
$\mathbb{R}$	set of all real numbers
supp	support (of a function or a probability measure)
$^\mathsf{T}$	transpose of a matrix (or vector)
$\wp(S)$	powerset, i. e. the set of all subsets of S
w. r. t.	with respect to

Introduction

The aim of this work is to analyse the convergence behaviour of a learning algorithm for a special version of stochastic Optimality Theory. Deterministic Optimality Theory as introduced by Prince and Smolensky (2004)[1] is a theory of Universal Grammar in the sense of generative linguistics and was originally developed in the context of phonology.

A grammar according to deterministic OT defines a relation between two sets of linguistic representations, the set of inputs and the set of outputs. Outputs represent linguistic surface forms (e. g. an utterance, the sound pattern of a word), inputs represent underlying forms (e. g. the meaning of an utterance, a sequence of consonants and vowels). Both sets are assumed to be universal, i. e. independent of any particular natural language. A language specific deterministic OT grammar assigns to each input a unique output, the grammatical form for expressing the given input, by searching for the best or optimal among a set of candidate outputs. The candidate sets are determined by the generator. This is a language independent subset of the product space of inputs and outputs. The generator defines the set of all admissible input output pairs and thus gives those relations between underlying forms and surface forms which must hold in all natural languages.

A further ingredient in building up a universal OT grammar is the set of constraints. Constraints are defined on the generator, i. e. on pairings of inputs with outputs. A constraint measures – along its specific grammatical dimension – the well-formedness of a particular output given an underlying input. Usually, the number of constraint violations caused by an input output pair is taken as measure of partial well-formedness. Constraints are violable and they may be in conflict with each other. Thus it may happen that for a given input there is no output among the candidates satisfying all constraints and that obeying one constraint means violating another.

A way of resolving conflicts among constraints is to rank them, that is to introduce an order of relative importance among them. It is then possible to find for each input an optimal output, where the optimality of an output depends on the current ranking of constraints. The component of grammar which assigns optimal outputs to underlying inputs according to a given constraint ranking is called evaluation component.[2] Genera-

[1] The book by Prince and Smolensky (2004) is a revised version of their original work published as technical report in 1993.

[2] For details on the evaluation component in deterministic OT see Prince and Smolensky (2004); also cf. example 2 in section 1.1.1 below.

tor, set of constraints and evaluation component constitute a universal deterministic OT grammar.[3] Any language specific grammar is determined by its particular ranking of constraints (Prince and Smolensky, 2004: p. 4):

> Universal Grammar provides a set of highly general constraints. These often conflicting constraints are *all* operative in individual languages. Languages differ primarily in the way they resolve the conflicts: in how they rank these universal constraints in strict domination hierarchies that determine the circumstances under which constraints are violated. A language-particular grammar *is* a means of resolving the conflicts among universal constraints.[4]

Stochastic OT differs from deterministic OT in that the evaluation component of a universal stochastic OT grammar yields for each input and a given ranking of constraints a probability distribution over the set of candidate outputs rather than a unique optimal output.

The learning problem for a universal stochastic OT grammar consists in finding a constraint ranking that best reflects a given empirical distribution on the set of admissible input output pairs. What counts as a good constraint ranking depends on the evaluation component of the OT grammar, that is on which version of stochastic OT is adopted.

A reasonable choice for how to determine the probability distributions over candidate sets is the following: Among all distributions which yield the same expected values of constraint violations select the one that maximizes entropy conditional on a given distribution of inputs. This is an application of Jayne's principle, also known as maximum entropy principle. Stemming from statistical physics, it has already been used in the field of linguistics (e.g. Berger et al. (1996) in the context of natural language processing, Goldwater and Johnson (2003) in the context of probabilistic OT).

The learning problem now amounts to searching for a constraint ranking that maximizes entropy conditional on the empirical distribution of inputs and yields the same average number of constraint violations as the empirical distribution. There are standard algorithms that converge to the unique solution of this optimization problem. The problem with these algorithms is that they require an a priori knowledge of the empirical distribution ("off-line" learning), which is implausible in case of human learners. For a different choice of the evaluation component, a gradual ("on-line") learning algorithm has been proposed (Boersma and Hayes, 2001). Unfortunately, convergence cannot be guaranteed for this algorithm due to the choice of the evaluation component.

In an attempt to overcome these difficulties, Jäger (2003) proposed an on-line version of one of the standard algorithms that can be applied to maximum entropy models, namely a stochastic gradient descent algorithm. The deterministic version of this algorithm converges provided an optimal constraint ranking exists. We will see under which conditions and in which sense Jäger's algorithm converges.

[3] If OT is empirically right and the right choices are made for generator, constraint set and evaluation component, then the resulting universal OT grammar is Universal Grammar.

[4] Emphasis as in the original.

The rest of this work is organized as follows. Chapter 1 explains the framework of stochastic Optimality Theory, in particular its maximum entropy version, and introduces Jäger's learning algorithm. Chapter 2 is dedicated to the convergence analysis of Jäger's algorithm, and chapter 3 summarizes and interprets the convergence results. Some material on convexity and optimization has been gathered in appendices A and B.

In section 1.1 we define the relevant concepts of stochastic OT alluded to above. The learning problem is introduced. We will see examples of different versions of stochastic OT. Boersma's version of stochastic OT is presented in greater detail in section 1.2, as it comes with a learning algorithm of its own (the GLA). Problems connected with the GLA are outlined. The purpose of section 1.3 is to derive the maximum entropy version of stochastic OT from fundamental principles of information theory, summarized in section 1.3.1. The derivation of the Gibbs family of probability distributions as maximum entropy model under linear constraints in section 1.3.2 is somewhat complicated by the fact that we consider distributions on denumerable, but possibly infinite sets. The dual problem associated with entropy maximization is stated in section 1.3.3. Jäger's algorithm, finally introduced in section 1.4, is designed for solving the dual optimization problem.

Chapter 2 opens with a summary of notions of stochastic convergence, compiled in section 2.1. The minima of the dual function are related to properties of the given constraints in section 2.2. A standing assumption on the underlying generator will be that candidate sets are finite, although the generator itself may be infinite. In section 2.3 convergence of Jäger's algorithm to stochastic equilibrium is proved for constant step size. Section 2.4 characterizes convergence as the step size, though still constant, is chosen smaller and smaller. Varying Jäger's algorithm by allowing the step size to decrease to zero in the course of learning, in section 2.5 we find ourselves in the situation studied by Robbins and Monro (1951).

Section 3.1 provides a summary of the results obtained in chapter 2. In section 3.2 we discuss why it was reasonable to work with infinitely many inputs and outputs. We sketch how Jäger's algorithm has to be modified if we want to allow for infinite candidate sets. Following Jäger and Rosenbach (2005), we present an application of maximum entropy OT to a syntactic phenomenon in section 3.3. Section 3.4 concludes our analysis.

Chapter 1

Stochastic Optimality Theory, maximum entropy and learning

Before presenting Jäger's algorithm in detail, we have to review the underlying theoretical framework. In section 1.1, we describe basic concepts of stochastic Optimality Theory together with the notation we shall adopt in the rest of this work. In particular, we define what is meant by an evaluation kernel and what a grammar in the sense of stochastic OT is. The learning problem arising in this context is introduced.

Section 1.2 presents Boersma's version of stochastic OT, which amounts to a special choice of the evaluation kernel. We briefly discuss a learning algorithm proposed by Boersma (1997) – the so-called Gradual Learning Algorithm (GLA) – with regard to its convergence behaviour.

A different version of stochastic OT can be derived by appeal to Jaynes's principle, also known as maximum entropy principle. Section 1.3.1 serves to recall this principle as well as Shannon's concept of entropy. The idea of entropy maximization under suitable constraints, developed in sections 1.3.2 and 1.3.3, leads to a family of distributions which can readily be interpreted as an evaluation kernel. The resulting OT version coincides with the one introduced by Goldwater and Johnson (2003).

Finally, in section 1.4, we present Jäger's learning algorithm for the maximum entropy version of stochastic OT and point out its connection with the gradient descent method.

1.1 Stochastic Optimality Theory and the learning problem

The main difference between deterministic and stochastic OT lies in the way the evaluation component works. Given an input and a ranking of constraints, the evaluation component of a stochastic OT grammar yields a probability distribution on the set of candidate outputs, while the deterministic variant determines a unique "best" output for each input.

In section 1.1.1 we give formal definitions of the ingredients of stochastic OT. Section 1.1.2 is concerned with the learning problem, which amounts to a procedure for se-

lecting a ranking of constraints in response to observations of input output pairs drawn from an empirical distribution.

1.1.1 Basic concepts of stochastic Optimality Theory

Let I, O be countable non-empty sets and $G \subset I \times O$ be a subset. By I, O we denote the set of *inputs* and *outputs*, respectively, and G is the set of admissible input output pairs, called the *generator*.

For each input i the generator determines a set of admissible outputs, the *output candidates*. Denote the set of output candidates for an input i by O_i, that is we let

$$O_i := \{o \in O \mid (i,o) \in G\}, \quad i \in I. \tag{1.1}$$

We require G to be such that O_i is non-empty for all $i \in I$. Let $N \in \mathbb{N}$ be a fixed natural number and $c_1, \ldots, c_N$ be functions $G \to \mathbb{N}_0$. Then N is the number of *constraints* $c_1, \ldots, c_N$. According to the above assumption, for each input i there is at least one admissible output. Write c for the compound function $G \to \mathbb{N}_0^N$ having as its components the functions $c_1, \ldots, c_N$. Let us refer to c as *feature function*.

A stochastic OT grammar works in the following way. Given an input $i \in I$ and a ranking of constraints, which may be represented as a vector $r \in \mathbb{R}^N$, the grammar assigns to each output $o \in O$ a probability based on the vector $c(i,o)$ of constraint violations and the vector r of corresponding constraint ranks, where positive probability may be ascribed only to outputs in O_i. Let us make this more precise.

Definition 1.1. An *evaluation kernel* for generator $G \subseteq I \times O$ and feature function $c : G \to \mathbb{N}_0^N$ is a mapping $p : \mathcal{R} \times I \times O \to [0,1]$, written $(r,i,o) \mapsto p_r(o|i)$, satisfying the following conditions:

(i) $\mathcal{R} \subseteq \mathbb{R}^N$ is open and such that $r + \tilde{r} \in \mathcal{R}$ for all $r \in \mathcal{R},\ \tilde{r} \in [0,\infty)^N$,

(ii) $$\text{for all } r \in \mathcal{R},\ i \in I, \quad \sum_{o \in O_i} p_r(o|i) \ = \ \sum_{o \in O} p_r(o|i) \ = \ 1,$$

(iii) for all $r \in \mathcal{R},\ i \in I,$ the mapping $O \ni o \mapsto p_r(o|i)$ is measurable w. r. t. the σ-algebra generated by $c(i,.)$,

(iv) for all $r \in \mathcal{R},\ (i,o) \in G,\ l \in \mathbb{N}_N,\ \delta > 0,$
$c_l(i,o) \ = \ \min\{c_l(i,\tilde{o}) \mid \tilde{o} \in O_i\}$ implies $p_{r+\delta e_l}(o|i) \ \geq \ p_r(o|i)$.

If, in addition, the mapping $\mathcal{R} \ni r \mapsto p_r(o|i)$ is continuous (differentiable) for each $(i,o) \in G$, then p is called a *continuous* (*differentiable*) evaluation kernel.

Recall that the sets I, O are assumed to be discrete. Therefore, requiring property (ii) is equivalent to saying that $(p_r(.|.))_{r \in \mathcal{R}}$ must be a family of Markov kernels from I to O with *support* in G, that is a transition from i to o can receive positive probability only if

$(i, o) \in G$. More generally, we may regard any Markov kernel from I to O as a *language* over $I \times O$ in the sense of stochastic OT. A language is compatible with a generator $G \subseteq I \times O$ whenever its support is contained in G.

In case p is continuous the dependence of the conditional probabilities on the ranking vector $r \in \mathcal{R}$ is regular in the sense that small changes in the ranking lead to small changes in the probabilities. Observe, however, that those changes need not be uniform over the set of inputs.

Condition (iii) in the above definition establishes a link between the constraint function and the assignment of probabilities. It amounts to saying that distinct outputs for a given input may receive different probabilities only if they differ in the number of violations of at least one constraint. In conjunction with property (ii) this implies that if all constraints are constant for a given input $i \in I$ then O_i must be finite and $p_r(.|i)$ the uniform distribution on O_i.

While requirements (ii) and (iii) are clearly necessary for capturing our intuition of how the evaluation part of an OT grammar should work, property (iv) is somewhat arbitrary. The idea is that those outputs which fare best with respect to a certain constraint must not become less probable when the constraint itself gets ranked higher.[1] According to property (i), promotion of constraints is always possible.

Definition 1.2. Let $I \times O$ be a set of input output pairs as above. A *universal stochastic OT grammar* over $I \times O$ is a triple (G, c, p) such that $G \subseteq I \times O$ is a generator, c a feature function on G and p an evaluation kernel for G and c. A *specific stochastic OT grammar* is a universal stochastic OT grammar (G, c, p) together with a vector $r \in \mathcal{R}$, where $\mathcal{R}$ is the ranking domain of p from definition 1.1.

Two specific OT grammars may differ not only as to the value of the ranking vector r, but also with respect to the underlying universal OT grammar. Nevertheless, they can give rise to the same language, i. e. to the same family of output probabilities conditional on the inputs.

Let us consider two simple versions of stochastic OT. Here, by a *version of stochastic OT* we mean a class of universal stochastic OT grammars. In giving a version of stochastic OT we will usually not specify the generator nor the constraints, but rather provide a prescription of how to build an evaluation kernel given a generator $G \subseteq I \times O$ and a feature function $c : G \to \mathbb{N}_0^N$.

Example 1. *Global linear models.* Let the ranking domain $\mathcal{R}$ be $(0, \infty)^N$. For $r \in \mathcal{R}$ and any input $i \in I$ denote by $\hat{O}_i(r)$ the set of outputs that produce a minimal amount of constraint violations in the direction of r, that is we set

$$\hat{O}_i(r) := \big\{o \in O_i \mid \langle r, c(i, o)\rangle = \min_{\tilde{o} \in O_i} \langle r, c(i, \tilde{o})\rangle\big\}.$$

[1] Note that high numbers correspond to high rankings, here; the opposite ordering is also found in the literature.

We notice that $\hat{O}_i(r)$ is well defined as the components of r and c are non negative. Assume generator G and feature function c are such that $\hat{O}_i(r)$ is finite for all $i \in I$, $r \in \mathcal{R}$. Define $p : \mathcal{R} \times I \times O \to [0,1]$ by

$$p_r(o|i) := \begin{cases} \frac{1}{\#\hat{O}_i(r)} & \text{if } o \in \hat{O}_i(r), \\ 0 & \text{else,} \end{cases} \qquad r \in \mathcal{R},\ i \in I,\ o \in O.$$

Clearly, p is an evaluation kernel for G and c, and (G, c, p) is a universal stochastic OT grammar. The evaluation kernel assigns equal probability to all outputs which minimize $\langle r, c(i,.)\rangle$ and discards any other alternative. $\Diamond$

A salient feature of deterministic Optimality Theory in the sense of Prince and Smolensky (2004) is that ranking of constraints means (total) ordering and that a higher ranked constraint overrules all constraints which are ranked lower. These properties can be captured in the present framework.[2]

Example 2. *Deterministic OT.* Let the ranking domain $\mathcal{R}$ be a non-empty subset of $\mathbb{R}^N$. Denote by $\nu_j(r)$, $r \in \mathcal{R}$, $j \in \mathbb{N}_N$, the number of components r_l of r such that $r_l \geq r_j$, that is we set

$$\nu_j(r) := \#\{l \in \mathbb{N}_N \mid r_l \geq r_j\}.$$

Define the map $\nu : \mathcal{R} \to \{1, \ldots, N\}^N$ by

$$\nu(r) := \big(\nu_1(r), \ldots, \nu_N(r)\big), \quad r \in \mathcal{R}.$$

Given $W \in \{1, \ldots, N\}^N$, $l \in \mathbb{N}_N$ and $i \in I$, define a *cost function* $K_l(W, i, .)$ on O_i by

$$K_l(W, i, o) := \sum_{j \in \mathbb{N}_N : W_j = l} c_j(i, o), \quad o \in O_i.$$

Next, given $i \in I$, $r \in \mathcal{R}$, introduce a preference relation $\succ_{i,r}$ on O_i by setting

$$\begin{aligned} o \succ_{i,r} \tilde{o} \quad :\Leftrightarrow \quad & d := \min\big\{l \in \mathbb{N}_N \mid K_l(\nu(r), i, o) \neq K_l(\nu(r), i, \tilde{o})\big\} \text{ is finite and} \\ & K_d(\nu(r), i, o) < K_d(\nu(r), i, \tilde{o}). \end{aligned}$$

Observe that the set $\{l \in \mathbb{N}_N \mid K_l(\nu(r), i, o) \neq K_l(\nu(r), i, \tilde{o})\}$ might be empty in which case we would have $d = \infty$ by convention. Denote by $\check{O}_i(r)$ those elements of O_i which are maximal with respect to $\succ_{i,r}$, that is we set

$$\check{O}_i(r) := \{o \in O_i \mid \nexists\, \tilde{o} \in O_i :\ \tilde{o} \succ_{i,r} o\}.$$

If $\check{O}_i(r)$ is finite for all $i \in I$, $r \in \mathcal{R}$, then we obtain an evaluation kernel p for G and c by requiring that $p_r(.|i)$ be the uniform distribution on $\check{O}_i(r)$.

Now assume that $\check{O}_i(r)$ is a singleton for all $i \in I$ and those $r \in \mathcal{R}$ with pairwise distinct components. Then (G, c, p) corresponds to a "universal" deterministic OT grammar. $\Diamond$

[2]See, for example, the lecture notes by M. Collins, MIT course 6.891, Fall 2003.

1.1.2 The learning problem

Suppose we are given a universal stochastic OT grammar and an empirical distribution of input output pairs. The learning problem then consists in finding a ranking vector such that the corresponding conditional distributions match as well as possible the empirical distribution.

To be more specific, let (G, c, p) be a universal stochastic OT grammar and p_{emp} a probability distribution on $I \times O$ with support in G. Throughout the rest of this work we will assume that the support of any empirical distribution p_{emp} contains at least one input output pair for each given input, that is for each input $i \in I$ there is an output $o \in O_i$ such that $p_{emp}(i, o) > 0$.

Denote by $p_{emp}(.|i)$ the conditional distribution which p_{emp} induces on O_i given an input $i \in I$. Then we are looking for a ranking vector $r \in \mathcal{R}$ such that the distributions $p_r(.|i)$ and $p_{emp}(.|i)$, $i \in I$, fit in a way that is optimal according to some criterion of goodness.

The empirical distribution p_{emp}, however, is not directly observable. Accessible to observation, rather, is a sequence of input output pairs distributed according to p_{emp}. Call this sequence $(X_n)_{n\in\mathbb{N}}$, write $X_n = (X_n^{in}, X_n^{out})$, $n \in \mathbb{N}$, and let $(\Omega, \mathcal{F}, \mathrm{P})$ denote the common probability space those random variables live on. We assume that observations are mutually independent, i. e. (X_n) is an i. i. d. sequence with $X_n \overset{d}{\sim} p_{emp}$. By definition, for all $n \in \mathbb{N}$ we have

$$\mathrm{P}(X_n^{in} = i, X_n^{out} = o) = p_{emp}(i, o), \qquad (i, o) \in G, \tag{1.3a}$$

$$\tilde{p}_{emp}(i) := \mathrm{P}(X_n^{in} = i) = \sum_{o \in O_i} p_{emp}(i, o), \qquad i \in I. \tag{1.3b}$$

The situation now is essentially the same as in statistical point estimation theory.[3] In particular, one could introduce a loss or error function and choose the ranking vector $r \in \mathcal{R}$ so as to minimize the error between $p_r(.|i)$ and $p_{emp}(.|i)$.

In section 1.3 we will follow a different approach and employ the idea of maximizing entropy subject to the requirement that the average number of constraint violations be preserved not only as a criterion for choosing the ranking vector, but as a starting point for constructing an evaluation kernel.

1.2 Boersma's version of stochastic Optimality Theory and the Gradual Learning Algorithm

A by now popular version of stochastic Optimality Theory is the one introduced in Boersma (1997, 1998) and further developed in Boersma and Hayes (2001). As for stochastic OT in general, motivation for the introduction of randomness comes from a twofold

[3] See, for example, the first chapter in Lehmann (1983).

desire: to be able to account for free variation among output candidates for the same input, and to render learning robust against errors or "noise" in the observations used for determining a ranking of constraints.

Recall example 2, which translates the original version of Optimality into our framework. Although we allow for constraint ranks taking values on a continuous scale, what is really needed in deterministic OT is just an ordering of the constraints. Accordingly, during evaluation no distinction is drawn between different ranking vectors as long as the ordering of their component values remains the same. The function ν as given in example 2 determines the corresponding equivalence relation among ranking vectors.

Boersma's idea in the definition of a class of universal stochastic OT grammars is to incorporate randomness by adding a perturbation to the current ranking of constraints each time a set of possible outputs is evaluated. At the level of probability distributions on the set of outputs, the random perturbations correspond to a randomization of the preference relation $\succ_{i,r}$ as defined in example 2. While the output candidates are evaluated with respect to a perturbed ranking of constraints, the ranking vector itself remains unchanged.

In making things more precise, let us assume that the set O_i of output candidates is finite for each input $i \in I$. Given an N-dimensional feature function c on a generator G, define the ordering map ν and the cost functions K_l, $l \in \mathbb{N}_N$, as for deterministic OT.

Example 3. *Boersma's stochastic OT.* Let the set of admissible ranking vectors $\mathcal{R}$ be $\mathbb{R}^N$. Let Y be an N-dimensional random variable with standard normal distribution on some probability space $(\Omega, \mathcal{F}, \mathrm{P})$.[4] Define a randomized preference relation $\succ_{i,r,\cdot}$ on O_i by setting for $\omega \in \Omega$

$$\begin{aligned} o \succ_{i,r,\omega} \tilde{o} \quad :\Leftrightarrow \quad & d := \min\bigl\{l \in \mathbb{N}_N \mid K_l\bigl(\nu\bigl(r + Y(\omega)\bigr), i, o\bigr) \neq K_l\bigl(\nu\bigl(r + Y(\omega)\bigr), i, \tilde{o}\bigr)\bigr\} \\ & \text{is finite and} \quad K_d(\nu(r), i, o) < K_d(\nu(r), i, \tilde{o}). \end{aligned}$$

Denote by $\check{O}_i(r, .)$ the random set of those elements of O_i which are maximal with respect to $\succ_{i,r,\cdot}$, that is we set

$$\check{O}_i(r, \omega) := \{o \in O_i \mid \nexists\, \tilde{o} \in O_i : \tilde{o} \succ_{i,r,\omega} o\}, \qquad \omega \in \Omega.$$

Observe that $\#\check{O}_i(r, \omega) = 1$ with probability one, because the event $\{Y = v\}$ has probability zero for each $v \in \mathbb{R}^N$. The evaluation kernel for c and G is now defined by

$$p_r(o|i) := \mathrm{P}\bigl(\{\omega \in \Omega \mid o \in \check{O}_i(r, \omega)\}\bigr), \qquad o \in O_i,\ i \in I,\ r \in \mathbb{R}^N.$$

Notice that the definition of $p_r(o|i)$ does not depend on the actual choice of the random variable Y, but only on its distribution.

Consider the case, where for each $l \in \mathbb{N}_N$ and each given $i \in I$ there is exactly one $o \in O_i$ minimizing $c_l(i, .)$ over O_i, that is the deterministic set $\check{O}_i(e_l)$ is a singleton for

[4] That is to say $Y = (Y_1, \ldots, Y_N)^\mathsf{T}$, where $Y_1, \ldots, Y_N$ are i. i. d. with distribution $N(0, 1)$.

each $l \in \mathbb{N}_N$. Then we can express $p_r(.|i)$ directly in terms of c and r, namely

$$p_r(o|i) \;=\; \sum_{l=1}^{N} \mathbf{1}_{\{o \in \check{O}_i(e_l)\}} \, \mathrm{P}\big(\{\omega \in \Omega \mid r_l + Y_l(\omega) > r_j + Y_j(\omega) \;\forall\, l \neq j\}\big), \quad o \in O_i.$$

The probabilities appearing on the right-hand side of the above equation can be calculated using the density function of the standard normal distribution. ◊

Boersma's version of stochastic OT comes with a learning algorithm, the so-called Gradual Learning Algorithm (GLA). The situation is as in section 1.1.2. We observe an i. i. d. random sequence $(X_n)_{n\in\mathbb{N}}$ on $(\Omega, \mathcal{F}, \mathrm{P})$ of input output pairs drawn from a probability distribution p_{emp} on the generator G. In order to compute a ranking vector reflecting the empirical distribution p_{emp}, we start with an arbitrary initial ranking.

The GLA then enters a recursive procedure: Take the input of the current observation, i. e. the input i of the pair $(i, o) = (X_n^{in}, X_n^{out})$ if we are currently at step n, and evaluate the output candidates for this input according to the current ranking of constraints. This yields a second output $\tilde{o}$. If the new output equals the one that has been observed, i. e. if $\tilde{o} = o$, then leave the ranking vector as it is. Else compare the constraint violations incurred by the two outputs (the associated input i being the same for both). Increase the rank of all those constraints favouring the observed output o over $\tilde{o}$ by a fixed small amount, and decrease the rank of all those constraints which favour the computed output $\tilde{o}$ over o by the same amount.

A constraint *favours* output o_1 over output o_2, both outputs being candidates for the same input, iff the number of violations of that constraint incurred by o_1 is smaller than the number of violations incurred by o_2. The amount of increase (resp. decrease) of a rank is referred to as *plasticity value*.

In order to write down formally the GLA, let $(\tilde{H}_n)_{n\in\mathbb{N}_0}$ be a sequence of measurable mappings $\Omega \times I \times \mathbb{R}^N \to O$ such that $(\tilde{H}_n(., i, r))_{n\in\mathbb{N}_0}$ is an i. i. d. random sequence with distribution $p_r(.|i)$ under P for all $i \in I$, $r \in \mathbb{R}^N$, where $(p_r)_{r\in\mathbb{R}^N}$ is the evaluation kernel of example 3. The random variables $\tilde{H}_n(., i, r)$ thus mimic the action of the evaluation component of Boersma's version of stochastic OT.

To start, choose a plasticity value $\eta > 0$ and an initial constraint ranking, that is an $\mathcal{F}$-measurable random variable R_0^η with values in $\mathbb{R}^N$ which is assumed to be independent of the other random variables. Compute a sequence $(R_n^\eta)_{n\in\mathbb{N}_0}$ of constraint rankings according to the recursion formula

$$R_{n+1}^\eta := \; R_n^\eta \;+\; \eta \cdot \mathrm{sgn}\big(c(X_n^{in}, \tilde{H}_n(X_n^{in}, R_n^\eta)) - c(X_n^{in}, X_n^{out})\big), \quad n \in \mathbb{N}_0, \tag{1.4}$$

where sgn is to be understood as the componentwise application of the sign function.

There are some difficulties connected with Boersma's version of stochastic OT, see Keller and Asudeh (2002). The most important problem in the context of learning is caused by *harmonic bounding*. Consider the following situation.

Example 4. *Problem with Boersma's version of stochastic OT.* Let the set of inputs $I := \{i\}$ be a singleton, let the set of outputs $O := \{o_1, o_2, o_3\}$ consist of exactly three elements, and let the generator G be $I \times O$. Hence $O = O_i$, that is the three outputs are candidates for i. Assume there are two constraints c_1, c_2, and let the evaluation kernel be as in example 3. Define the feature function $c = (c_1, c_2)^\mathsf{T}$ according to the table below, where the number of constraint violations incurred by an input output pair is given by the number of asterisks in the respective cell.

	c_1	c_2
(i, o_1)	$****$	
(i, o_2)	$**$	$*$
(i, o_3)	$*$	$**$

With this feature function, the pair (i, o_2) will have probability zero whatever the ranking of c_1, c_2 is. If c_1 outranks c_2 (in evaluation mode, after addition of a random perturbation to the current ranking), then o_3 will be optimal. Else if c_2 is ranked higher than c_1, then o_1 will be the best candidate. Therefore, o_2 will never be optimal and will not be produced by the evaluation component. Yet output o_2 is neither the candidate incurring the highest total number of constraint violations nor is it the worst choice under any single constraint. ◊

We can describe how Boersma's GLA works in the situation of example 4 by looking at recursion formula (1.4), the update rule of the GLA. Imagine we are given an empirical distribution which ascribes non-zero probability p_k to the pair (i, o_k), $k \in \{1, 2, 3\}$, such that $p_1 \leq p_2 \leq p_3$. Fix a gain parameter $\eta > 0$, and denote by $r_s \in \mathbb{R}^2$ the deterministic initial ranking.

Now suppose we are at the n-th learning step. With probability p_1 the learner observes pair (i, o_1). Since o_2 cannot be produced by the learner, the output generated according to the current ranking is either o_1 or o_3. In the former case, the current ranking of constraints remains unchanged, in the latter case the vector of constraint ranks is shifted by $\eta(-1, 1)^\mathsf{T}$. Similarly, if pair (i, o_3) is observed, then any non-zero update must be a shift by $\eta(1, -1)^\mathsf{T}$. Lastly, if the learner observes pair (i, o_2), then the current ranking will be updated either by $\eta(-1, 1)^\mathsf{T}$ or $\eta(1, -1)^\mathsf{T}$. If c_1 currently outranks c_2, then update $\eta(-1, 1)^\mathsf{T}$ will be more probable, else update $\eta(1, -1)^\mathsf{T}$ will be more likely to occur.

The GLA therefore shifts the current ranking of constraints along the straight line $\{r_s + \lambda(1, -1)^\mathsf{T} \mid \lambda \in \mathbb{R}\}$. This is not a problem in itself. Notice that in Boersma's version of stochastic OT it is the difference between the ranks of constraints that counts, not their absolute value. Moreover, we have

$$\big\langle (1, 1)^\mathsf{T}, c(i, o_2) - c(i, o_3) \big\rangle \;=\; 0,$$

that is in direction $(1,1)^\mathsf{T}$ the constraints for (i,o_2) and (i,o_3) cancel out. We will come back to this point in section 2.2 below in the context of maximum entropy stochastic OT. Besides the fact that no specific grammar learned by the GLA can produce a distribution such that (i,o_2) has non-zero probability, the difficulty in example 4 is this: The probability ratio between p_1 and p_3 causes the GLA to adjust the ratio between the ranks of c_1 and c_2 in such a way that if $p_1 < p_3$ then c_1 outranks c_2. The occurrence of pair (i,o_2), on the other hand, pushes the constraint ranking back to a point, where the ranks of c_1 and c_2 are equal.

1.3 Maximum entropy approach to stochastic Optimality Theory

From section 1.1.2 we know that learning a ranking of constraints amounts to choosing a family of distributions $\{p(.|i) \mid i \in I\}$, where $p(.|i)$ is a probability distribution on the set of possible outputs O with support contained in O_i. Any such family induces – by means of the marginal $\partial_I p_{emp} = \tilde{p}_{emp}$ on I of the empirical distribution p_{emp} – a probability distribution p on $I \times O$, namely

$$p(i,o) := \tilde{p}_{emp}(i) \cdot p(o|i), \qquad (i,o) \in I \times O.$$

It is now reasonable to demand that the family of conditional distributions be such that the induced distribution p yields the same average number of constraint violations as does the empirical distribution p_{emp}, i. e. we require

$$(*) \qquad \sum_{(i,o)\in G} c(i,o) \cdot \big(p(i,o) - p_{emp}(i,o)\big) \;=\; 0.$$

Thus, instead of letting some evaluation kernel generate distributions over candidate outputs given a certain ranking of constraints, we can try directly to determine a probability distribution p on $I \times O$ such that $\mathrm{supp}(p) \subseteq G$, $\partial_I p = \tilde{p}_{emp}$, and $(*)$ holds.

There is, in general, more than one distribution compatible with the above requirements, so that we need a criterion for preferring one compatible distribution to any other. Such a criterion is provided by Jaynes's principle which generalizes and mathematically justifies Laplace's "principle of insufficient reason" and the ancient rule that two alternatives which are mutually exclusive should be regarded as being equally likely if nothing else is known about them.[5]

In order to get in a position to state Jaynes's principle we have to recall the concept of entropy as introduced in Shannon (1948), which provides a measure of uncertainty. This is done in section 1.3.1. In subsections 1.3.2 and 1.3.3, Jaynes's principle, also known as maximum entropy principle, is applied to the problem at hand. A resulting dual or conjugate problem conduces us back to the original task of learning a ranking of constraints.

An exposition of the maximum entropy approach in the field of computational linguistics can be found in Berger et al. (1996). Note, however, that the setting here is more general in that we allow for distributions over countably infinite sets.

[5] See Jaynes (1957a,b).

1.3.1 Shannon entropy and Jaynes's principle

Let A be a countable set. In information theory, A plays the rôle of the *alphabet*. With Shannon (1948) define the entropy H of a probability distribution p on A as[6]

$$\mathrm{H}(p) := -\sum_{i \in A} p(i) \cdot \ln p(i), \tag{1.5}$$

where $0 \cdot \ln(0) = 0$, i. e. $x \mapsto x \cdot \ln(x)$ is interpreted as a continuous function on $[0, 1]$. Notice that $\mathrm{H}(.)$ is well defined as a function over probability distributions on A and takes its values in $[0, \infty]$.

Imagine we were confronted with a random experiment with outcomes labelled by elements of A and that we had to bet on the result of the experiment. Suppose the probabilities $p(i)$ of outcome with label i were known to us for all $i \in A$. Then we should regard $\mathrm{H}(p)$ as indicating the degree of uncertainty inherent in the random experiment governed by the law p.

Assume, for the moment, that A is finite. The most extreme cases are, on the one hand, point distributions, i. e. $p(i) = 1$ for some $i \in A$, and on the other hand the uniform distribution, i. e. $p(i) = \frac{1}{n}$ for all $i \in A$, where $n = \#A$ is the number of elements of A. In the former case we have $\mathrm{H}(p) = 0$, in the latter case $\mathrm{H}(p) = \ln(n)$. Observe that $0 \leq \mathrm{H}(p) \leq \ln(\#A)$ for any distribution p on A.

Consequently, $\mathrm{H}(.)$ is minimal whenever one outcome is almost certain, and maximal when all outcomes are equally likely. Of course, in case $p(i) = 1$ for some $i \in A$, knowing p eliminates uncertainty about the result of the experiment since the outcome must be the one labelled i. In case p is uniformly distributed there is no bias in the distribution that would help us in placing our bet (or prediction), while in all other cases, we could exploit such a bias.

If the alphabet A is countably infinite, then it still holds that entropy is minimal (and zero) for point distributions. The maximum value of $\mathrm{H}(.)$, however, now goes to infinity as there is no uniform distribution on an infinite set.[7]

The above definition of entropy might seem arbitrary at first glance. Yet the form of $\mathrm{H}(.)$ as given by (1.5) is determined – up to a constant positive factor – by three basic axioms, see theorem 2 in Shannon (1948) or appendix A in Jaynes (1957a). The multiplicative constant corresponds to a choice of the base for the logarithm appearing in (1.5). For our purposes, the natural logarithm is more convenient, while in computer science one usually works with the binary logarithm.[8]

Important for us are the concavity / convexity properties of the Shannon entropy. Let p, q be probability distributions on A, and $\lambda \in [0, 1]$. Then $\lambda p + (1-\lambda)q$ also is a probability

[6] In Shannon (1948), entropy is defined over probability distributions on finite sets. See, for example, Harremoës and Topsøe (2001) for properties of the extension to distributions on countable sets.

[7] Probability distributions with infinite entropy fall in the class of *hyperbolic* distributions, although there are hyperbolic distributions with finite entropy, cf. Harremoës and Topsøe (2001).

[8] Entropy $\mathrm{H}(p)$ then gives the minimal average code length in **b**inary dig**its** which is needed for the noiseless transmission of messages generated by a random source with output distribution p.

distribution on A, and it holds that

$$\begin{aligned}
& \mathrm{H}\big(\lambda p + (1-\lambda)q\big) \\
= \quad & -\sum_{i\in A} \lambda p(i)\cdot \ln\big(\lambda p(i) + (1-\lambda)q(i)\big) \;-\; \sum_{i\in A} (1-\lambda)q(i)\cdot \ln\big(\lambda p(i) + (1-\lambda)q(i)\big) \\
= \quad & \lambda \sum_{i\in A} p(i)\cdot \Big(\ln\big(\tfrac{1}{p(i)}\big) + \ln\big(\tfrac{p(i)}{\lambda p(i)+(1-\lambda)q(i)}\big)\Big) \\
& +\; (1-\lambda)\sum_{i\in A} q(i)\cdot \Big(\ln\big(\tfrac{1}{q(i)}\big) + \ln\big(\tfrac{q(i)}{\lambda p(i)+(1-\lambda)q(i)}\big)\Big) \\
= \quad & \lambda\,\mathrm{H}(p) \;+\; (1-\lambda)\mathrm{H}(q) \;+\; \lambda\,\mathrm{D}\big(p \parallel \lambda p + (1-\lambda)q\big) \;+\; (1-\lambda)\mathrm{D}\big(q \parallel \lambda p + (1-\lambda)q\big).
\end{aligned}$$

Here, $D(.\|.)$ denotes the *Kullback-Leibler divergence*, also known as *relative entropy*, which is defined for probability distributions p, q on A by

$$D(p\,\|\,q) := \sum_{i\in A} p(i)\cdot \ln\big(\tfrac{p(i)}{q(i)}\big). \tag{1.6}$$

Kullback-Leibler divergence is well defined as a function on $\mathrm{M}^1_+(A)\times \mathrm{M}^1_+(A)$ and takes its values in $[0,\infty]$. This can be seen by applying the logarithm inequality

$$x\cdot \ln\big(\tfrac{x}{y}\big) \;\geq\; x-y \quad \text{for all } x,y\in[0,1].$$

Moreover, $D(p\|q) = 0$ if and only if $p(i) = q(i)$ for all $i \in A$. The Kullback-Leibler divergence may be regarded as a quantity indicating how different two probability distributions are. Notice, though, that $D(p\|q)$ is not symmetric in p, q. We see that entropy $H(.)$ is strictly concave, that is

$$\mathrm{H}\big(\lambda p + (1-\lambda)q\big) \quad\geq\quad \lambda\mathrm{H}(p) \;+\; (1-\lambda)\mathrm{H}(q) \quad \text{for all } \lambda\in[0,1], \tag{1.7}$$

where equality holds if and only if $p(i) = q(i)$ for all $i \in A$ or $\lambda \in \{0,1\}$. On the other hand, we can obtain an upper bound for $\mathrm{H}(\lambda p + (1-\lambda)q)$, for it holds that

$$\begin{aligned}
\mathrm{D}\big(p \parallel \lambda p + (1-\lambda)q\big) \quad &= \quad \sum_{i\in A} p(i)\cdot \ln\big(\tfrac{p(i)}{\lambda p+(1-\lambda)q}\big) \\
\leq \quad \sum_{i\in A} p(i)\cdot \ln\big(\tfrac{p(i)}{\lambda p}\big) \quad &= \quad \ln\big(\tfrac{p(i)}{\lambda p}\big)\sum_{i\in A} p(i) \quad = \quad \ln\big(\tfrac{1}{\lambda}\big).
\end{aligned}$$

Similarly, it holds that

$$\mathrm{D}\big(q \parallel \lambda p + (1-\lambda)q\big) \;\leq\; \ln\big(\tfrac{1}{1-\lambda}\big).$$

Since $\lambda \in [0,1]$, we have $\lambda \ln(\frac{1}{\lambda}) \leq \frac{1}{e}$, wherefore

$$\lambda \ln\big(\tfrac{1}{\lambda}\big) \;+\; (1-\lambda)\ln\big(\tfrac{1}{1-\lambda}\big) \;\leq\; \tfrac{2}{e}.$$

Putting everything together, one obtains

$$\mathrm{H}\big(\lambda p + (1-\lambda)q\big) \quad\leq\quad \lambda\mathrm{H}(p) \;+\; (1-\lambda)\mathrm{H}(q) \;+\; \frac{2}{e} \quad \text{for all } \lambda\in[0,1]. \tag{1.8}$$

Consider now the case that the alphabet A is a subset of the product of two countable sets. More specifically, let $A = G$, where $G \subseteq I \times O$ and I, O satisfy the assumptions of section 1.1. Define the *conditional entropy* of a probability distribution p on G with respect to a distribution $\tilde{p}$ on I as

$$\mathrm{H}(p \,|\, \tilde{p}) := -\sum_{(i,o)\in G} \tilde{p}(i) \cdot p(o|i) \cdot \ln p(o|i), \quad \text{where} \quad p(o|i) := \frac{p(i,o)}{\sum_{\tilde{o}\in O_i} p(i,\tilde{o})}.$$

Thus, $\mathrm{H}(p|\tilde{p})$ may be interpreted as the uncertainty about the outcome of a chance experiment with law p remaining when the result of that part of the experiment which is governed by the law $\tilde{p}$ has been revealed.

We finally turn to Jaynes's principle of statistical inference. The situation can be described as follows (Jaynes, 1957a: p. 622):

> "Just as in applied statistics the crux of the problem is often the devising of some method of sampling that avoids bias, our problem is that of finding a probability assignment which avoids bias, while agreeing with whatever information is given."

The means for choosing among probability assignments that are compatible with the given information is Shannon's measure of entropy (Jaynes, 1957a: p. 623):

> "It is now evident how to solve our problem; in making inferences on the basis of partial information we must use that probability distribution which has maximum entropy subject to whatever is known. This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have."

We can represent the information at our disposal as a set of probability distributions $\mathcal{P}_0 \subseteq \mathrm{M}_+^1(A)$. According to Jaynes's principle, we must choose $p_* \in \mathcal{P}_0$ such that $\mathrm{H}(p_*) = \max\{\mathrm{H}(p) \mid p \in \mathcal{P}_0\}$. The probability distribution p_* then is the best description of the phenomenon at hand given the information we have.

In this generality, it may happen that for a given model $\mathcal{P}_0$ there is more than one maximum entropy distribution or that, on the contrary, there is no probability distribution at all which would maximize entropy over $\mathcal{P}_0$.

1.3.2 Maximum entropy under linear constraints

Let p_{emp} be an empirical distribution, that is $p_{emp} \in \mathrm{M}_+^1(I \times O)$ with support in G. Let us assume that p_{emp} has finite entropy and that, for each constraint, the average number of violations under p_{emp} is finite. Accordingly, we suppose that

$$\mathrm{supp}(p_{emp}) \subseteq G, \tag{1.9a}$$

$$\mathrm{H}(p_{emp}) < \infty, \tag{1.9b}$$

$$\mathrm{E}_{emp}(c_j) = \sum_{(i,o)\in G} c_j(i,o) \cdot p_{emp}(i,o) < \infty \quad \text{for all } j \in \mathbb{N}_N. \tag{1.9c}$$

Denote by $\tilde{p}_{emp}$ the marginal distribution of p_{emp} on the set of inputs I, i. e. $\tilde{p}_{emp} := \partial_I p_{emp}$. Define two sets of probability distributions on $G \subseteq I \times O$ by

$$\mathcal{P} := \big\{p \in \mathrm{M}_+^1(G) \;\big|\; \partial_I p = \tilde{p}_{emp} \;\wedge\; \mathrm{H}(p) < \infty \;\wedge\; \mathrm{E}_p(c_j) < \infty \;\forall\, j \in \mathbb{N}_N\big\}, \tag{1.10a}$$

$$\mathcal{P}_0 := \big\{p \in \mathcal{P} \;\big|\; \mathrm{E}_p(c) \;=\; \mathrm{E}_{emp}(c)\big\}. \tag{1.10b}$$

Clearly, $\mathcal{P}_0 \subseteq \mathcal{P}$, and p_{emp} is in $\mathcal{P}_0$. Moreover, both $\mathcal{P}_0$ and $\mathcal{P}$ are convex subsets of the vector space $\ell^1(G)$ as well as $\ell_c^1(G)$, where

$$\ell_c^1(G) := \Big\{\phi\colon G \to \mathbb{R} \;\Big|\; \sum_{(i,o)\in G} c_j(i,o)\cdot\phi(i,o) < \infty \;\forall\, j \in \mathbb{N}_N\Big\}. \tag{1.11}$$

We are now looking for a distribution $p_* \in \mathcal{P}_0$ such that $\mathrm{H}(p_*)$ is maximal over $\mathcal{P}_0$. Then p_* would be a maximal entropy distribution under the linear condition $(*)$. As usual, in order to find p_* we apply Lagrange's method. To this end, define a real-valued function f on $\mathcal{P}$ by

$$f(p) := \;-\mathrm{H}(p), \quad p \in \mathcal{P}.$$

Since entropy is non-negative and finite on $\mathcal{P}$, the function f takes its values in $(-\infty, 0]$. As $\mathrm{H}(.)$ is strictly concave, f is strictly convex. Our aim is to minimize f over $\mathcal{P}_0$, that is to find a distribution $p_* \in \mathcal{P}_0$ such that $p_* = \min\{f(p) \mid p \in \mathcal{P}_0\}$.

By convexity of $\mathcal{P}_0$ and strict convexity of f, it follows that there can be at most one minimizing distribution p_* in $\mathcal{P}_0$. Next, define an $\mathbb{R}^N$-valued function g on $\mathcal{P}$ by

$$g(p) := \;\mathrm{E}_p(c) - \mathrm{E}_{emp}(c), \quad p \in \mathcal{P}.$$

Note that $\mathcal{P}_0 = \{p \in \mathcal{P} \mid g(p) = 0\}$, and for each $j \in \mathbb{N}_N$

$$g_j(p) \quad = \quad \sum_{(i,o)\in G} c_j(i,o)\cdot\big(p(i,o) - p_{emp}(i,o)\big), \quad p \in \mathcal{P},$$

is a real-valued affine-linear function. Instead of minimizing f over $\mathcal{P}_0$, following Lagrange's method, one minimizes a Lagrangian function $L : \mathcal{P} \times \mathbb{R}^N \to \mathbb{R}$ over $\mathcal{P}$, where L is defined as

$$L(p,\lambda) := \;f(p) \;+\; \langle\lambda, g(p)\rangle, \quad p \in \mathcal{P},\; \lambda \in \mathbb{R}^N.$$

Notice that, for each $\lambda \in \mathbb{R}^N$, $L(.,\lambda)$ is a convex real-valued function on the convex set $\mathcal{P}$. If $L(.,\lambda)$ attains its global minimum over $\mathcal{P}$ at p and if, in addition, $g(p) = 0$, then by Lagrange's lemma, see proposition B.1 in the appendix, p is also a minimum position of f on $\mathcal{P}_0$, that is f restricted to $\mathcal{P}_0$ attains its global minimum at $p = p_*$.

Clearly, the expectation operator g can be extended to an affine linear function on the entire vector space $\ell_c^1(G)$. To this end, set

$$\Gamma(\phi) := \sum_{(i,o)\in G} \phi(i,o)\cdot c(i,o), \quad \phi \in \ell_c^1(G). \tag{1.12}$$

Then $\Gamma - \mathrm{E}_{emp}(c)$ is the affine linear extension of g to $\ell_c^1(G)$. If $\mathrm{E}_{emp}(c)$ is a "regular value" of Γ in an algebraic sense and if there is a distribution $p_* \in \mathcal{P}_0$ such that f restricted to $\mathcal{P}_0$

attains its minimum at p_*, then a version of Lagrange's theorem guarantees the existence of a Lagrange multiplier λ such that $L(.,\lambda)$ attains its minimum over $\mathcal{P}$ at p_*.

Proposition B.2 is a suitable special case of Lagrange's theorem. The regularity condition requires that $\mathrm{E}_{emp}(c)$ be a relative interior point of $\Gamma(\mathcal{P})$, cf. definition A.2 in appendix A. Under that regularity assumption we have equivalence between the minimization of f over $\mathcal{P}_0$ and the parametrized minimization of $L(.,\lambda)$ over $\mathcal{P}$.

According to the fundamental theorem of convex optimization, see theorem B.1 in the appendix, a necessary *and sufficient* condition for $L(.,\lambda)$ to attain its minimum at $p \in \mathcal{P}$ is that

$$L'_+(p, q-p; \lambda) \;\geq\; 0 \quad \text{for all } q \in \mathcal{P}, \tag{1.13}$$

where $L'_+(p, q-p; \lambda)$ is the right-hand Gâteaux derivative of $L(.,\lambda)$ at p in direction $q-p$, see appendix B. Let us compute $L'_+(p, q-p; \lambda)$ for $p, q \in \mathcal{P}$, $\lambda \in \mathbb{R}^N$. We have

$$\begin{aligned} L'_+(p, q-p; \lambda) &= \lim_{t\downarrow 0} \frac{1}{t} \Big(\mathrm{H}(p) - \mathrm{H}\big(p+t(q-p)\big) \;+\; \big\langle \lambda, \mathrm{E}_{p+t(q-p)}(c) - \mathrm{E}_p(c) \big\rangle \Big) \\ &= -\,\mathrm{H}'_+(p, q-p) \;+\; \sum_{(i,o)\in G} (q-p)(i,o) \cdot \big\langle \lambda, c(i,o) \big\rangle, \end{aligned} \tag{1.14}$$

where the integrability of c under p and q, respectively, has been exploited. As to the Gâteaux derivative of entropy, it holds that

$$\begin{aligned} \mathrm{H}'_+(p, q-p) &= \lim_{t\downarrow 0} \frac{1}{t} \Big(\mathrm{H}\big(p + t(q-p)\big) \;-\; \mathrm{H}(p) \Big) \\ = \lim_{t\downarrow 0} \frac{1}{t} \Bigg(& \sum_{q(i,o)>0} t \cdot q(i,o) \cdot \ln\Big(\tfrac{1}{(1-t)p(i,o)+tq(i,o)}\Big) \\ & + \sum_{p(i,o)>0} (1-t)p(i,o) \cdot \ln\Big(\tfrac{1}{(1-t)p(i,o)+tq(i,o)}\Big) \;-\; \sum_{p(i,o)>0} p(i,o) \cdot \ln\Big(\tfrac{1}{p(i,o)}\Big) \Bigg) \\ & \Big| \text{ rearrangement o. k. because of inequality (1.8)} \\ = \lim_{t\downarrow 0} \Bigg(& \sum_{q(i,o)>0} q(i,o) \cdot \ln\Big(\tfrac{1}{(1-t)p(i,o)+tq(i,o)}\Big) \;-\; \sum_{p(i,o)>0} p(i,o) \cdot \ln\Big(\tfrac{1}{(1-t)p(i,o)+tq(i,o)}\Big) \\ & - \sum_{p(i,o)>0} \frac{p(i,o)}{t} \cdot \ln\Big(1 + t\tfrac{q(i,o)-p(i,o)}{p(i,o)}\Big) \Bigg) \\ & \Big| \text{ apply monotone convergence to first sum, dominated conv. to last two sums} \\ = & \sum_{q(i,o)>0} q(i,o) \cdot \ln\Big(\tfrac{1}{p(i,o)}\Big) \;-\; \sum_{p(i,o)>0} p(i,o) \cdot \ln\Big(\tfrac{1}{p(i,o)}\Big) \;-\; 0. \end{aligned}$$

Observe that $\mathrm{H}'_+(p, q-p)$ exists as an element of $(-\infty, \infty]$. If we had $\mathrm{H}'_+(p, q-p) = \infty$ for some $q \in \mathcal{P}$, then equation (1.14) would imply that there could be no $\lambda \in \mathbb{R}^N$ satisfying

condition (1.13). Thus, in order for $\mathrm{H}'_+(p, q-p)$ to be finite for all $q \in \mathcal{P}$, it is necessary (though not sufficient) that p has full support. Accordingly, we assume that $\mathrm{supp}(p) = G$. Then

$$\mathrm{H}'_+(p, q-p) \quad = \quad -\sum_{(i,o)\in G} (q-p)(i,o) \cdot \ln\big(q(i,o)\big), \tag{1.15}$$

where the sum still has to be understood as a limit in $\mathbb{R} \cup \{\infty\}$. Plunging (1.15) into equation (1.14) yields

$$\begin{aligned} L'_+(p, q-p; \lambda) \quad &= \quad -\mathrm{H}'_+(p, q-p) \; + \sum_{(i,o)\in G} (q-p)(i,o) \cdot \big\langle \lambda, c(i,o) \big\rangle \\ &= \sum_{(i,o)\in G} (q-p)(i,o) \cdot \Big(\ln\big(p(i,o)\big) + \big\langle \lambda, c(i,o) \big\rangle \Big) \\ &= \sum_{i\in I} \tilde{p}_{emp}(i) \sum_{o\in O_i} \big(q(o|i) - p(o|i)\big) \cdot \Big(\ln\big(p(o|i)\big) + \big\langle \lambda, c(i,o) \big\rangle + \ln\big(\tilde{p}_{emp}(i)\big) \Big). \end{aligned}$$

By definition of $\mathcal{P}$, it holds that

$$\sum_{o\in O_i} p(o|i) \; = \; \tilde{p}_{emp}(i) \; = \; \sum_{o\in O_i} q(o|i) \quad \text{for all } i \in I. \tag{1.16}$$

Therefore

$$L'_+(p, q-p; \lambda) \quad = \quad \sum_{i\in I} \tilde{p}_{emp}(i) \sum_{o\in O_i} \big(q(o|i) - p(o|i)\big) \cdot \Big(\ln\big(p(o|i)\big) + \big\langle \lambda, c(i,o) \big\rangle \Big). \tag{1.17}$$

Relation (1.16) leads to the following reformulation of condition (1.13). Suppose we had

$$\ln\big(p(o_1|i)\big) \; + \; \big\langle \lambda, c(i,o_1) \big\rangle \; = \; \ln\big(p(o_2|i)\big) \; + \; \big\langle \lambda, c(i,o_2) \big\rangle \quad \text{for all } o_1, o_2 \in O_i, \; i \in I. \tag{1.18}$$

Then equation (1.16) would imply $L'_+(p, q-p; \lambda) = 0$ for all $q \in \mathcal{P}$. If, on the other hand, there were distinct elements o_1, o_2 in O_i for some $i \in I$ such that

$$\ln\big(p(o_1|i)\big) \; + \; \big\langle \lambda, c(i,o_1) \big\rangle \; \neq \; \ln\big(p(o_2|i)\big) \; + \; \big\langle \lambda, c(i,o_2) \big\rangle,$$

then we could find $q \in \mathcal{P}$ which would render $L'_+(p, q-p; \lambda)$ negative.[9] Therefore, conditions (1.18) and (1.13) are really equivalent.

Define constants in $(0, \infty]$ by

$$Z_r(i) := \sum_{\tilde{o}\in O_i} \exp\big(-\langle r, c(i,\tilde{o}) \rangle\big), \quad r \in \mathbb{R}^N, \; i \in I. \tag{1.19}$$

Let $\mathcal{R}$ denote the set of all parameters r such that $Z_r(i)$ is finite for all $i \in I$, that is

$$\mathcal{R} := \; \big\{ r \in \mathbb{R}^N \;\big|\; Z_r(i) < \infty \; \forall \, i \in I \big\}. \tag{1.20}$$

[9] Such a distribution q could be taken equal to p on $G \setminus \{(i,o_1), (i,o_2)\}$. By choosing $q(i,o_1)$ either smaller or bigger than $p(i,o_1)$ and setting $q(i,o_2) := p(i,o_1) + p(i,o_2) - q(i,o_1)$ one would achieve $L'_+(p, q-p; \lambda) < 0$.

We notice that if $r \in \mathcal{R}$ and $\tilde{r} \in [0,\infty)^N$ then also $r + \tilde{r} \in \mathcal{R}$. For $r \in \mathcal{R}$ we define a probability distribution on each of the sets O_i, $i \in I$, and a compound distribution conditional on $\tilde{p}_{emp}$ on G by

$$p_r(o|i) := \frac{1}{Z_r(i)} \exp\big(-\langle r, c(i,o)\rangle\big), \quad o \in O_i, \tag{1.21a}$$

$$p_r(i,o) := p_r(o|i) \cdot \tilde{p}_{emp}(i), \quad (i,o) \in G. \tag{1.21b}$$

We see that $Z_r(i)$ is a normalizing constant for the *Gibbs distribution* $p_r(.|i)$ on O_i. If the Lagrange multiplier λ is in $\mathcal{R}$, then the compound Gibbs distribution p_λ is in $\mathcal{P}$ and condition (1.18) is fulfilled with $p = p_\lambda$.

Conversely, if a probability distribution $p \in \mathcal{P}$ satisfies condition (1.18), where $\lambda \in \mathcal{R}$, then p must be of the form (1.21b); more precisely, $p = p_\lambda$ must hold.

Lastly, if $\lambda \notin \mathcal{R}$, then condition (1.18) will not be satisfied for any $p \in \mathcal{P}$. We have thus established the following result.

Theorem 1.1. *Suppose the sets I, O of inputs and outputs, respectively, the generator $G \subseteq I \times O$ and the feature function $c : G \to \mathbb{R}^N$ meet the requirements of section 1.1.1. Let the set $\mathcal{P}$ of probability distributions on G of finite entropy and finite expectation for c be given by* (1.10a)*, and define $\ell^1_c(G)$ according to* (1.11)*.*

Let $p_{emp} \in \mathcal{P}$ be given. Define the probability model $\mathcal{P}_0$ according to (1.10b)*. Let $\Gamma : \ell^1_c \to \mathbb{R}^N$ be as given by* (1.12)*; in particular, $\mathcal{P}_0 = \{p \in \mathcal{P} \mid \Gamma(p) = \mathrm{E}_{emp}(c)\}$. Let $\mathcal{R}$ be the set of parameters determined by* (1.20)*, and for $r \in \mathcal{R}$ let p_r denote the Gibbs distribution conditional on p_{emp} as given by* (1.19) *and* (1.21)*.*

If there is $r \in \mathcal{R}$ such that $p_r \in \mathcal{P}_0$, then p_r is the maximum entropy distribution of $\mathcal{P}_0$. To obtain the converse statement, suppose that $\mathrm{E}_{emp}(c)$ is a relative interior point of $\Gamma(\mathcal{P})$. Under this assumption, if the model $\mathcal{P}_0$ possesses a maximum entropy distribution p_*, then $p_* = p_r$ for some $r \in \mathcal{R}$.

We conclude this subsection by making two observations regarding the regularity condition on $\mathrm{E}_{emp}(c)$ in theorem 1.1.

First, suppose $\mathrm{E}_{emp}(c)$ is a relative interior point of $\Gamma(\mathcal{P})$ and the model $\mathcal{P}_0$ has a maximum entropy distribution p_*. From a convexity argument we already know that p_* is uniquely determined. According to theorem 1.1, we have $p_* = p_r$ for some $r \in \mathcal{R}$. The Gibbs parameter r, however, is not necessarily unique; it may happen that $p_* = p_r = p_{\hat{r}}$ for two distinct elements r, $\hat{r}$ of $\mathcal{R}$. Whether this is possible depends on the feature function c.[10]

As to the second observation, suppose the regularity condition on $\mathrm{E}_{emp}(c)$ is not met. Then $\mathcal{P}_0$ might still possess a maximum entropy distribution p_*, only that we have no guarantee for p_* to be of the Gibbs form (1.21). Yet we may hope to be able to obtain p_* as a limit of Gibbs distributions p_{r_n} for some parameter sequence $(r_n)_{n\in\mathbb{N}} \subset \mathcal{R}$.[11]

[10] The constraints must be such that c is "separating", cf. section 2.2.

[11] Cf. footnote 2 in Berger et al. (1996); attention there is restricted to probability distributions with finite support.

As a simple example for the second point consider the case of a one-dimensional constraint function c, just one input i and exactly two possible outputs o_1, o_2. Suppose $c(i, o_1) = 0$, $c(i, o_2) = 1$, and assume that the empirical distribution p_{emp} is concentrated on (i, o_1). Let $(r_n)_{n\in\mathbb{N}}$ be a sequence of real numbers such that $r_n \to \infty$ as n tends to infinity. Then p_{r_n} converges to p_{emp} as n tends to infinity, but there is no maximum entropy distribution of Gibbs form for p_{emp}, because $\mathrm{E}_{emp}(c) = 0$, while $\mathrm{E}_{p_r}(c) > 0$ for all ranking parameters $r \in \mathbb{R}$. We notice that p_{emp} has entropy zero.

1.3.3 The dual optimization problem

In section 1.3.2 we have seen that the entropy maximization problem under linear constraints leads to the choice of a probability distribution from a family of Gibbs distributions $(p_r)_{r\in\mathcal{R}}$, where p_r is given by (1.21) and $\mathcal{R}$ by (1.20). The set of parameters $\mathcal{R} \subset \mathbb{R}^N$ can readily be interpreted as a set of constraint rankings in the sense of section 1.1.

At this point, we have to find a way of how to determine a maximum entropy parameter $r \in \mathcal{R}$, that is a parameter r such that p_r is the maximum entropy distribution for the model $\mathcal{P}_0$ as given by (1.10b). To this purpose, define the function $f_{pot} : \mathbb{R}^N \to \mathbb{R} \cup \{\infty\}$ by

$$f_{pot}(r) := \begin{cases} -\mathrm{E}_{emp}(\ln p_r), & \text{if } r \in \mathcal{R}, \\ \infty, & \text{else.} \end{cases} \tag{1.22}$$

In particular, for all $r \in \mathcal{R}$ we have

$$\begin{aligned} f_{pot}(r) &= \big\langle r, \mathrm{E}_{emp}(c)\big\rangle + \sum_{i\in I} \tilde{p}_{emp}(i) \cdot \ln Z_r(i) \\ &= \sum_{(i,o)\in G} p_{emp}(i,o) \cdot \big(\langle r, c(i,o)\rangle + \ln Z_r(i)\big). \end{aligned} \tag{1.23}$$

Denote by $\mathrm{int}(\mathcal{R})$ the interior of $\mathcal{R}$. The next proposition lists some properties of f_{pot}.

Proposition 1.1. *Define the function f_{pot} on $\mathbb{R}^N$ according to* (1.22)*. Then it holds that*

(i) *f_{pot} takes its values in $[0, \infty]$, and*

$$f_{pot}(r) = \mathrm{D}(p_{emp}\|p_r) + \mathrm{H}(p_{emp}) \quad \textit{for all } r \in \mathcal{R},$$

where $\mathrm{D}(.\|.)$ is the Kullback-Leibler divergence,

(ii) *f_{pot} is convex,*

(iii) *f_{pot} is twice continuously differentiable on $\mathrm{int}(\mathcal{R})$.*

The first and second order partial derivatives of f_{pot} at $r \in \mathrm{int}(\mathcal{R})$ are

$$\frac{\partial}{\partial r_j} f_{pot}(r) = \mathrm{E}_{emp}(c_j) - \mathrm{E}_{p_r}(c_j), \qquad j \in \{1, \ldots, N\}, \tag{1.24a}$$

$$\frac{\partial^2}{\partial r_j \partial r_k} f_{pot}(r) = \mathrm{E}_{\tilde{p}_{emp}}\left(\mathrm{cov}_{p_r(.|i)}\big(c_j(i,.), c_k(i,.)\big)\right), \quad j, k \in \{1, \ldots, N\}. \tag{1.24b}$$

Proof. Recall from section 1.3.1 how the Kullback-Leibler divergence $\mathrm{D}(.\|.)$ was defined. The identity for f_{pot} is established by rewriting equation (1.23). Kullback-Leibler divergence as well as entropy are non-negative. Convexity is a consequence of Hölder's inequality and the monotonicity of the logarithm.

Termwise differentiation in equation (1.23) leads to the right-hand sides in (1.24). Notice that the right-hand side of (1.24b) written out reads

$$\sum_{i\in I} \tilde{p}_{emp}(i) \left(\Big(\sum_{o\in O_i} c_j(i,o)\cdot c_k(i,o)\cdot p_r(o|i) \Big) \;-\; \Big(\sum_{o\in O_i} c_j(i,o)\cdot p_r(o|i) \Big) \cdot \Big(\sum_{o\in O_i} c_k(i,o)\cdot p_r(o|i) \Big) \right).$$

The sums appearing in (1.23) and (1.24), respectively, correspond to limits of sequences that are non-decreasing; Dini's theorem establishes uniform convergence of the partial sums, which in turn guarantees differentiability of the original function. □

Theorem 1.1 and (1.24a) together imply that if f_{pot} possesses a minimum at $\hat{r} \in \mathcal{R}$ then $\mathcal{P}_0$ possesses a maximum entropy distribution p_* and $p_* = p_{\hat{r}}$. Minimizing the function f_{pot} over $\mathcal{R}$ thus constitutes the dual to the problem of maximizing entropy over the probability model $\mathcal{P}_0$. This fact has been exploited in the design of Jäger's learning algorithm and will be essential for the proofs of convergence in chapter 2.

From proposition 1.1 we see that minimizing the dual function f_{pot} amounts to minimizing the Kullback-Leibler distance or relative entropy between p_{emp} and the set of Gibbs distributions $\{p_r \mid r \in \mathcal{R}\}$. The value $f_{pot}(r)$ is the *cross entropy* between p_{emp} and p_r. Minimization of cross entropy or, equivalently, relative entropy is a standard procedure in computational linguistics (cf. Manning and Schütze, 1999: pp. 73-77).

Let us also mention that the maximum entropy distribution – provided there is any – is at the same time a maximum likelihood estimate of the empirical distribution.[12]

To conclude this section, we list the names used in physics for the quantities introduced above. We may assume there is exactly one input i associated with only a finite number of possible outputs, that is O_i is finite and coincides with the set of outputs O. The case of more than one input can be regarded as a mixture of systems, each system corresponding to one input and its output candidates.

Gibbs distribution (1.21a) gives the probability distribution of a system allowing for $\#O_i$ microstates in thermal equilibrium at *temperature* T, where state (i,o) – or just o – has *energy* $kT\langle r, c(i,o)\rangle$. Here, k is *Boltzmann's constant*. The normalizing constant $Z_r(i)$ is called *partition function* or *Zustandssumme*.

Relative entropy $D(q\|p)$ measures the difference in *free energy* between an arbitrary probability distribution q and the equilibrium distribution p. This seems to fit well our needs: setting $q := p_{emp}$ and $p := p_r$, where p_r is of Gibbs form, hence an equilibrium distribution, we find that f_{pot} corresponds (up to an additive constant) to the difference in free energy between p_{emp} and p_r. Note, however, that we have to minimize $f_{pot}(r)$ over $r \in \mathbb{R}^N$. This cannot be interpreted as merely adjusting temperature. A different

[12] For details see Berger et al. (1996), for example.

interpretation of f_{pot} can be found in Jaynes (1957a: p. 624), where an ensemble made up of particles of N different types is considered.

1.4 Jäger's algorithm for maximum entropy learning

Assume we are given a generator $G \subseteq I \times O$ together with a feature function $c\colon G \to \mathbb{R}^N$. We choose an evaluation kernel according to the maximum entropy version of stochastic OT. To this end, let $\mathcal{R}$ be the set of Gibbs parameters as defined in (1.20). In the convergence analysis of chapter 2 we will assume c and G are such that $\mathcal{R} = \mathbb{R}^N$. In chapter 3, especially in section 3.2, we will discuss implications of this assumption and what should be done in case we have $\mathcal{R} = (0, \infty)^N$.

The family $(p_r(.|i))_{i\in I, r\in\mathcal{R}}$ of probability distributions defined according to (1.19) and (1.21a) gives an evaluation kernel for G and c, and (G, c, p) is a universal stochastic OT grammar. Given observations of an empirical distribution p_{emp} on G, we need a procedure for selecting a specific stochastic OT grammar that best reflects p_{emp}.

Jäeger's idea for such a procedure (Jäger, 2003) is similar to what is behind Boersma's GLA. Start with an arbitrary constraint ranking. Each time an input output pair is observed, draw an output for the observed input according to the current ranking of constraints. If the computed output is different from the one observed, compare the number of constraint violations incurred by the two outputs. Increase the rank of those constraints favouring the observed output over the generated one, decrease the rank of those constraints having the opposite effect (cf. section 1.2). The amount of rank promotion and demotion, respectively, is proportional to the difference in the number of constraint violations. It also depends on a gain parameter regulating the step size of the algorithm.[13]

As in section 1.1.2, let (X_n) be an i. i. d. sequence of input output pairs on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ such that $X_n = (X_n^{in}, X_n^{out})$ has distribution p_{emp} under P. The random sequence (X_n) thus represents the observations of the empirical distribution available to the learner. Let $(H_n)_{n\in\mathbb{N}_0}$ be a sequence of measurable mappings $\Omega \times I \times \mathbb{R}^N \to O$ such that $(H_n(.,i,r))_{n\in\mathbb{N}_0}$ is an i. i. d. random sequence with distribution $p_r(.|i)$ under P for all $i \in I$, $r \in \mathbb{R}^N$. The random variables $H_n(.,i,r)$ mimic the action of the evaluation component, which generates outputs $o \in O_i$ given an input i and a ranking vector r.

Jäger's algorithm for learning constraint rankings can now be expressed as follows. Choose a gain parameter $\eta > 0$ and an initial constraint ranking, that is an $\mathcal{F}$-measurable random variable R_0^η with values in $\mathbb{R}^N$ which is assumed to be independent of the other random variables. Of course, R_0^η might be constant, corresponding to a deterministic initial ranking. Compute a sequence $(R_n^\eta)_{n\in\mathbb{N}_0}$ of constraint rankings according to the recursion formula

$$R_{n+1}^\eta := R_n^\eta + \eta \cdot \big(c(X_n^{in}, H_n(X_n^{in}, R_n^\eta)) - c(X_n^{in}, X_n^{out})\big), \quad n \in \mathbb{N}_0. \tag{1.25}$$

[13] In section 1.2 the gain parameter was referred to as plasticity value. For rank promotion or demotion the GLA takes into account only the sign of the difference in the number of violations of a certain constraint, not the value of that difference.

If the initial ranking is deterministic, i. e. $R_0^\eta = r$ for some $r \in \mathcal{R}$, then $(R_n^\eta)_{n\in\mathbb{N}_0}$ as defined by (1.25) is a sequence of random variables on $(\Omega, \mathcal{F}, \mathrm{P})$ taking values in the countable space $r + \eta\mathbb{Z}^N$. The sequence of constraint rankings thus defines a random walk on an N-dimensional Euclidean grid. In order to obtain an implementable algorithm, we would have to specify a halting condition, e. g. to stop computation after a certain fixed number of iterations.

A glance at recursion formula (1.25) shows that Jäger's algorithm and Boersma's GLA are quite similar. Indeed, if all constraints are $\{0, 1\}$-valued, recursion formula (1.25) is equivalent to formula (1.4). Nevertheless, the two learning procedures are essentially different due to the fact that they work with two different methods for output evaluation or, more generally, with two different versions of stochastic OT .

Constraint rankings are learned if the sequence (R_n^η) converges in some sense. Learning is successful if the limit of the sequence of constraint rankings is close to the minimum position of f_{pot} provided η is small. In chapter 2, we will find an appropriate notion of convergence and give a precise meaning to what "close to" should be.

In the convergence proofs of chapter 2 the following observation will play a fundamental rôle. Recall from section 1.3.3 that entropy maximization is linked to the dual optimization problem of minimizing the function f_{pot} as given by (1.22). The sequence $(R_n^\eta)_{n\in\mathbb{N}_0}$ defined above can be regarded as a process in discrete time. The function f_{pot} is a *potential* for this process in the sense that for all $n \in \mathbb{N}_0$ we have[14]

$$\mathrm{E}\big(R_{n+1}^\eta - R_n^\eta \mid R_n^\eta\big) \quad = \quad -\eta \cdot \nabla f_{pot}(R_n^\eta), \tag{1.26}$$

where $\eta > 0$ is the gain parameter inducing a certain time scale. In view of (1.26), if no random elements were involved, then Jäger's algorithm would reduce to an application of the gradient descent method for finding the minimum of the function f_{pot}. As we are confronted with a stochastic algorithm, we have to show that the "mean dynamics" described by equation (1.26) are predominant at least for small values of the gain parameter.

[14] Cf. Jäger (2003) for a less formal way of stating (1.26).

Chapter 2

Convergence of Jäger's algorithm

The central theme of this chapter are proofs of convergence for Jäger's algorithm under varying assumptions. As we are dealing with a stochastic algorithm, we have to distinguish between different notions of convergence. Definitions of convergence which come into question in the present context are summarized in section 2.1. The notion of weak convergence turns out to be the one that is most appropriate to the problem at hand.

Since Jäger's learning algorithm is, in effect, a procedure for finding the minimum positions of the function f_{pot} associated with the empirical distribution p_{emp} and the feature function c, we have to study the interrelation between the minima of f_{pot} on the one hand and the distribution p_{emp} and feature function c on the other. This is the purpose of section 2.2.

In section 2.3 we prove convergence of the sequence of constraint rankings to a probabilistic counterpart of an equilibrium or steady state, namely convergence to a stationary distribution. We will make use of the discrete nature of the constraint rankings arising from the hypothesis that the feature function takes as its values only non-negative integers. Section 2.3.1 is a selection of results from the theory of discrete Markov chains. It also helps in providing insight into some of the ideas of stability theory. The theory is put into application in section 2.3.2.

In the remaining sections we will rely on more general results concerning iterative stochastic algorithms and approximation, which will be stated only briefly. A comprehensive exposition can be found in Kushner and Yin (2003).

Section 2.4 deals with convergence of the stationary distributions as the gain parameter η, which up to that point shall be considered a constant, approaches zero. In section 2.4.1 we present a general approximation result which relates Markov chains to deterministic and stochastic differential equations. So equipped, we will return to Jäger's algorithm in section 2.4.2.

In section 2.5 we will not wait for the algorithm to settle in the stationary regime; instead we let η change with each iteration according to some suitable annealing scheme. The situation there is similar to the one studied in the classical work of Robbins and Monro (1951), although we will not rely on that result.

The main idea for the proofs of convergence in section 2.3, and also in section 2.5, is that f_{pot} is not only the function we have to minimize; it also serves as a Lyapunov function. Consider a path or trajectory of rankings as produced by Jäger's algorithm, i. e. a realization of the random sequence of constraint rankings. The function f_{pot} is a Lyapunov function in that it decreases along trajectories which have gone too far away from the position of the global minimum.[1] We will see that the problem is indeed reducible to the case when f_{pot} has exactly one (local and global) minimum at a unique position. The fact that f_{pot} decreases along deviating trajectories implies that – at least on average – trajectories cannot "run away" from the position of the minimum.

Throughout this chapter we assume that the parameter set $\mathcal{R}$ as given by (1.20) in section 1.3.2 coincides with $\mathbb{R}^N$. An immediate consequence of this hypothesis is that there can be only finitely many outputs for a given input.[2] In the notation of section 1.1 this means that O_i is finite for all $i \in I$. The set O of all outputs may nevertheless be infinite, because we still allow for a (countably) infinite set of inputs. Notice that the number of elements of O_i is not required to be uniformly bounded in $i \in I$.

2.1 Notions of stochastic convergence

Here, we collect a number of definitions that describe the convergence behaviour of a sequence of random variables.[3] For each $n \in \mathbb{N}$ let X_n be an $\mathbb{R}^N$-valued random variable on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. We are interested in conditions under which the sequence (X_n) can reasonably be taken to converge to an $\mathbb{R}^N$-valued random variable X defined on the same probability space.

The first notion is that of almost sure convergence. In that case, we have convergence in $\mathbb{R}^N$ for "almost all" scenarios ω, that is for all $\omega \in \Omega$ not in a set of probability zero.

Definition 2.1. The sequence $(X_n)_{n\in\mathbb{N}}$ is said to *converge* P-*almost surely* to a random variable X iff there is set $N \in \mathcal{F}$ such that

(i) $\mathrm{P}(N) = 0$,

(ii) $X_n(\omega) \overset{n\to\infty}{\longrightarrow} X(\omega)$ for all $\omega \in \Omega \setminus N$.

A set with probability zero is also called a *null set* or, more precisely, a P-null set, where P is the underlying probability measure. The requirement of "scenariowise" convergence can be weakened in the following way.

Definition 2.2. The sequence $(X_n)_{n\in\mathbb{N}}$ is said to *converge in probability* to a random variable X iff it holds that

$$\mathrm{P}\big(\{\omega \in \Omega \mid \|X_n(\omega) - X(\omega)\| \geq \varepsilon\}\big) \overset{n\to\infty}{\longrightarrow} 0 \quad \text{for each } \varepsilon > 0.$$

[1] More precisely, f_{pot} need not decrease along all trajectories which have left some neighbourhood of the minimum position, but it will do so on average.

[2] See section 3.2 for a discussion of the case of infinitely many output alternatives for a given input.

[3] The definitions we cite are standard; see Bauer (1991), for example.

Almost sure convergence implies convergence in probability, while the converse is not true in general. However, a sequence that converges in probability has the property that any of its subsequences contains in turn a subsequence which is almost surely convergent. Definitions 2.1 and 2.2 can both be generalized to work for random variables taking values in arbitrary metric or topological spaces.

If the random variables involved have finite pth moments, then convergence with respect to those averages can be defined.

Definition 2.3. Let $\infty > p \geq 1$. Suppose that $\mathrm{E}(\|X\|^p) < \infty$ and $\mathrm{E}(\|X_n\|^p) < \infty$ for all $n \in \mathbb{N}$. The sequence $(X_n)_{n\in\mathbb{N}}$ is said to *converge in L^p* to a random variable X iff it holds that

$$\mathrm{E}\big(\|X_n - X\|^p\big) \overset{n\to\infty}{\longrightarrow} 0.$$

If $p = 2$ one also speaks of *mean square convergence*. Convergence in L^p implies convergence in probability. Observe that L^p-convergence requires integrability of the random variables $\|X_n\|^p$ and $\|X\|^p$.

There is an even weaker notion of convergence than that of convergence in probability. It only depends on the distributions of the random variables. Assume X_n, $n \in \mathbb{N}$, X take values in the measurable space $(S, \mathcal{B}(S))$, where S is a topological space and $\mathcal{B}(\mathcal{S})$ its Borel σ-algebra, i. e. the system of measurable sets generated by the open subsets of S.

Definition 2.4. The sequence $(X_n)_{n\in\mathbb{N}}$ is said to *converge in distribution* to a random variable X iff it holds that

$$\mathrm{E}\big(\phi(X_n)\big) \overset{n\to\infty}{\longrightarrow} \mathrm{E}\big(\phi(X)\big) \quad \text{for all bounded and continuous functions } \phi\colon S \to \mathbb{R}.$$

In case of convergence, let us write $X_n \overset{w}{\to} X$ as $n \to \infty$.

Convergence in distribution makes sense even if the random variables are defined on different probability spaces. To see that definition 2.4 is really a notion of convergence for the accompanying distributions, let Y be an S-valued random variable on $(\Omega, \mathcal{F}, P)$ and notice that

$$\mathrm{E}\big(\phi(Y)\big) \;=\; \int_S \phi(y)\, d\,\mathrm{P}_Y(y),$$

where ϕ is any appropriate function on S (e. g. bounded and measurable) and P_Y is the distribution of Y under P. Thus, P_Y is a probability measure on $\mathcal{B}(S)$. Accordingly, one defines convergence of probability measures or distributions on the state space S.

Definition 2.5. The sequence $(\mathrm{P}_n)_{n\in\mathbb{N}}$ of probability measures on $\mathcal{B}(\mathcal{S})$ is said to *converge weakly* to a probability measure P on $\mathcal{B}(\mathcal{S})$ iff it holds that

$$\int_S \phi(y)\, d\,\mathrm{P}_n(y) \overset{n\to\infty}{\longrightarrow} \int_S \phi(y)\, d\,\mathrm{P}(y) \quad \text{for all bounded continuous functions } \phi\colon S \to \mathbb{R}.$$

In case of convergence, let us write $\mathrm{P}_n \overset{w}{\to} \mathrm{P}$.

Now consider the case that S is discrete. If we endow S with the discrete topology, then any function on S is continuous, and weak convergence is equivalent to the – in general stronger – notion of convergence in total variation.

Definition 2.6. Let ν_n, $n \in \mathbb{N}$, and μ be distributions on the countable set S. The sequence (ν_n) *converges in total variation* to μ iff

$$\lim_{n\to\infty} \sum_{i\in S} |\nu_n(i) - \mu(i)| \;=\; 0.$$

All we really need is the notion of weak convergence. The reason is that we are interested in distributions of constraint rankings, not in the evolution of the sequence of constraint rankings given a particular sequence of input output pairs.

From definition 2.5 it is evident that weak convergence is preserved under continuous transformations. If (X_n) converges weakly to X and ψ is a continuous mapping $S \to \tilde{S}$, where $\tilde{S}$ is a second topological space, then the transformed sequence $(\psi(X_n))$ converges weakly to $\psi(X)$, the transformation of X. As a consequence, we have the following continuity property of maximum entropy OT grammars.

Proposition 2.1. *Given an empirical distribution p_{emp} and a feature function c on G, let the parameter set $\mathcal{R}$ be as defined by* (1.20)*. Assume that $\mathcal{R} = \mathbb{R}^N$. In particular, O_i is supposed to be finite for each $i \in I$. Let the family of Gibbs distributions p_r, $r \in \mathbb{R}^N$, be as defined by* (1.21)*. Let R_n, $n \in \mathbb{N}$, R be $\mathbb{R}^N$-valued random variables. Then $R_n \xrightarrow{w} R$ implies $p_{R_n} \xrightarrow{w} p_R$ as n tends to infinity.*

Proof. For $n \in \mathbb{N}$ let $\tilde{\mathrm{P}}_n$ denote the probability measure induced by R_n, and let $\tilde{\mathrm{P}}$ denote the probability measure induced by R. Let $\psi\colon \mathrm{M}_+^1(G) \to \mathbb{R}$ be a bounded and continuous function, where we choose as topology for $\mathrm{M}_+^1(G)$ the topology of weak convergence.[4] We have to show that

$$\int_{\mathbb{R}^N} \psi(p_r)\, d\tilde{\mathrm{P}}_n(r) \;\overset{n\to\infty}{\longrightarrow}\; \int_{\mathbb{R}^N} \psi(p_r)\, d\tilde{\mathrm{P}}(r).$$

By hypothesis, we know that

$$\int_{\mathbb{R}^N} \phi(r)\, d\tilde{\mathrm{P}}_n(r) \;\overset{n\to\infty}{\longrightarrow}\; \int_{\mathbb{R}^N} \phi(r)\, d\tilde{\mathrm{P}}(r)$$

for all bounded and continuous functions $\phi\colon \mathbb{R}^N \to \mathbb{R}$. Since ψ is bounded, the function $\mathbb{R}^N \ni r \mapsto \psi(p_r)$ is bounded, too. Hence, it is sufficient to show that $r \mapsto \psi(p_r)$ is continuous. Let $(r_n) \subset \mathbb{R}^N$ be a sequence of real vectors such that $r_n \to r$ as n tends to infinity. Convergence of the sequence (r_n) to r must imply weak convergence $p_{r_n} \xrightarrow{w} p_r$ of the corresponding Gibbs distributions. Since G is countable, we will check that

$$\sum_{(i,o)\in G} \left|p_{r_n}(i,o) - p_r(i,o)\right| \;\overset{n\to\infty}{\longrightarrow}\; 0. \tag{2.1}$$

[4] Since G is countable, the topology of weak convergence coincides with that of convergence in total variation.

We have

$$\sum_{(i,o)\in G} \big|p_{r_n}(i,o) - p_r(i,o)\big| \quad = \quad \sum_{i\in I} \tilde{p}_{emp}(i) \sum_{o\in O_i} \big|p_{r_n}(o|i) - p_r(o|i)\big|$$

$$= \quad \sum_{i\in I} \tilde{p}_{emp}(i) \sum_{o\in O_i} \exp\Big(-\big\langle r, c(i,o)\big\rangle\Big) \cdot \left|\frac{1}{Z_{r_n}(i)} \exp\Big(-\big\langle r - r_n, c(i,o)\big\rangle\Big) - \frac{1}{Z_r(i)}\right|.$$

Clearly,

$$\exp\Big(-\big\langle r - r_n, c(i,o)\big\rangle\Big) \overset{n\to\infty}{\longrightarrow} 1 \quad \text{for all } (i,o) \in G.$$

Now O_i is supposed to be finite, whence for all $i \in I$

$$Z_{r_n}(i) \;=\; \sum_{\tilde{o}\in O_i} \exp\Big(-\big\langle r_n, c(i,\tilde{o})\big\rangle\Big) \overset{n\to\infty}{\longrightarrow} \sum_{\tilde{o}\in O_i} \exp\Big(-\big\langle r, c(i,\tilde{o})\big\rangle\Big) \;=\; Z_r(i).$$

The limit in equation (2.1) follows by dominated convergence. □

2.2 Minima of the dual function

Recall from section 1.3.3 the definition of the dual function f_{pot}. We already know that f_{pot} is a non-negative convex function; given an empirical distribution p_{emp} and a feature function c we want to check whether f_{pot} has a global minimum. Let p_r, $r \in \mathbb{R}^N$, denote the family of Gibbs distributions associated with c and p_{emp}.

We notice the linear dependence between the ranking vector r and the feature function c in the exponent appearing in (1.21a), the definition of the Gibbs distribution $p_r(.|i)$ on O_i conditional on the input $i \in I$. For this reason, the families $p_r(.|i)$, $r \in \mathbb{R}^N$, are also referred to as log-linear probability models.

Observe that Jäger's algorithm as introduced in section 1.4 is driven by the difference in the number of constraint violations between two outputs for the same input. Define two sets V_c, W_c of ranking vectors by

$$\text{(2.2a)} \quad V_c := \big\{r \in \mathbb{R}^N \;\big|\; \forall\,(i,o_1) \in \mathrm{supp}(p_{emp}), (i,o_2) \in G :\; \big\langle r,\, c(i,o_1) - c(i,o_2)\big\rangle = 0\big\},$$

$$\text{(2.2b)} \quad W_c := \big\{r \in \mathbb{R}^N \;\big|\; \forall\,(i,o_1) \in \mathrm{supp}(p_{emp}), (i,o_2) \in G :\; \big\langle r,\, c(i,o_1) - c(i,o_2)\big\rangle \le 0\big\}.$$

First, notice that V_c is a linear subspace of $\mathbb{R}^N$, while W_c, in general, is a convex subset of $\mathbb{R}^N$ containing V_c. It is, of course, possible that $V_c = W_c$. Indeed, this is certainly the case when $\mathrm{supp}(p_{emp}) = G$, that is whenever the empirical distribution has full support.

Recalling definitions (1.19) and (1.21) we find that the distributions p_r are invariant under translation along V_c, that is[5]

$$\text{(2.3)} \qquad p_r(i,o) \;=\; p_{r+v}(i,o) \quad \text{for all } r \in \mathbb{R}^N,\; v \in V_c,\; (i,o) \in G.$$

[5] Remember our assumption concerning the support of p_{emp}, namely that for each input $i \in I$ there is at least one output $o \in O_i$ such that $(i,o) \in \mathrm{supp}(p_{emp})$, cf. section 1.1.

As an immediate consequence we see that f_{pot} is constant along the affine-linear subspace $r + V_c$ for each $r \in \mathbb{R}^N$, that is

$$f_{pot}(r) \;=\; f_{pot}(r+v) \quad \text{for all } r \in \mathbb{R}^N,\ v \in V_c. \tag{2.4}$$

Denote by $V_c^\perp$ the orthogonal complement of V_c. Let (R_n^η) be a sequence of constraint rankings according to (1.25). It is clear that

$$R_{n+1}^\eta - R_n^\eta \in V_c^\perp \quad \text{P-a. s.} \quad \text{for all } n \in \mathbb{N}_0. \tag{2.5}$$

If the initial ranking is deterministic, i. e. $R_0^\eta = r$ for some $r \in \mathbb{R}^N$, then the sequence (R_n^η) of constraint rankings is contained in the affine-linear subspace $r + V_c^\perp$.

Let v be an element of V_c. If the sequence of constraint rankings with initial ranking r converges in distribution to a random variable $R(r)$, then the ranking sequence with initial constraint ranking $r + v$ converges in distribution to $v + R(r)$. Moreover, because of (2.3), the distribution-valued random variables $p_{R(r)}$ and $p_{v+R(r)}$ are equal.

Next, let $v \in \mathbb{R}^N \setminus W_c$, where we assume that $W_c \neq \mathbb{R}^N$. We find $(i, o_1) \in \operatorname{supp}(p_{emp})$, $(i, o_2) \in G$ such that $\langle v, c(i,o_1) - c(i,o_2)\rangle > 0$, and for all $t > 0$ it holds that

$$\begin{aligned} f_{pot}(t \cdot v) \quad &\geq \quad p_{emp}(i,o_1) \cdot t \cdot \big\langle v, c(i,o_1)\big\rangle \;+\; p_{emp}(i,o_1) \cdot \ln\!\left(\sum_{o \in O_i} \exp\Big(-t\big\langle v, c(i,o)\big\rangle\Big)\right) \\ &> \quad p_{emp}(i,o_1) \cdot t \cdot \Big(\big\langle v, c(i,o_1) - c(i,o_2)\big\rangle\Big) \end{aligned}$$

as a consequence of equation (1.23) and the strict monotonicity of the logarithm. By construction, $p_{emp}(i,o_1) > 0$ as well as $\langle v, c(i,o_1) - c(i,o_2)\rangle > 0$. Hence, we have

$$f_{pot}(t \cdot v) \;\to\; \infty \quad \text{as} \quad t \to \infty \quad \text{for all } v \in \mathbb{R}^N \setminus W_c. \tag{2.6}$$

Now, suppose that $V_c = W_c$. By (2.6), we see that $f_{pot}(t \cdot v) \to \infty$ as $t \to \infty$ for all $v \in V_c^\perp \setminus \{0\}$. By lemma A.1, f_{pot} restricted to $V_c^\perp$ possesses a unique global minimum.[6] Because of (2.4), the same holds true for the restrictions of f_{pot} to $r + V_c^\perp$ with $r \in \mathbb{R}^N$. The function f_{pot} itself attains its global minimum value on the entire affine-linear subspace $\hat{r} + V_c$, where $\hat{r} \in V_c^\perp$ is the position of the global minimum of f_{pot} restricted to $V_c^\perp$.

Lastly, suppose $V_c \neq W_c$. Then we would find ourselves in the situation that the regularity condition of theorem 1.1 on $\mathrm{E}_{emp}(c)$ is violated.[7] To check this, recalling the notation of section 1.3, we have to indicate $r \in \mathbb{R}^N$ such that $\mathrm{E}_{emp}(c) + t \cdot r$ is in the affine hull of $\Gamma(\mathcal{P})$ for all $t \in \mathbb{R}$, while $\mathrm{E}_{emp}(c) - t \cdot r \notin \Gamma(\mathcal{P})$ for $t > 0$.

Since p_{emp} is in $\mathcal{P}$ and $\Gamma(p)$ for $p \in \mathcal{P}$ is just the expectation of the feature function c under p, we see that $\operatorname{hull}_{\mathrm{aff}} \Gamma(\mathcal{P}) = \mathrm{E}_{emp}(c) + V_c^\perp$. As a consequence of our hypothesis that $V_c \neq W_c$, the intersection of W_c and $V_c^\perp \setminus \{0\}$ is non-empty. Thus, we are able to choose $r \in W_c \cap V_c^\perp \setminus \{0\}$. If there is $v \in \mathbb{R}^N$ such that for any $t > 0$ and all $p \in \mathcal{P}$

$$\big\langle \mathrm{E}_{emp}(c) - \mathrm{E}_p(c), v\big\rangle \;\neq\; t \cdot \langle r, v\rangle,$$

[6] To check strict convexity of f_{pot} on $V_c^\perp$ recall the condition for equality in Hölder's inequality.

[7] See end of section 1.3.2.

then we will have established $\mathrm{E}_{emp}(c) - t \cdot r \notin \Gamma(\mathcal{P})$ for $t > 0$; since $r \in V_c^{\perp} \setminus \{0\}$, the non-regularity of $\mathrm{E}_{emp}(c)$ would follow. As test vector v take r itself. Then $t \cdot \langle r, v\rangle = t \cdot \langle r, r\rangle > 0$ for all $t > 0$, while

$$\begin{aligned} & \big\langle \mathrm{E}_{emp}(c) - \mathrm{E}_p(c), r \big\rangle \\ = & \sum_{i \in I} \tilde{p}_{emp}(i) \left(\sum_{(i,o) \in \mathrm{supp}(p_{emp})} p_{emp}(o|i) \cdot \big\langle r, c(i,o) \big\rangle \; - \sum_{(i,o) \in G} p(o|i) \cdot \big\langle r, c(i,o) \big\rangle \right) \\ = & \sum_{i \in I} \tilde{p}_{emp}(i) \left(\kappa(i) \; - \sum_{(i,o) \in G} p(o|i) \cdot \big\langle r, c(i,o) \big\rangle \right) \quad \leq \quad 0, \end{aligned}$$

where $\kappa(i) := \langle r, c(i,o)\rangle$ for some $(i,o) \in \mathrm{supp}(p_{emp})$, which is well defined, because $r \in W_c$. By construction, $\kappa(i) \leq \langle r, c(i,o)\rangle$ for all $(i,o) \in G$.

In case $V_c \neq W_c$ we cannot expect the sequence (R_n^{η}) of constraint rankings to converge. Instead, we would have to study the sequence of induced Gibbs distributions.

In the sequel, however, we will generally assume that the expectation vector $\mathrm{E}_{emp}(c)$ is regular in the sense of being a relative interior point of $\mathrm{E}_{emp}(c) + V_c^{\perp}$.

2.3 Constant step size and convergence on a grid

Recall that Jäger's algorithm is driven by the difference in the number of constraint violations incurred by two alternative outputs for the same input. As a consequence, if the initial constraint ranking R_0^{η} is deterministic, then the random sequence (R_n^{η}) of constraint rankings defined in section 1.4 is an instance of a random walk on an N-dimensional grid. The evolution of this random walk is governed by a convex potential, given in our case by the function f_{pot}.

Notice that the mesh size of the N-dimensional grid is determined by the gain parameter η. In this section we take η to be a constant. Before proving convergence of the ranking sequence in section 2.3.2, we review a bit of theory.

2.3.1 Invariant distributions and Foster's drift criterion

Here, we collect a number of results concerning homogeneous Markov chains. Our main reference is Brémaud (1999), where proofs can be found.

Let $(X_n)_{n \in \mathbb{N}_0}$ be a random sequence on the measurable space $(\Omega, \mathcal{F})$ with values in a countable set S. Let $\mathbf{P} = (p_{ij})_{i,j \in S}$ be a transition matrix on S. For every distribution ν on S let P_{ν} be a probability measure on $\mathcal{F}$ such that $(X_n)_{n \in \mathbb{N}_0}$ is a homogeneous Markov chain under P_{ν} with transition matrix $\mathbf{P}$ and initial distribution ν, where the latter means that $\mathrm{P}_{\nu}(X_0 = i) = \nu(i)$ for all $i \in S$.

If ν is a point distribution concentrated at some $i \in S$, write P_i for the corresponding probability measure on $\mathcal{F}$; we then have $\mathrm{P}_i(X_0 = i) = 1$. When it does not matter which initial distribution ν one chooses, we just write P instead of P_{ν}.

An entry p_{ij} of the transition matrix $\mathbf{P}$ gives the probability for the homogeneous chain (X_n) to get from state i to state j, that is

$$p_{ij} \;=\; \mathrm{P}\big(X_{n+1} = j \mid X_n = i\big) \quad \text{for all } n \in \mathbb{N}_0.$$

If the Markov chain reaches every state in S from any other state with positive probability in a finite number of steps, then the chain is called *irreducible*. We are interested in the long-term behaviour of irreducible Markov chains. As a first step, one studies the recurrence properties of the chain.

Definition 2.7. For $i \in S$ denote by τ_i the random time of return to state i, that is we set

$$\tau_i(\omega) := \; \min\{n \in \mathbb{N} \mid X_0(\omega) = i \;\wedge\; X_n(\omega) = i\}, \quad \omega \in \Omega,$$

where $\tau_i(\omega) := \infty$ if $X_0(\omega) = i$, but $X_n(\omega) \neq i$ for all $n \in \mathbb{N}$. In this notation, the chain (X_n) is *recurrent* iff $\mathrm{P}_i(\tau_i < \infty) = 1$ for all $i \in S$, while it is *positive recurrent* iff $\mathrm{E}_i(\tau_i) < \infty$ for all $i \in S$.

Positive recurrence thus means that for any state the expected time of return to this state is finite, whereas recurrence only guarantees that return is almost certain.

The following sufficient condition for positive recurrence of an irreducible homogeneous Markov chain will be useful, cf. Brémaud (1999: pp. 167-169).

Theorem 2.1 (Foster). *Suppose $(X_n)_{n\in\mathbb{N}_0}$ is irreducible. If there exist $\varepsilon > 0$, a function $h\colon S \to \mathbb{R}$ and a* finite *set $B \subset S$ such that*

(i) $\inf\{h(j) \mid j \in S\} > -\infty$, *that is h is bounded from below,*

(ii) $\sum_{j\in S} p_{ij}h(j) \;<\; \infty$ *for all $i \in B$,*

(iii) $\sum_{j\in S} p_{ij}h(j) \;\leq\; h(i) - \varepsilon$ *for all $i \in S \setminus B$,*

then (X_n) is positive recurrent.

In case the requirements of Foster's theorem are met, the function h is called a *Lyapunov function*, the finite set B is called the *refuge*. Note that for all $i \in S$, $n \in \mathbb{N}_0$ we have

$$\sum_{j\in S} p_{ij}h(j) \;=\; \mathrm{E}\big(h(X_{n+1}) \mid X_n = i\big) \quad \text{on the event } \{\omega \in \Omega \mid X_n(\omega) = i\}.$$

A distribution π is a *stationary distribution* for (X_n) or $\mathbf{P}$ iff $\pi^{\mathsf{T}} = \pi^{\mathsf{T}}\mathbf{P}$. In this case, (X_n) is a stationary process under P_π, that is the distribution of (X_n) is invariant under time shifts.

As far as irreducible Markov chains are concerned, positive recurrence is equivalent to the existence of a stationary distribution (Brémaud, 1999: p. 104).

Theorem 2.2 (Stationary distribution criterion). *Suppose $(X_n)_{n\in\mathbb{N}_0}$ is irreducible. Then (X_n) is positive recurrent if and only if there exists a stationary distribution for (X_n). A stationary distribution π for (X_n), whenever it exists, is unique, and $\pi(i) > 0$ for all $i \in S$.*

The stationary distribution of an irreducible and positive recurrent time homogeneous Markov chain can be represented in terms of the expected times of return (Brémaud, 1999: pp. 104-105).

Theorem 2.3 (Mean return time and stationary distribution). *Suppose $(X_n)_{n\in\mathbb{N}_0}$ is irreducible and positive recurrent. Denote by τ_i the time of return to state i and by π the stationary distribution of (X_n). Then it holds for all $i \in S$ that*

$$\mathrm{E}_i(\tau_i) \;<\; \infty \quad \textit{and} \quad \pi(i) \;=\; \frac{1}{\mathrm{E}_i(\tau_i)}.$$

Once we have established that a chain possesses a stationary distribution, we also know its long-term behaviour. The only additional property we have to check regards the periodicity of the chain. Recall that a Markov chain in S is called *aperiodic* iff

$$d_i := \gcd\{n \in \mathbb{N} \mid p_{ii}(n) > 0\} \;=\; 1 \quad \text{for all } i \in S,$$

where $p_{ii}(n)$ is the probability to return to state i after exactly n steps. The (possibly infinite) number d_i is called the *period* of state i. The states of an irreducible chain all have the same period.

An irreducible, positive recurrent and aperiodic homogeneous Markov chain is called *ergodic*. Such a chain is *mixing* in the sense that it converges to its stationary distribution regardless of its initial state (Brémaud, 1999: p. 130).

Theorem 2.4 (Convergence to stochastic equilibrium). *Suppose the chain $(X_n)_{n\in\mathbb{N}_0}$ is ergodic, and denote by π its stationary distribution. Then the distributions of (X_n) converge in total variation to π for any choice of the initial distribution, that is*

$$\lim_{n\to\infty} \sum_{i\in S} |\,\mathrm{P}_\nu(X_n = i) - \pi(i)| \;=\; 0 \quad \textit{for all distributions } \nu \textit{ on } S.$$

In case the irreducible, positive recurrent chain (X_n) had period $d > 1$, we would look for convergence in total variation with respect to the d-step transition matrix $\mathbf{P}^d$, cf. Brémaud (1999: p. 131).

2.3.2 Convergence to stochastic equilibrium

We now apply the results presented in section 2.3.1 in order to show that the sequence of constraint rankings produced by Jäger's learning algorithm converges – under mild assumptions – to a stationary distribution which depends on the gain parameter η, but is the same for different choices of the initial ranking.

With $\eta > 0$, let R_0^η be a deterministic initial ranking of constraints, i. e. $R_0^\eta = r_s$ for some $r_s \in \mathbb{R}^N$, and let (R_n^η) be the corresponding random sequence of constraint rankings on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ as defined recursively by formula (1.25). The random variables R_n^η, $n \in \mathbb{N}_0$, take values in the countable state space $S_\eta := r_s + \eta\mathbb{Z}^N$. Thus, S_η is the evenly spaced orthogonal grid of mesh size η containing r_s.

Once the state space has been fixed, we can allow for a random initial ranking, that is R_0^η is a random variable with values in S_η. The sequence (R_n^η) satisfies a recursion equation of the form

$$R_{n+1}^\eta(\omega) \quad = \quad R_n^\eta(\omega) \; + \; \eta \cdot L_n\big(\omega, R_n^\eta(\omega)\big) \qquad \text{for all } n \in \mathbb{N}_0,\ \omega \in \Omega, \tag{2.7}$$

where $(L_n)_{n\in\mathbb{N}_0}$ is a sequence of measurable mappings $\Omega \times \mathbb{R}^N \to \mathbb{Z}^N$ such that

(i) $L_n(.,r)$ has distribution μ_r with respect to P for all $r \in \mathbb{R}^N$ and all $n \in \mathbb{N}_0$, where $(\mu_r)_{r\in\mathbb{R}^N}$ is a family of probability distributions on $\mathbb{Z}^N$,

(ii) $(L_n(.,r))_{n\in\mathbb{N}_0, r\in\mathbb{R}^N}$ is an independent family.

Conversely, we may prescribe a family $(\mu_r)_{r\in\mathbb{R}^N}$ of probability distributions on the grid $\mathbb{Z}^N$, choose $\eta > 0$, fix the state space S_η and use (2.7) in order to define a sequence $(R_n^\eta)_{n\in\mathbb{N}_0}$ of S_η-valued random variables.

To this end, take an independent family $(L_n(.,r))_{n\in\mathbb{N}_0, r\in\mathbb{R}^N}$ on a suitable probability space $(\Omega, \mathcal{F}, \mathrm{P})$ such that $L_n(.,r)$ has distribution μ_r under P for all $r \in \mathbb{R}^N$ and all $n \in \mathbb{N}_0$.[8] For each $\eta > 0$ choose an initial distribution ν_η on S_η and, if necessary, augment the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ so as to make it carry a second independent family $(R_0^\eta)_{\eta>0}$, where R_0^η has distribution ν_η under P.

With (R_0^η) and (L_n) as above, define for each $\eta > 0$ a random sequence $(R_n^\eta)_{n\in\mathbb{N}_0}$ according to recursion formula (2.7). We notice that $(R_n^\eta)_{n\in\mathbb{N}_0}$ is a homogeneous Markov chain on $(\Omega, \mathcal{F}, \mathrm{P})$ with state space S_η and transition probabilities

$$\mathrm{P}(r \to r + \eta z) \quad = \quad \mathrm{P}\big(R_{n+1}^\eta = r + \eta z \mid R_n^\eta = r\big) \quad = \quad \mu_r(z), \qquad z \in \mathbb{Z}^N, \tag{2.8}$$

where $r \in S_\eta$ and $n \in \mathbb{N}_0$ is arbitrary. The next proposition states that if the Markov chain of constraint rankings $(R_n^\eta)_{n\in\mathbb{N}_0}$ induced by (μ_r) is irreducible and follows a potential which has a global minimum and grows in an appropriate way, then the chain possesses a stationary distribution provided the gain parameter η is sufficiently small.

Proposition 2.2. *Let $(\mu_r)_{r\in\mathbb{R}^N}$ be a family of probability distributions on $\mathbb{Z}^N$ and define the corresponding random family (L_n) on some suitable probability space $(\Omega, \mathcal{F}, \mathrm{P})$ as above. For $\eta > 0$ and an initial ranking R_0^η with values in S_η define (R_n^η) according to* (2.7)*. Let $f : \mathbb{R}^N \to \mathbb{R}$ be a twice continuously differentiable function. Suppose that*

[8] Kolmogorov's consistency theorem guarantees existence of a family (L_n) with the desired properties, see for example Bauer (1991: pp. 303-310).

(H1) $(R_n^\eta)_{n\in\mathbb{N}_0}$ *is irreducible for all* $\eta > 0$,

(H2) f *has a global minimum at* $\hat{r} \in \mathbb{R}^N$,

(H3) for all $r \in \mathbb{R}^N$

$$L_n(.,r) \text{ is integrable and } \quad \sum_{z\in\mathbb{Z}^N} \mu_r(z)\cdot z \;=\; -\nabla f(r),$$

(H4) there are positive constants $\hat{\eta}$, κ_1, κ_2 *and a* bounded *set* $B \subset \mathbb{R}^N$ *such that for all* $r \in \mathbb{R}^N \setminus B$

$$\|\nabla f(r)\|^2 \;\geq\; \kappa_1 \quad \text{and} \quad \sum_{z\in\mathbb{Z}^N} \mu_r(z)\cdot\|z\|^2 \cdot \sup_{v\in[r,r+\hat{\eta}z]} \|H_f(v)\| \;\leq\; \kappa_2\|\nabla f(r)\|^2,$$

where H_f *is the Hessian of* f.

Set $\eta_0 := \min\{\hat{\eta}, \frac{2}{\kappa_2}\}$. *If* $\eta \in (0,\eta_0)$, *then the Markov chain* $(R_n^\eta)_{n\in\mathbb{N}_0}$ *possesses a stationary distribution.*

Proof. First observe that by construction of (L_n) we have for all $r \in \mathbb{R}^N$ and all $n \in \mathbb{N}_0$

$$\sum_{z\in\mathbb{Z}^N} \mu_r(z)\cdot z \quad = \quad \mathrm{E}\big(L_n(.,r)\big), \qquad \sum_{z\in\mathbb{Z}^N} \mu_r(z)\cdot\|z\|^2 \quad = \quad \mathrm{E}\big(\|L_n(.,r)\|^2\big).$$

Let $\eta > 0$. By construction and hypothesis, $(R_n^\eta)_{n\in\mathbb{N}_0}$ is an irreducible homogeneous Markov chain with countable state space S_η and transition probabilities given by (2.8).

According to theorem 2.2, the stationary distribution criterion, we have to show that (R_n^η) is positive recurrent. We will apply Foster's drift condition, which is theorem 2.1, here. More precisely, we show that there exist a positive constant ε dependent on η and a function $h\colon \mathbb{R}^N \to \mathbb{R}$ bounded from below such that

$$\mathrm{E}\Big(h\big(r+\eta L_n(.,r)\big)\Big) \;\leq\; h(r) - \varepsilon \quad \text{for all } r \in \mathbb{R}^N \setminus B,\; n \in \mathbb{N}_0. \tag{2.9}$$

The above expectation is really independent of $n \in \mathbb{N}_0$ as $(L_n(.,r))_{n\in\mathbb{N}_0}$ is an i. i. d. sequence. For Lyapunov function h we simply take the function f itself, which is bounded from below by its minimum value $f(\hat{r})$.

Let $\hat{\eta}$, κ_1, κ_2 be positive constants and $B \subset \mathbb{R}^N$ a bounded set such that the inequalities in hypothesis (H4) are satisfied. Assume that $\eta \in (0,\hat{\eta})$. Let $r \in \mathbb{R}^N \setminus B$. Appealing to

Taylor's formula, we find an $\mathcal{F}$-measurable function $\phi_r\colon \Omega \to [0,1]$ such that

$$\begin{aligned}
& \mathrm{E}\Big(f\big(r+\eta L_n(.,r)\big)\Big) \\
= \;& \mathrm{E}\Big(f(r) \,+\, \eta\,\big\langle \nabla f(r), L_n(.,r)\big\rangle \,+\, \tfrac{\eta^2}{2}\,\big\langle L_n(.,r), H_f\big(r+\eta\phi_r(.)L_n(.,r)\big)L_n(.,r)\big\rangle\Big) \\
& \;\big|\; \text{Taylor's formula} \\
= \;& f(r) \,-\, \eta\,\big\langle \nabla f(r), \nabla f(r)\big)\big\rangle \,+\, \tfrac{\eta^2}{2}\,\mathrm{E}\Big(\big\langle L_n(.,r), H_f\big(r+\eta\phi_r(.)L_n(.,r)\big)L_n(.,r)\big\rangle\Big) \\
& \;\big|\; \text{by hypothesis (H3)} \\
\leq \;& f(r) \,-\, \eta\,\big\|\nabla f(r)\big\|^2 \,+\, \tfrac{\eta^2}{2}\,\mathrm{E}\Big(\big\|H_f\big(r+\eta\phi_r(.)L_n(.,r)\big)\big\|\big\|L_n(.,r)\big\|^2\Big) \\
& \;\big|\; \text{Cauchy-Schwarz inequality} \\
\leq \;& f(r) \,-\, \big\|\nabla f(r)\big\|^2\cdot\big(\eta-\tfrac{\eta^2}{2}\kappa_2\big) \qquad \big|\; \text{by hypothesis (H4)} \\
\leq \;& f(r) \,-\, \kappa_1\big(\eta-\tfrac{\eta^2}{2}\kappa_2\big) \quad =:\; f(r) \,-\, \varepsilon.
\end{aligned}$$

If $\eta < \frac{2}{\kappa_2}$, then $\varepsilon = \varepsilon(\eta) > 0$. Before applying theorem 2.1 we observe that $B \cap S_\eta$ is a finite set and that for all $r \in S_\eta$, $n \in \mathbb{N}_0$ we have

$$\mathrm{E}\big(h(R^\eta_{n+1}) \mid R^\eta_n = r\big) \;=\; \mathrm{E}\big(h\big(r+\eta L_n(.,r)\big)\big) \quad \text{on the event } \{\omega\in\Omega \mid R^\eta_n(\omega) = r\}. \qquad \square$$

If the hypotheses of proposition 2.2 are satisfied and, in addition, $(R^\eta_n)_{n\in\mathbb{N}_0}$ is aperiodic, then for $\eta > 0$ small enough $(R^\eta_n)_{n\in\mathbb{N}_0}$ converges in total variation to its stationary distribution as a consequence of theorem 2.4.

It is clear that the dual function f_{pot} will take over the rôle of the function f in proposition 2.2. By proposition 1.1, we know that f_{pot} is convex and twice differentiable with partial derivatives given by (1.24). In the notation of section 2.2, if $\mathrm{E}_{emp}(c)$ is regular and $V_c = \{0\}$, then f_{pot} has a unique global minimum. In case $V_c \neq \{0\}$, we consider the projection of the constraint rankings onto $V_c^\perp$.

In order to be able to apply proposition 2.2 to the problem of finding the maximum entropy constraint ranking, we have to specify the family $(\mu_r)_{r\in\mathbb{R}^N}$ of probability distributions on $\mathbb{Z}^N$. In the notation of section 1.4, for $r \in \mathbb{R}^N$ set

(2.10)
$$\mu_r(z) := \;\mathrm{P}\Big(\big\{\omega\in\Omega \mid c\big(X^{in}_n(\omega), H_n(X^{in}_n(\omega), r)\big) - c\big(X^{in}_n(\omega), X^{out}_n(\omega)\big) = z\big\}\Big), \quad z\in\mathbb{Z}^N,$$

the definition being independent of $n \in \mathbb{N}$. Taking into account the distributions of (X^{in}_n, X^{out}_n) and $H_n(X^{in}_n, r)$, respectively, and their mutual independence conditional on X^{in}_n, we find that

$$\mu_r(z) \;=\; \sum_{i\in I} \tilde{p}_{emp}(i) \sum_{o_1,o_2\in O_i} p_{emp}(o_1|i)\cdot p_r(o_2|i)\cdot \mathbf{1}_{\{z\}}\big(c(i,o_2)-c(i,o_1)\big), \quad z\in\mathbb{Z}^N. \tag{2.11}$$

Let us assume for the moment that the second partial derivatives of f_{pot} are bounded or, equivalently, that the Hessian of f_{pot} has globally bounded matrix norm. Then hypothesis (H4) of proposition 2.2 imposes a bound on $\sum \mu_r(z)\|z\|^2$, and by (2.11) we have

$$\sum_{z\in\mathbb{Z}^N} \mu_r(z)\cdot\|z\|^2 \;=\; \sum_{i\in I} \tilde{p}_{emp}(i) \sum_{o_1,o_2\in O_i} p_{emp}(o_1|i)\cdot p_r(o_2|i)\cdot \big\|c(i,o_2)-c(i,o_1)\big\|^2. \tag{2.12}$$

Notice that hypothesis (H3) is satisfied by the choice of f_{pot} in place of f. The first part of hypothesis (H4) is fulfilled by virtue of f_{pot} being convex and growing to infinity – provided $\mathrm{E}_{emp}(c)$ is regular and $V_c = \{0\}$, see lemma A.2.

We conclude this section by a summary of conditions that guarantee convergence of Jäger's algorithm to stochastic equilibrium. The summability conditions could be relaxed, as becomes clear by comparison of propositions 2.2 and 2.3. But observe that the case of greatest applicability is probably the one where the feature function c is bounded, which certainly holds true when the generator G is finite.

Proposition 2.3. *Let $c\colon G \to \mathbb{N}_0^N$ be a feature function and p_{emp} an empirical distribution with support in $G \subseteq I \times O$ as in section 1.1. Let V_c be the subspace of $\mathbb{R}^N$ determined by* (2.2a). *Assume that O_i is finite for all $i \in I$ and that $\mathcal{R} = \mathbb{R}^N$. For each $r \in \mathbb{R}^N$ let μ_r be the probability distribution on $\mathbb{Z}^N$ satisfying* (2.11). *For $\eta > 0$ let S_η be an η-spaced orthogonal grid in $\mathbb{R}^N$ and let R_0^η be an initial ranking with values in S_η. Define the sequence $(R_n^\eta)_{n\in\mathbb{N}_0}$ of constraint rankings according to* (1.25). *Suppose that $V_c = \{0\}$ and that*

(H1) $\mathrm{E}_{emp}(c)$ is regular in the sense of theorem 1.1,

(H2) $(\mu_r)_{r\in\mathbb{R}^N}$ is such that $(R_n^\eta)_{n\in\mathbb{N}_0}$ is irreducible and aperiodic for all $\eta > 0$,

(H3) there is a positive constant κ such that for all $r \in \mathbb{R}^N$, $j,k \in \{1,\dots,N\}$

$$\mathrm{E}_{\tilde{p}_{emp}} \big|\mathrm{cov}_{p_r(.|i)}\big(c_j(i,.), c_k(i,.)\big)\big| \;\le\; \kappa,$$

(H4) there is a positive constant κ_2 and a bounded *set $B \subset \mathbb{R}^N$ such that for all $r \in \mathbb{R}^N \setminus B$*

$$\sum_{z\in\mathbb{Z}^N} \mu_r(z)\cdot\|z\|^2 \;\le\; \frac{\kappa_2}{\kappa}\cdot\|\mathrm{E}_{emp}(c) - \mathrm{E}_{p_r}(c)\|^2,$$

where κ is the constant from hypothesis (H3).

If $\eta \in (0, \frac{2}{\kappa_2})$, then (R_n^η) possesses a stationary distribution π_η and R_n^η converges in total variation to π_η as n tends to infinity.

In case $V_c \neq \{0\}$ analogous assertions obtain for the projection of (R_n^η) to any of the affine-linear subspaces $r + V_c^\perp$ with $r \in \mathbb{R}^N$. The induced Gibbs distributions $p_{R_n^\eta}$, $n \in \mathbb{N}_0$, are convergent in any case.

If c is bounded, then the covariances in hypothesis (H3) of proposition 2.3 become arbitrarily small for all ranking parameters of sufficiently large norm, cf. lemma A.2. It follows that a stationary distribution exists for all choices of the gain parameter $\eta > 0$, and convergence to stochastic equilibrium is guaranteed.

2.4 The limit of small constant step size

We have seen that the sequence of constraint rankings converges to the stationary regime, when the gain parameter η is held constant. What we have not obtained, yet, are statements concerning the form of the stationary distribution. Recalling that Jäger's algorithm is essentially a procedure for finding the minimum positions of the dual function f_{pot}, we should expect that the stationary distributions are concentrated about the positions of those minima provided η is small enough and the expectation of the feature function under the empirical distribution is regular.

In section 2.3, the discrete nature of the feature function c allowed us to consider any sequence (R_n^η) of constraint rankings as a Markov chain with discrete state space S_η. It was thus possible to get along with a modest amount of theory. There are several drawbacks to this approach, though.

Firstly, we were forced to assume that the initial ranking R_0^η is a random variable taking values in an N-dimensional grid, rather than in the entire space $\mathbb{R}^N$. Secondly, the state space depends on the gain parameter η, so each ranking sequence (R_n^η) has its own state space, which makes it difficult to compare ranking sequences for different values of the gain parameter. Thirdly, the approach would not do if the single constraints c_j were allowed to take values in an interval $I \subseteq [0,\infty)$, instead of only non-negative integer values.

Therefore, let us drop all limitations on the state space and work directly with $\mathbb{R}^N$-valued constraint rankings. The results of this section are easily adapted to the case when c is a function $G \to [0,\infty)^N$, though we will stick to the hypotheses of section 1.1. Before we further analyze Jäger's algorithm in section 2.4.2, we recall a general convergence result from the literature.

2.4.1 Mean ODE approximation

Let R_0^η, $\eta > 0$, be a family of $\mathbb{R}^N$-valued random variables on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. For each η, define (R_n^η) according to (1.25). In order to be able to compare the distributions of ranking sequences for different values of η, a common time scale is needed. The natural choice is the time scale induced by the gain parameter itself. Accordingly, we set

$$R^\eta(t) := \sum_{n=0}^{\infty} R_n^\eta \cdot \mathbf{1}_{[n\eta,(n+1)\eta)}(t) \;=\; R^\eta_{\lfloor \frac{t}{\eta} \rfloor}, \quad t \geq 0, \tag{2.13}$$

where $\lfloor . \rfloor$ denotes the Gauß bracket, that is $\lfloor x \rfloor$ is the biggest integer not greater than x. Thus, (2.13) defines a stochastic process on $(\Omega, \mathcal{F}, \mathrm{P})$ with piecewise continuous paths. As path space one usually takes $D_\infty^N := \mathrm{D}_{\mathbb{R}^N}([0,\infty))$, the space of all $\mathbb{R}^N$-valued *càdlàg*-functions on $[0,\infty)$ endowed with the Skorokhod topology.[9]

[9] A càdlàg-function is a right-continuous function with finite left-hand limits. For definition and properties of the Skorokhod space D_∞^N see Ethier and Kurtz (1986: §3.5).

We already know that the sequences of constraint rankings follow on average a potential given by the function f_{pot}. The same holds true for the time-interpolated processes $R^\eta(.)$. As the gain parameter η tends to zero, the impact of the potential should become predominant.

In fact, we will see that if the initial rankings R_0^η converge in distribution to some $\mathbb{R}^N$-valued random variable R_0 as η goes to zero, then the processes $R^\eta(.)$ will converge in distribution to a process $R(.)$, where $R(.,\omega)$ is determined as the unique solution to

$$\dot{x}(t) \;=\; -\nabla f_{pot}\big(x(t)\big), \quad t \geq 0, \qquad x(0) \;=\; R_0(\omega). \tag{2.14}$$

The ordinary differential equation in (2.14) is referred to as the *mean ODE* of the underlying algorithm.

To prove convergence, we could invoke theorems 8.2.1 and 8.5.1, respectively, from Kushner and Yin (2003: §§ 8.1, 8.5).[10] Recall recursion formula (2.7), which we may rewrite as

$$R_{n+1}^\eta \quad = \quad R_n^\eta \;-\; \eta \cdot \nabla f_{pot}\big(R_n^\eta\big) \;+\; \partial M_{n+1}, \qquad n \in \mathbb{N}_0, \tag{2.15}$$

where ∂M_{n+1} is a *martingale difference* noise term defined as

$$\partial M_{n+1} := \; L_n(.,R_n^\eta) \;-\; \mathrm{E}\big(L_n(.,R_n^\eta) \;\big|\; R_l^\eta, L_{l-1}(.,r),\, l \in \{0,\ldots,n\}, r \in \mathbb{R}^N\big).$$

Instead of further pursuing this approach, we will rely on the following extremely simplified version of a general result concerning the approximation of Itô diffusions by discrete Markov chains; for details see Ethier and Kurtz (1986: §7.4) or Stroock and Varadhan (1979: §11.2).

Theorem 2.5 (Mean ODE approximation). *Let $b\colon \mathbb{R}^N \to \mathbb{R}^N$ be a locally Lipschitz continuous function. For each $\eta > 0$ let (Y_n^η) be a homogeneous Markov chain with state space $\mathbb{R}^N$ and transition function $\mu^\eta\colon \mathbb{R}^N \times \mathcal{B}(\mathbb{R}^N) \to [0,1]$. Define the truncated coefficient functions a^η, b^η and the time-interpolated processes $X^\eta(.)$ by*

$$\begin{aligned}
a^\eta(x) &:= \; \frac{1}{\eta} \int_{\|y-x\|\leq 1} (y-x)(y-x)^\mathsf{T}\, \mu^\eta(x,dy), && x \in \mathbb{R}^N,\\
b^\eta(x) &:= \; \frac{1}{\eta} \int_{\|y-x\|\leq 1} (y-x)\, \mu^\eta(x,dy), && x \in \mathbb{R}^N,\\
X^\eta(t) &:= \; Y^\eta_{\lfloor \frac{t}{\eta} \rfloor}, && t \geq 0.
\end{aligned}$$

Suppose that for each $\rho > 0$ and each $\varepsilon > 0$

$$\sup_{\|x\|\leq\rho} \big\|a^\eta(x)\big\| \;\overset{\eta\to 0}{\to}\; 0, \quad \sup_{\|x\|\leq\rho} \big\|b^\eta(x) - b(x)\big\| \;\overset{\eta\to 0}{\to}\; 0, \quad \frac{1}{\eta} \cdot \sup_{\|x\|\leq\rho} \mu^\eta\big(x, \mathbb{R}^N \setminus \mathrm{B}_\varepsilon(x)\big) \;\overset{\eta\to 0}{\to}\; 0.$$

If the family (Y_0^η) converges in distribution to some $\mathbb{R}^N$-valued random variable Y_0 as η tends to zero, then the processes $X^\eta(.)$ converge in distribution to the process $X(.)$, where $X(.,\omega)$ is determined as the unique solution to the deterministic initial value problem

$$\dot{x}(t) \;=\; b\big(x(t)\big), \quad t \geq 0, \qquad x(0) \;=\; Y_0(\omega).$$

[10] The most important hypothesis to check would be the tightness of the family $\{R^\eta(.) \mid \eta > 0\}$.

In a more general version of theorem 2.5 the functions a^η could converge to a continuous matrix-valued function a. The limit process $X(.)$ would then be given as solution to a stochastic differential equation (SDE) determined by the coefficient functions a and b. The function b, called *drift vector*, would still represent the deterministic part of the dynamics, while a would be the *diffusion matrix*, corresponding to the noise part of the SDE.[11]

The original result, which is theorem 11.2.3 in Stroock and Varadhan (1979), makes use of the so-called martingale problem.[12] Well-posedness of the martingale problem in the situation of theorem 2.5 reduces to well-posedness of the deterministic initial value problem for the function b.

2.4.2 Convergence in distribution and sojourn probabilities

In order to apply theorem 2.5, we have to compute the transition function associated with (R_n^η) for each $\eta > 0$. For $r \in \mathbb{R}^N$ let μ_r be as given by (2.10) and (2.11). In case the range of c were continuous we would define $\mu_r : \mathcal{B}(\mathbb{R}^N) \to [0,1]$, $r \in \mathbb{R}^N$, in analogy to (2.10) and find that

(2.16)
$$\mu_r(\Gamma) \;=\; \sum_{i\in I} \tilde{p}_{emp}(i) \sum_{o_1,o_2\in O_i} p_{emp}(o_1|i)\cdot p_r(o_2|i)\cdot \mathbf{1}_{\{\Gamma\}}\big(c(i,o_2)-c(i,o_1)\big), \quad \Gamma\in\mathcal{B}(\mathbb{R}^N).$$

The transition function associated with the Markov chain (R_n^η) thus reads

(2.17)
$$\mu^\eta(r,\Gamma) \;=\; \begin{cases} \sum\limits_{z\in\mathbb{Z}^N} \mu_r(z)\cdot \mathbf{1}_\Gamma(r+\eta\cdot z), & \text{if range}(c)\subseteq\mathbb{Z}^N, \\ \int\limits_{z\in\mathbb{R}^N} \mathbf{1}_\Gamma(r+\eta\cdot z)\,\mu_r(dz), & \text{in general,} \end{cases} \qquad r\in\mathbb{R}^N,\ \Gamma\in\mathcal{B}(\mathbb{R}^N).$$

Restricting attention to integer-valued constraints, we compute the associated truncated coefficient functions a^η, b^η. For $x\in\mathbb{R}^N$ it holds that

(2.18)
$$\begin{aligned} a^\eta(x) &= \frac{1}{\eta}\int_{\|y-x\|\le 1} (y-x)(y-x)^{\mathsf{T}}\,\mu^\eta(x,dy) \\ &= \frac{1}{\eta}\sum_{z\in\mathbb{Z}^N:\,\|z\|\le\frac{1}{\eta}} \mu_x(z)\cdot(x+\eta\cdot z-x)(x+\eta\cdot z-x)^{\mathsf{T}} \\ &= \eta\cdot\sum_{z\in\mathbb{Z}^N:\,\|z\|\le\frac{1}{\eta}} \mu_x(z)\cdot zz^{\mathsf{T}}. \end{aligned}$$

[11] The functions b and a should be compared with two fundamental characteristics of a random variable, namely the expectation or mean, on the one hand, and the variance, usually denoted by σ^2, on the other.

[12] The martingale problem is the problem of finding a probability measure on the path space which makes a family of processes induced by an abstract operator into martingales, where a martingale is a process preserving conditional expectations. For an SDE a differential operator can be associated with its coefficient functions, and the martingale problem establishes a link between that differential operator, Itô diffusions and the infinitesimal generator of a Markov process.

Notice that zz^T is a symmetric $N \times N$-matrix for each $z \in \mathbb{Z}^N$. As to b^η, we have for $x \in \mathbb{R}^N$

$$\begin{aligned} b^\eta(x) &= \frac{1}{\eta} \int_{\|y-x\| \le 1} (y-x)\, \mu^\eta(x, dy) \\ &= \frac{1}{\eta} \sum_{z \in \mathbb{Z}^N:\, \|z\| \le \frac{1}{\eta}} \mu_x(z) \cdot (x + \eta \cdot z - x) \quad = \quad \sum_{z \in \mathbb{Z}^N:\, \|z\| \le \frac{1}{\eta}} \mu_x(z) \cdot z. \end{aligned} \tag{2.19}$$

Recalling the discussion of section 2.3.2 or, more directly, by considering (2.11) together with equation (1.26), we see that

$$\lim_{\eta \to 0} b^\eta(x) \;=\; \sum_{z \in \mathbb{Z}^N} \mu_x(z) \cdot z \quad = \quad -\nabla f_{pot}(x)$$

for all those $x \in \mathbb{R}^N$ such that the sum exists. This is clearly true for all $x \in \mathbb{R}^N$ whenever the feature function c is bounded. In this case uniform convergence of $b^\eta(.)$ is easily checked and one also finds that $a^\eta(x) \to 0$ uniformly in $x \in \mathbb{R}^N$ as η tends to zero.

If we do not have boundedness of c, then we just assume summability conditions on $\mu_r(.)$ sufficient so as to satisfy the hypotheses of theorem 2.5. The following proposition summarizes our findings.

Proposition 2.4. *Let $c\colon G \to \mathbb{N}_0^N$ be a feature function and p_{emp} an empirical distribution with support in $G \subseteq I \times O$ as in section 1.1. Let V_c be the subspace of $\mathbb{R}^N$ determined by* (2.2a)*. Assume that O_i is finite for all $i \in I$ and that $\mathcal{R} = \mathbb{R}^N$. For each $r \in \mathbb{R}^N$ let μ_r be the probability distribution on $\mathbb{Z}^N$ satisfying* (2.11)*. For each $\eta > 0$ let R_0^η be an arbitrary initial ranking, i.e. an $\mathbb{R}^N$-valued random variable on the common probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Define the sequence $(R_n^\eta)_{n \in \mathbb{N}_0}$ of constraint rankings according to* (1.25)*, and let the corresponding time-interpolated process be given by* (2.13)*. Suppose that*

(H1) for each $\rho > 0$

$$\eta \cdot \sup_{\|r\| \le \rho} \sum_{z \in \mathbb{Z}^N:\, \|z\| \le \frac{1}{\eta}} \mu_r(z) \cdot \|z\|^2 \;\overset{\eta \to 0}{\longrightarrow}\; 0,$$

(H2) for each $\rho > 0$

$$\sup_{\|r\| \le \rho} \sum_{z \in \mathbb{Z}^N:\, \|z\| \ge \frac{1}{\eta}} \mu_r(z) \cdot \|z\| \;\overset{\eta \to 0}{\longrightarrow}\; 0.$$

If the family (R_0^η) converges in distribution to some $\mathbb{R}^N$-valued random variable R_0 as η tends to zero, then the processes $R^\eta(.)$ converge in distribution to the process $R(.)$, where $R(., \omega)$ is determined as the unique solution to the deterministic initial value problem (2.14)*.*

Assume, in addition, that $\mathrm{E}_{emp}(c)$ is regular in the sense of theorem 1.1 and that $V_c = \{0\}$. Let $\hat{r} \in \mathbb{R}^N$ be the position of the global minimum of f_{pot}. Then for $\varepsilon > 0$ and $\delta > 0$ one finds $\eta_0 > 0$, $t_0 > 0$ and a family $(T^\eta)_{\eta \in (0, \eta_0]} \subset (0, \infty)$ such that

$$T^\eta \to \infty \quad \textit{as} \quad \eta \searrow 0,$$

$$\mathrm{P}\Big(\sup\big\{\|R_n^\eta - \hat{r}\| \;\big|\; \tfrac{t_0}{\eta} \le n \le \tfrac{t_0 + T^\eta}{\eta}\big\} \le \varepsilon\Big) \quad > \quad 1 - \delta \qquad \textit{for all } \eta \in (0, \eta_0].$$

If $V_c \neq \{0\}$, then an analogous assertion holds for the sequence of constraint rankings projected to $V_c^{\perp}$.

Proof. The convergence assertion is, of course, a consequence of theorem 2.5. Notice that $\|zz^{\mathsf{T}}\| = \|z\|^2$ if we choose the matrix norm induced by the Euclidean distance. The requirement regarding $\mu^\eta(x, \mathbb{R}^N \setminus \mathrm{B}_\varepsilon(x))$ is implied by hypothesis (H2), because

$$\mu^\eta\big(x, \mathbb{R}^N \setminus \mathrm{B}_\varepsilon(x)\big) \;=\; \sum_{z \in \mathbb{Z}^N:\, \|z\| \geq \frac{\varepsilon}{\eta}} \mu_r(z) \quad \text{for all } \varepsilon > 0,\ \eta > 0,\ r \in \mathbb{R}^N.$$

If $\mathrm{E}_{emp}(c)$ is regular and $V_c = \{0\}$, then f_{pot} possesses a unique global minimum. By (2.14), we have $R(t,\omega) \to \hat{r}$ as $t \to \infty$ for all $\omega \in \Omega$. Given $\varepsilon > 0$, $\delta > 0$, we can choose a compact set $B \subset \mathbb{R}^N$ such that $\mathrm{P}(R_0 \in B) > 1 - \frac{\delta}{2}$. Then there is $t_0 > 0$ such that $\mathrm{P}(\|R(t) - \hat{r}\| \leq \frac{\varepsilon}{2}) > 1 - \frac{\delta}{2}$ for all $t \geq t_0$.

For $t \geq 0$ let π_t be the canonical coordinate projection $D_\infty^N \to \mathbb{R}^N$, that is we set

$$\pi_t(\phi) := \; \phi(t), \quad \phi \in D_\infty^N.$$

Although π_t is not continuous on D_∞^N (w. r. t. the Skorokhod topology), π_t is P_R-almost surely continuous, because $R(.)$ has continuous sample paths. For $T > 0$ define

$$\xi_T \colon \mathrm{D}_\infty^N \to [0, \infty] \quad \text{by} \quad \xi_T(\phi) := \sup\big\{\|\phi(t) - \hat{r}\| \;\big|\; t \in [t_0, t_0 + T]\big\}.$$

Notice that ξ_T is continuous and finite on $\mathrm{C}([0,\infty), \mathbb{R}^N)$, while it is only P_R-almost surely continuous and finite on D_∞^N. Let g_ε be a continuous real-valued function on $\mathbb{R}^N$ such that $0 \leq g_\varepsilon \leq 1$, $g_\varepsilon(x) = 1$ whenever $\|x\| \leq \frac{\varepsilon}{2}$, $g_\varepsilon(x) = 0$ whenever $\|x\| > \varepsilon$. For $T > 0$, $n \in [\frac{t_0}{\eta}, \frac{t_0+T}{\eta}] \cap \mathbb{N}$, $\eta > 0$ we have

$$\begin{aligned}
& \mathrm{P}\Big(\sup\big\{\|R_n^\eta - \hat{r}\| \;\big|\; \tfrac{t_0}{\eta} \leq n \leq \tfrac{t_0+T}{\eta}\big\} \leq \varepsilon\Big) \\
\geq \;& \int_\Omega \mathbf{1}_{\{\|R^\eta(t,\omega) - \hat{r}\| \leq \varepsilon \,|\, t \in [t_0, t_0+T]\}} \, d\,\mathrm{P}(\omega) \\
= \;& \int_{\mathrm{D}_\infty^N} \mathbf{1}_{\{\|\pi_t(\phi) - \hat{r}\| \leq \varepsilon \,|\, t \in [t_0, t_0+T]\}} \, d\,\mathrm{P}_{R^\eta}(\phi) \\
\geq \;& \int_{\mathrm{D}_\infty^N} g_\varepsilon\big(\xi_T(\phi)\big) \, d\,\mathrm{P}_{R^\eta}(\phi) \qquad\qquad \xrightarrow{\eta \downarrow 0} \quad \int_{\mathrm{D}_\infty^N} g_\varepsilon\big(\xi_T(\phi)\big) \, d\,\mathrm{P}_R(\phi) \\
\geq \;& \mathrm{P}\Big(\sup\big\{\|R(t) - \hat{r}\| \;\big|\; t \in [t_0, t_0+T]\big\}\big\} \leq \frac{\varepsilon}{2}\Big) \qquad > \quad 1 - \frac{\delta}{2}.
\end{aligned}$$

Therefore, given $T > 0$, we can choose $\eta(T) > 0$ such that

$$\mathrm{P}\Big(\sup\big\{\|R_n^\eta - \hat{r}\| \;\big|\; \tfrac{t_0}{\eta} \leq n \leq \tfrac{t_0+T}{\eta}\big\} \leq \varepsilon\Big) \quad > \quad 1 - \delta \qquad \text{for all } \eta \in (0, \eta(T)].$$

Clearly, $\eta(T)$ can be chosen decreasing in $T > 0$. Now, let $(T_n)_{n\in\mathbb{N}_0} \subset (0,\infty)$ be any increasing sequence such that $T_n \to \infty$ as n tends to infinity. Set $\eta_n := \eta(T_n)$, $n \in \mathbb{N}_0$. In particular, we have found $\eta_0 > 0$. We obtain the required sequence $(T^\eta)_{\eta\in(0,\eta_0]}$ by setting

$$T^\eta := T_n \quad \text{iff} \quad \eta \in (\eta_{n+1}, \eta_n], \qquad\qquad \eta \in (0, \eta_0].$$

By construction, we have $T^\eta \to \infty$ as $\eta \searrow 0$ as well as the asserted probability estimate. □

Observe that proposition 2.4 does *not* guarantee existence of $n_0 \in \mathbb{N}$ such that

$$\mathrm{P}\big(\|R_n^\eta - \hat{r}\| \le \varepsilon\big) \;>\; 1-\delta \quad \text{for all } n \ge n_0.$$

In fact, it is not true in general that R_n^η remains in an ε-neighbourhood of $\hat{r}$ for *all* n big enough (η arbitrarily small). Waiting long enough one will usually witness excursions away from $\hat{r}$ with probability greater than δ. There is, so to speak, a residual risk that the algorithm will some time or other depart from a small neighbourhood of the minimum position of f_{pot}.

2.5 Variable step size tending to zero

We now turn to the case that the gain parameter η in Jäger's algorithm is variable. In the notation of section 2.3.2, the sequence of constraint rankings then satisfies – instead of (2.7) – the recursion equation

$$R_{n+1}(\omega) \;=\; R_n(\omega) \;+\; \eta_n \cdot L_n\big(\omega, R_n(\omega)\big) \quad \text{for all } n \in \mathbb{N}_0,\ \omega \in \Omega, \tag{2.20}$$

where $(L_n)_{n\in\mathbb{N}_0}$ is again a sequence of measurable mappings from $\Omega \times \mathbb{R}^N$ to $\mathbb{Z}^N$ – or $\mathbb{R}^N$ if c has continuous range – such that

(i) $L_n(.,r)$ has distribution μ_r with respect to P for all $r \in \mathbb{R}^N$ and all $n \in \mathbb{N}_0$, where μ_r for $r \in \mathbb{R}^N$ is given by (2.11), respectively (2.16) for non-discrete range of c,

(ii) $(L_n(.,r))_{n\in\mathbb{N}_0, r\in\mathbb{R}^N}$ is an independent family.

Note that we write (R_n) instead of (R_n^η), because here the $n+1$-th iterate of the constraint sequence depends not on a fixed value η, but on the values of the parameter sequence up to η_n. We refrain from making this dependency explicit in the formulae and simply take the sequence (η_n) as specified in advance. Similarly, we must be given an initial constraint ranking R_0, i. e. an $\mathbb{R}^N$-valued random variable.

If the parameter sequence (η_n) goes slowly enough to zero and if the dual function f_{pot} has a unique global minimum, then we may hope for convergence of (R_n) to the position of that minimum.

The ideas presented in section 2.3.1 can be extended considerably. The next result, which is theorem 4.5.3 in Kushner and Yin (2003: §4.5), should be compared with Foster's drift

condition, cited here as theorem 2.1, and our proposition 2.2, which establishes existence of a stationary distribution. Since the gain parameter is no longer constant, we are aiming at almost sure convergence to a point, rather than weak convergence to the stationary regime.

Theorem 2.6 (Stochastic Stability). *Let $(X_n)_{n\in\mathbb{N}_0}$ be a sequence of $\mathbb{R}^N$-valued random variables on a probability space $(\Omega, \mathrm{P}, \mathcal{F})$. Denote by $(\mathcal{F}_n)$ the filtration generated by (X_n). Let h, k be non-negative real-valued functions on $\mathbb{R}^N$ such that*

(H1) h is continuous and $h(0) = 0$,

(H2) for each $\varepsilon > 0$ there is $\delta > 0$ such that $k(x) \geq \delta$ whenever $\|x\| \geq \varepsilon$,

(H3) for each $\varepsilon > 0$ there is $\delta(\varepsilon) > 0$ such that $h(x) \geq \delta(\varepsilon)$ whenever $\|x\| \geq \varepsilon$, and $\delta(\varepsilon)$ can be chosen non-decreasing in ε,

(H4) $\mathrm{E}(h(X_0)) < \infty$ and integrability of $h(X_n)$ implies integrability of $k(X_n)$ for all $n \in \mathbb{N}_0$.

Suppose there are positive constants α, κ, κ_2, an $(\mathcal{F}_n)$-adapted sequence (ε_n) of positive random variables and a sequence (Y_n) of $\mathbb{R}^N$-valued random variables such that

(H5) P*-a. s.,*

$$\varepsilon_n \overset{n\to\infty}{\longrightarrow} 0 \quad \textit{and} \quad \sum\nolimits_{n\geq 0} \varepsilon_n \;=\; \infty,$$

(H6)

$$\mathrm{E}\Big(\sum\nolimits_{n\geq 0} \varepsilon_n^2 \cdot \|Y_n\|^2 \cdot \mathbf{1}_{\|X_n\|\leq\alpha}\Big) \;<\; \infty,$$

(H7) for all $n \in \mathbb{N}_0$, on the event $\{\|X_n\| \geq \alpha\}$,

$$\mathrm{E}\big(\|Y_n\|^2 \mid \mathcal{F}_n\big) \;\leq\; \kappa_2 \cdot k(X_n),$$

(H8) for all $n \in \mathbb{N}_0$, P*-a. s.,*

$$\mathrm{E}\big(h(X_{n+1}) \mid \mathcal{F}_n\big) \;-\; h(X_n) \quad\leq\quad -\varepsilon_n \cdot k(X_n) \;+\; \kappa \cdot \varepsilon_n^2 \cdot \mathrm{E}\big(\|Y_n\|^2 \mid \mathcal{F}_n\big).$$

Then $X_n \to 0$ as $n \to \infty$ with probability one.

Suppose the dual function f_{pot} has a unique global minimum at $\hat{r} \in \mathbb{R}^N$. Assume further that the – possibly random – sequence of gain parameters (η_n) satisfies hypothesis (H5) above. Under conditions analogous to those of proposition 2.3 we can apply theorem 2.6 to the sequence $X_n = R_n - \hat{r}$ by choosing $h(x) := f_{pot}(x+\hat{r}) - f_{pot}(\hat{r})$, $k(x) := \|\nabla f_{pot}(x+\hat{r}\|^2$ and letting $Y_n := L_n(., R_n)$, thereby obtaining P-almost sure convergence of $R_n \to \hat{r}$ as n tends to infinity.

The most important hypothesis of theorem 2.6 to check is hypothesis (H8). This can be done using Taylor's formula in the same way as in the proof of proposition 2.2.

Chapter 3

Interpretation of the convergence results

Section 3.1 serves to summarize the convergence properties of Jäger's algorithm. In section 3.2 we discuss the question of whether there is any need for allowing the generator of a stochastic OT grammar to be infinite. As above, a distinction is drawn between generators, where the set of inputs is infinite, but the set of output candidates is finite for each input, and generators allowing for an infinite number of inputs, each of them with a possibly infinite set of output candidates. Results by Jäger and Rosenbach (2005) are presented as an application of the maximum entropy version of stochastic OT to a syntactic phenomenon in section 3.3. Indicating some open questions and possible extensions, we conclude our analysis with section 3.4.

3.1 A robust and flexible algorithm

Jäger's algorithm converges under mild integrability conditions on the feature function and mild regularity assumptions on the empirical distribution as was shown in chapter 2. The regularity assumption concerning the empirical distribution p_{emp} translates into requiring that p_{emp} provide positive and negative evidence for all constraints or grammatical dimensions. A standing hypothesis about the underlying generator was that there were only finitely many output candidates associated with any given input; as we will see in section 3.2.2, this hypothesis can be relaxed if the algorithm is modified by adding an appropriate reflection term.

We distinguished between two variants of Jäger's algorithm: one, where the gain parameter η regulating the size of the learning steps is kept constant, the other, where the gain parameter itself is variable and tends to zero as learning proceeds.

In the former case the sequence of constraint rankings delivered by the algorithm weakly converges to the unique stationary distribution provided η is sufficiently small. If the underlying generator is finite or the feature function is globally bounded, convergence holds for all $\eta > 0$, i. e. for all choices of the gain parameter. Under more general

conditions, proposition 2.3 in section 2.3.2 provides a threshold so that convergence is guaranteed for all η smaller than that threshold, which is expressed in terms of the dual function f_{pot} and is thus connected to the feature function c as well as the empirical distribution p_{emp}.

In the latter case, convergence of the sequence of constraint rankings is to the maximum entropy ranking corresponding to the given empirical distribution provided the sequence of gain parameters tends to zero slowly enough. We have much freedom in the choice of the gain parameter sequence, as it only has to satisfy hypothesis (H5) of theorem 2.6 in section 2.5.

In both cases Jäger's algorithm converges independently of the initial ranking of constraints, which may be deterministic or random. By initializing the algorithm with a randomly distributed constraint ranking one can model the learning behaviour of a whole population of learners, where each learner starts with his or her own choice of a specific OT grammar. Convergence then means that after a sufficiently long time a uniquely determined distribution of specific OT grammars is reached. In case the gain parameter decreases to zero as learning proceeds, all individual grammars will converge to the same specific OT grammar, namely the maximum entropy grammar corresponding to the given empirical distribution.

One reason for introducing randomness into Optimality Theory is the robustness of stochastic learning algorithms. Here, robustness is understood primarily in the sense of stability against small variations in the learning data. Even if the language that has to be learned were produced by a (specific) deterministic grammar, the data available to the learner would in a realistic setting deviate from what the grammar yields. Hence, learning data should be modeled as a sequence of random samples drawn from a fixed probability distribution, namely the empirical distribution, on interpreted linguistic forms, where not all of the data has to belong to the target language.

Similarly, while any language produced by a grammar of the maximum entropy version of OT has compound Gibbs form (cf. section 1.3.2), the empirical distribution need not be of that form in order for the learning process to be successful. The deviation of the empirical distribution from the type of distributions generated by a certain version of stochastic OT can be taken to mirror the effect of errors in linguistic production, transmission and comprehension. More specifically, the empirical distribution can account for mistakes the learner makes in pairing outputs with underlying inputs. This partially justifies our assumption that the learner is able to observe input output pairs, not only outputs, because what is observed is the result of the learner's own interpretation. Note, however, that pairing errors of this kind are necessarily independent of the current state of grammar, i. e. independent of the current ranking of constraints.

Robustness can also be understood in a more technical way, meaning that the algorithm is insensitive to small errors due to rounding or approximation in the update step. Although this is not what we want robustness to mean, Jäger's algorithm is robust in that sense, too. To show this, we could insert an additional error term in recursion formula

(1.25). The convergence theory as developed in Kushner and Yin (2003) allows to handle such algorithms, too.

An important point which we have been ignoring up to now is the question of how fast Jäger's algorithm converges. Generally speaking, it is the geometry of the dual function f_{pot} that determines rate and speed of convergence. We just briefly sketch one approach to analyzing the rate of convergence; for details and, in particular, the case of variable gain parameter tending to zero see Kushner and Yin (2003: ch. 10).

Let (R_n^η) be a sequence of constraint rankings computed according to recursion formula (1.25) with constant gain parameter η. If $\eta > 0$ is small, then R_n^η will be concentrated about the maximum entropy ranking $\hat{r}$ for all $n \geq n_\eta$ provided n_η is big enough. Asymptotically, the distribution of R_n^η, $n \geq n_\eta$, where η has to be small and n_η big, can be characterized as solution to a linear SDE. The drift matrix A of that SDE is the negative of the Hessian of f_{pot} at $\hat{r}$, that is $A = -H_{f_{pot}}(\hat{r})$. This becomes clear if we think of a one-dimensional function having a global minimum. In a neighbourhood of the minimum position the geometry of the function is determined by its second derivative at that position. Likewise, the geometry of f_{pot} in a neighbourhood of $\hat{r}$ is determined by its Hessian at $\hat{r}$. Recall that the Hessian is positive semi-definite. The bigger its smallest eigenvalue, the more concentrated will be the distribution of R_n^η about $\hat{r}$ or, changing perspective, the faster convergence of R_n^η to $\hat{r}$ will be in the gain parameter.

3.2 Generator with infinitely many input output pairs

In deriving the maximum entropy version of stochastic OT in chapter 1 we had to employ more mathematical machinery than is usually found in the literature[1], because we allowed the generator of our stochastic grammar to consist of an infinite number of input output pairs. Here and in the sequel, if nothing else is said, infinite shall be understood as countably infinite.

In chapter 2 convergence of Jäger's algorithm was studied under the hypothesis that for a given input there were only finitely many output candidates. We first discuss why it is desirable to allow for an infinite number of different inputs. Later on in section 3.2.1 we consider the case, where not only the set of inputs, but also sets of output candidates may be infinite. A way of adapting Jäger's algorithm to the more general situation is described in section 3.2.2.

3.2.1 Infinitely many inputs and infinite candidate sets

To get a linguistic interpretation of the generator, think of the inputs as meanings in some semantic (truth-conditional) or pragmatic sense and of the outputs as corresponding linguistic expressions or forms. It is easy to create an infinitude of different meanings by

[1] Cf. Berger et al. (1996), textbooks like Aarts and Korst (1989) or the original papers by Shannon and Jaynes. See Harremoës and Topsøe (2001) for a different general approach.

repetition and recombination of a small number of basic elements. Take a binary predicate like MOTHER and two proper names, say MARIA and JULIA. We can form sentences like:

(E1) a. Julia is Maria's mother.

b. Julia is Maria's mother's mother.

c. Julia is Maria's mother's mother's mother.

Clearly, the list in (E1) can be continued ad infinitum by inserting ever more generations of mothers and daughters between Julia and Maria. It is therefore reasonable to allow the set of inputs to be infinite.

A way out would be to choose a symbolic system for representing meanings and to prescribe an upper bound on the length of representations in that system. The set of all meanings having a representation not longer than the given upper bound would clearly be finite. This approach has the disadvantage of being arbitrary in two related respects. Neither is there a canonical symbolic system for meaning representation, nor does a natural upper bound on the lengths of such representations exist, while both choices have an impact on what meanings are available to the grammar. What is more, Optimality Theory is not a theory of representations. Consequently, we should not incorporate anything pertaining to the level of representations into the set-up of our OT grammars.

A further advantage of working with infinite sets of inputs and outputs is that, in doing so, we allow for unseen material coming up after any finite number of learning or production steps. This is especially true for the maximum entropy version of stochastic OT. Recall that a probability distribution of Gibbs type on a set of output candidates necessarily has full support. Thus some probability mass is always reserved for hitherto unseen input output pairs.

When the generator is infinite, one might wonder how to compute in practice the number of constraint violations. Of course, it is not possible to store the values $c_j(i,o)$ of the j-th constraint for all input output pairs (i,o). But we may assume that the feature function c is defined by recursive application of certain rules so that $c_j(i,o)$ can be calculated in a finite number of steps for any given input output pair (i,o) and each of the finitely many constraints $c_1,\ldots,c_N$. As expressions like (E1a) through (E1c) are generated recursively, so the corresponding numbers of constraint violations should be computed recursively.

Let us now turn our attention to the output side of the generator, retaining the interpretation of inputs as meanings and outputs as associated expressions or linguistic forms. Clearly, a given meaning may have different linguistic realizations. Instead of (E1a), for example, one could say (E2a). Similarly, one obtains (E2b) instead of (E1b) and (E2c) as a reformulation of (E1c). Combining the two possessive constructions, we get sentences like (E2d) or (E2e).

(E2) a. Julia is the mother of Maria.

b. Julia is the mother of the mother of Maria.

c. Julia is the mother of the mother of the mother of Maria.

d. Julia is the mother of Maria's mother.

e. Julia is the mother of Maria's mother's mother.

We have been vague about what the inputs to expressions like those in (E1) and (E2) should be. If we take meanings in the sense of traditional truth-conditional meanings and, in addition, assume that part of the semantics of the predicate MOTHER is a one-to-many relation between mothers and children, then the sentences in (E2) are really equivalent formulations of sentences in (E1). If we wanted the inputs to carry more structure, e. g. information structure like focus, we would have to specify additional features of the output candidates, e. g. the stress or the intonation pattern.

At this point it seems that candidate sets containing more than one output arise from combination of different grammatical constructions available in the natural language under consideration. In this way we would get a finite number of outputs. But the generator of an OT grammar is supposed to be much less restrictive. It should be universal in the sense of not being language-specific except for lexical items. Hence the generator should propose as output candidates all combinations of the given elements which can possibly occur in any natural language. So, for example, all expressions in (E3) are output candidates for the input corresponding to sentence (E1a), where we take the input to consist of the predicate MOTHER, the individual constants JULIA and MARIA, the argument structure corresponding to (E1a), and some information regarding tense, mood and aspect.

(E3)
a. Julia is mother Maria.
b. Maria's mother is Julia.
c. Julia is Maria's mother.
d. of Maria mother Julia is.
e. to Maria Julia mother.
f. Maria Julia is mother of.
g. Julia mother.
h. she Maria mother.
i. Maria's mother Julia.
j. Maria Julia is be mother.
k. Julia is mother of Maria.
l. is.

The empty utterance can also be a candidate for expressing the meaning of (E1a). According to an even more radical idea of the generator, every word appearing in (E3) can be replaced by any lexical item. The generator produces templates which have to be filled in with (language-specific) lexical material. By inserting new material or repeating elements, an infinite list of possible outputs can be created by the generator component of an OT grammar. Omitting material required for expressing relations, names or other information present in the input or adding material which has no equivalent in the input leads to violation of *faithfulness* constraints. In this way, generally only a few output candidates have probability not close to zero.

3.2.2 A constrained algorithm for infinite candidate sets

We turn to the question of how Jäger's algorithm has to be modified if it is to handle possibly infinite sets of output candidates. Let (G, c, p) be a universal stochastic OT grammar, where the evaluation kernel $p = (p_r)_{r\in\mathcal{R}}$ is of maximum entropy type, the feature function c is N-dimensional and the generator $G \subseteq I \times O$ is such that for some input $i \in I$ the candidate set O_i is infinite.

In chapter 2 we worked under the hypothesis that the set of admissible ranking parameters $\mathcal{R}$ was maximal in the sense that $\mathcal{R} = \mathbb{R}^N$. For the maximum entropy version of stochastic OT, however, the set $\mathcal{R}$ is given by – or must at least be a subset of – the set of Gibbs parameters defined by (1.20). The problem is whether the normalizing constant $Z_r(i)$ converges if O_i is infinite. Since c is non-negative, for such input i we have

$$Z_r(i) \;=\; \sum_{\tilde{o}\in O_i} \exp\big(-\langle r, c(i,\tilde{o})\rangle\big) \;\geq\; \sum_{\tilde{o}\in O_i} 1 \;=\; \infty \quad \text{for all } r \in (-\infty, 0]^N.$$

For parameters $r \in \mathbb{R}^N \setminus (-\infty, 0]^N$ convergence of $Z_r(i)$ depends on the feature function c. We will be more restrictive than necessary in limiting attention to elements of $(0,\infty)^N$ as possible ranking vectors. Set

$$\mathcal{R}_+ := \big\{ r \in (0,\infty)^N \;\big|\; Z_r(i) < \infty \;\forall\, i \in I \big\}. \tag{3.1}$$

From a conceptual point of view, restriction of the set of admissible constraint rankings to $\mathcal{R}_+$ may even be an advantage. Take two output candidates which differ only in the number of violations of a single constraint. Then the output which incurs less violations should be more probable. But if the decisive constraint has negative rank, it will be the other way round: the output incurring more constraint violations will also be more probable.

Recall from section 1.4 that Jäger's algorithm produces a random sequence $(R_n^\eta)_{n\in\mathbb{N}_0}$ of constraint rankings by iteratively applying formula (1.25). The update rule is

$$R_{n+1}^\eta := \; R_n^\eta \;+\; \eta \cdot \big(c(X_n^{in}, H_n(X_n^{in}, R_n^\eta)) - c(X_n^{in}, X_n^{out})\big),$$

where η is the gain parameter and (X_n^{in}, X_n^{out}) provides the n-th input output pair drawn at random from a given empirical distribution. For an input i and a ranking of constraints r the random variable $H_n(i, r)$ yields the corresponding output drawn at the n-th step according to the Gibbs distribution $p_r(.|i)$ on the candidate set O_i.

We have already made use of the fact that the sequence of constraint rankings produced by the algorithm moves on an N-dimensional Euclidean grid with mesh size η. When working with $\mathcal{R}_+$ instead of $\mathcal{R} = \mathbb{R}^N$ there is a new difficulty: Updating the current constraint ranking according to (1.25) might lead out of the set $\mathcal{R}_+$. In this case the Gibbs distributions on the candidate sets would not be guaranteed to be well defined any more. To prevent such a situation from occurring, we add a reflection term to the update rule. This term neutralizes any update which would take the constraint ranking outside

the set $\mathcal{R}_+$ of admissible rankings. To be more precise, define a family of $\mathbb{R}^N$-valued random variables $Y_n^\eta(i,o,r)$ by

(3.2)
$$Y_n^\eta(i,o,r) := \begin{cases} \eta \cdot \big(c(i,o) - c(i,H_n(i,r))\big) & \text{if} \quad r + \eta\,\big(c(i,H_n(i,r)) - c(i,o)\big) \notin \mathcal{R}_+, \\ 0 & \text{else,} \end{cases}$$

where $n \in \mathbb{N}_0$, $(i,o,r) \in I \times O \times \mathcal{R}_+$. Then recursion formula (1.25) gets replaced by

$$R_{n+1}^\eta := R_n^\eta \;+\; \eta \cdot \big(c(X_n^{in}, H_n(X_n^{in}, R_n^\eta)) - c(X_n^{in}, X_n^{out})\big) \;+\; Y_n^\eta(X_n^{in}, X_n^{out}, R_n^\eta). \tag{3.3}$$

We refrain from carrying out a convergence analysis of the modified algorithm along the lines of chapter 2. Such an analysis is essentially not more difficult than what we have already done. In Kushner and Yin (2003) all important results are stated for unbounded as well as constrained algorithms. Although notation is more cumbersome, inclusion of a reflection term may make life easier. The results of section 2.3, in particular, are easily transferred to the new situation. If the random variables Y_n^η were redefined so as to force the sequence of constraint rankings to stay in a compact subset of $\mathbb{R}^N$, existence of a stationary distribution would be an almost trivial consequence.[2]

Of course, some assumptions on the feature function c and the empirical distribution are needed. In particular, c must be such that the set of admissible rankings $\mathcal{R}_+$ is non-empty. As in theorem 1.1, the empirical distribution should have finite entropy. Since $\mathcal{R}_+$ is now allowed to be a proper subset of $\mathcal{R}$, there may exist a maximum entropy distribution which, however, cannot be learned by the algorithm as defined in (3.3).

There is also another problem connected with the normalizing constants $Z_r(i)$ in case O_i is infinite, because then $Z_r(i)$ is an infinite sum. As such it can only be approximated. The normalizing constant $Z_r(i)$ is needed in order to compute the Gibbs distribution $p_r(.|i)$ on O_i. Computation of those Gibbs distributions in turn is essential for the evaluation component of any maximum entropy OT grammar.[3] It seems we are borne back to where we started, namely to the hypothesis that any set of output candidates must be finite. But still there is an advantage in allowing candidate sets to be infinite, as the quality of approximation, e. g. the number of iteration steps in calculating $Z_r(i)$, may then be made dependent not only on the input i, but also on the current ranking r. When the norm of r is small, we may allow more output candidates to be considered than when $\|r\|$ is big.

What is more, even if O_i is finite, the constant $Z_r(i)$ can in general only be approximated, for example by Monte Carlo simulation. The same method allows to directly simulate a Gibbs distribution on O_i without any need for computing $Z_i(r)$ (cf. Brémaud, 1999: ch. 7). A random experiment giving outputs according to a Gibbs distribution $p_r(.|i)$ on a finite set O_i can also be realized as an artificial neural net (cf. Aarts and Korst, 1989).

[2] One only needs irreducibility of the Markov chain corresponding to the sequence of constraint rankings.

[3] In recursion formulae (1.25) and (3.3) the family of random variables H_n takes over the rôle of the evaluation kernel.

3.3 An application to syntax

Following Jäger and Rosenbach (2005), we present an application of the maximum entropy version of stochastic OT to the phenomenon of English genitive variation. The study also illustrates an important property of stochastic OT grammars, namely that they may exhibit *cumulativity effects*. Before going into details, we make two remarks concerning the application of OT type grammars to linguistic phenomena.

Firstly, in most applications of Optimality Theory only small part of what would be a universal OT grammar is considered. The generator is a "small" set of input output pairs, and only few constraints are taken into account. Apart from technical considerations and convenience, this course of action is justified by the way the evaluation component and learning work – at least for the versions of stochastic OT we have been considering here. A constraint which is constant on the entire set of output candidates for a given input has no impact on the distribution of outputs. Similarly, learning can change the rank of a constraint only if an input output pair is encountered such that not all the other output candidates satisfy the constraint in question equally well.[4] It is therefore reasonable to restrict attention to a subset of admissible input output pairs and few constraints.

The second remark concerns the structure of grammar. If universal grammar consists of a chain of independent submodules, then it is possible to represent each submodule by a universal stochastic OT grammar of its own. At the very least one would have a syntactic and a phonological module, the outputs of the first serving as inputs to the second. Notice that Optimality Theory was originally developed in the context of phonology. Inputs are, for example, sequences of consonants and vowels, outputs syllabified consonant vowel sequences (cf. Prince and Smolensky, 2004).

In English, there are two genitive constructions for expressing a possessive relation: the *s*-genitive and the prepositional *of*-genitive. In many situations both constructions are acceptable, although variation is not entirely free. The study by Jäger and Rosenbach (2005) compares determiner *s*-genitive and *of*-genitives, where the possessor is a complement.[5] Consider the following pairs of alternative expressions (table 2 in Jäger and Rosenbach, 2005):

(E4) a. the boy's eyes / the eyes of the boy

b. the mother's future / the future of the mother

c. a girl's face / the face of a girl

d. a woman's shadow / the shadow of a woman

e. the chair's frame / the frame of the chair

f. the bag's contents / the contents of the bag

[4] This is true of Jäger's algorithm as well as Boersma's GLA.

[5] Genitive constructions expressing an attributive relation like *a man of honour* are not comparable to genitives expressing possession, as the corresponding inputs are necessarily different.

g. a lorry's wheels / the wheels of a lorry

h. a car's fumes / the fumes of a car

The alternatives in (E4) can be classified according to three factors, each corresponding to a binary feature:

- $[\pm a]$: animacy of the possessor (entity with a soul),
- $[\pm t]$: topicality of the possessor (reference is unique),
- $[\pm p]$: prototypicality of the relation between possessor and possessum (part-of-relation is prototypical).

Combination of these features yields eight types of possessive relations, each represented by one of the construction pairs in (E4). Pair (E4c), for example, is an instance of feature combination $[+a][-t][+p]$.

In an empirical study native speakers were presented with short text passages containing a possessive relation the informants had to express by choosing between the *s*- and the *of*-genitive.[6] As a result it was found that the relative frequency with which *s*-genitive constructions were preferred over *of*-genitives decreases from possessive relations of type $[+a][+t][+p]$ to type $[-a][-t][-p]$ if the ordering

$$\text{animacy} \quad \succ \quad \text{topicality} \quad \succ \quad \text{prototypicality}$$

is assumed. The relative frequencies were recorded for each construction pair in (E4). For the input underlying (E4a) the *s*-genitive was opted for in 89.3% of all occurrences, while the preferred way of saying (E4h) was in 88.1% of all cases the *of*-genitive. Note that there are 16 input output pairs: two output candidates for each of the eight different inputs. Incorporating a distribution over inputs, one obtains an empirical distribution over the input output pairs.

In order to reproduce the results by means of a stochastic OT grammar, take constraints of the form

$$* \begin{bmatrix} \pm a \\ \pm t \\ \pm p \end{bmatrix} /[\pm\text{prenom}], \tag{3.4}$$

where the asterisk means "avoid", a, t, p stand for the three features with values $+$ or $-$, $/[+\text{prenom}]$ means "in combination with prenominal possessor" and $/[-\text{prenom}]$ means "in combination with postnominal possessor". The constraint $*[+p]/[+\text{prenom}]$, for example, is violated iff a prototypical possessive relation is realized as *s*-genitive construction. Notice that the value of all features referred to in (3.4) must be derivable from any given input output pair.

[6] See references in Jäger and Rosenbach (2005).

Jäger's algorithm run over samples drawn from a simulated corpus mirroring the empirical distribution obtained in the experiments produces the following ranking of constraints (see Jäger and Rosenbach, 2005):

$*[+a]/[+prenom]$	9.476	$*[-a]/[+prenom]$	10.644
$*[+a]/[-prenom]$	10.524	$*[-a]/[-prenom]$	9.356
$*[+t]/[+prenom]$	9.746	$*[-t]/[+prenom]$	10.374
$*[+t]/[-prenom]$	10.254	$*[-t]/[-prenom]$	9.626
$*[+p]/[+prenom]$	9.895	$*[-p]/[+prenom]$	10.225
$*[+p]/[-prenom]$	10.105	$*[-p]/[-prenom]$	9.775

Observe that in each competition only six of the twelve constraints are active[7] – depending on the feature values of the underlying input. The acquired ranking induces a distribution on input output pairs which gives a good approximation to the empirical distribution: Kullback-Leibler distance (with logarithmic base 2) is about 0.00472 bit. Moreover, animacy is learned as the strongest factor, followed by topicality and prototypicality.

The variation between *s*-genitive and *of*-genitive illustrates the effect of *ganging-up* cumulativity if we abstract away from specific outputs and regard only the genitive construction. Going from an input of type $[+a][+t][-p]$ to one of type $[+a][-t][-p]$ the probability assigned to the *of*-construction increases, because the two $[+t]$-constraints, which together favour the *s*-genitive, become inactive, while the two $[-t]$-constraints, which together favour the *of*-construction, are "switched on". Note that $*[+a]/[-prenom]$ is higher ranked than all other active constraints. Ganging-up cumulativity in general means that lower ranked constraints matter and can conspire to overrule a high ranked constraint.

Further experiments were carried out with the aim of establishing which impact the *weight* of the possessor phrase had on the choice of the genitive construction. In a corpus analysis, weight was measured as the number of prenominal modifiers in the possessor phrase.[8] Of the three features animacy, topicality, prototypicality only animacy of the possessor was retained. Examples of genitive constructions with animate possessors of different weight are

(E5) a. Pauline's birthday / the birthday of Pauline

b. the doctor's daughter / the daughter of the doctor

c. the other person's nose / the nose of the other person

d. right honourable gentleman's policy / the policy of the right honourable gentleman

[7] A constraint is *active* in a competition if it is not constant on the set of candidate outputs.

[8] The possessum was kept at constant length. The impact of the *relative weight* between possessor and possessum was also tested.

Although with animate possessor the *s*-genitive is usually preferred over the *of*-construction, this preference changes as the possessor gets longer. Thus, *s*-genitives were found in 84.2% of all instances of (E5a), while in cases like (E5d) the portion of *s*-genitives fell to 35.7%. With inanimate possossors there is a strong tendency towards the *of*-genitive, but if the possessor is short, the *s*-genitive has about 12% probability.

The resulting empirical distribution can be faithfully reproduced by a maximum entropy OT grammar using Jäger's learning algorithm, if there are four constraints of the form $*[\pm a]/[\pm prenom]$ plus one constraint, say $*[premod][+prenom]$, which counts the number of (prenominal) modifiers of a prenominal possessor. That constraint is not violated if the output is an *of*-genitive, since then the possessor follows the possessum.

After learning, constraint $*[premod][+prenom]$ is ranked considerably lower than the other constraints. Multiple violations of $*[premod][+prenom]$, however, strengthen its effect so that it becomes predominant. This is referred to as *counting cumulativity*.

The work by Jäger and Rosenbach (2005) provides abstract definitions of both cumulativity effects. We should mention that Boersma's version of stochastic OT allows for ganging-up cumulativity, but not for counting cumulativity, while the maximum entropy version allows for both. The fact that maximum entropy OT grammars exhibit ganging-up as well as counting cumulativity is clear by the log-linear nature of the Gibbs distributions.

3.4 Possible extensions and conclusions

Up to now our perspective on OT has been that of production. An input – corresponding to a meaning – was given and the grammar had to provide appropriate linguistic expressions as outputs. This is the *speaker* or *expressive* perspective. Changing point of view by taking an output o and looking for all inputs i such that (i, o) is in the generator G, we would be confronted with the problem of computing a probability distribution over the possibly infinite set

$$I_o := \{i \in I \mid (i, o) \in G\}.$$

A probability distribution on I_o can be calculated according to the maximum entropy version, Boersma's version of stochastic OT or deterministic OT provided I_o is finite. If I_o is infinite, there may again be a problem with convergence of the normalizing constant in case the maximum entropy version is used. If Boersma's version or deterministic OT is applied, then one must make sure that with probability one there are only finitely many optimal outputs for each input.

Starting on the output side of the generator means adopting the *interpretative* or *hearer* perspective. This is reminiscent of what is called "lexicon optimization" in Prince and Smolensky (2004: §9.3), and it is an essential ingredient in *bidirectional Optimality Theory*, where speaker and hearer perspective are combined, see Blutner (2004). In the context of learning, by taking into account the hearer perspective, one can model how the learner associates observed outputs with underlying inputs. In particular, pairing errors in the

sense of section 3.1 can then vary according to the current state of the learner's grammar and need not be specified a priori by incorporating them into the empirical distribution.

Little has been said about convergence speed and rate of convergence of Jäger's algorithm (cf. section 3.1), implementational issues and connections of stochastic OT with artificial neural networks (cf. section 3.2.1). We briefly sketched how Jäger's algorithm has to be adapted in order to deal with the general situation, where a single input may have infinitely many output candidates (cf. section 3.2.2).

We gave a unified definition of stochastic Optimality Theory as a conception of universal grammar, which enabled us to derive various versions of probabilistic Optimality Theory as special cases. We described and stated the learning problem as it arises for stochastic OT grammars, namely the problem of learning the "right" ranking of constraints given an empirical distribution of learning data. The maximum entropy version of stochastic OT, which has been our main concern, was derived in great generality from the principle of entropy maximization. Jäger's algorithm, a stochastic learning procedure for maximum entropy OT, was introduced, and its connections with classical iterative algorithms such as gradient descent and stochastic descent were explained.

Two variants of Jäger's algorithm can be distinguished, one, where the gain parameter corresponding to the size of learning steps is constant, the other, where the gain parameter tends to zero as learning proceeds. Convergence in an appropriate sense was proved for both variants under mild assumptions on learning data and constraints, which are determined by the underlying universal OT grammar. Assumptions were spelled out in detail. In deriving the maximum entropy version of stochastic OT and in the proofs of convergence of Jäger's algorithm, a considerable amount of mathematical machinery was needed due to the fact that we allowed the generator of our OT grammar to consist of infinitely many input output pairs.

Jäger's algorithm was shown to be a robust and flexible procedure for solving the learning problem associated with the maximum entropy version of stochastic OT, which in turn directly derives from fundamental principles of information theory.

Appendix A

Convex sets and convex functions

Here, we collect definitions and results related to the concept of convexity. Let V be a real vector space of possibly infinite dimension. In case V is a normed space, denote by $\mathrm{B}_\rho(x)$ the open ball and by $\overline{\mathrm{B}_\rho(x)}$ the closed ball centered at $x \in V$ with radius ρ.

Definition A.1. A subset $S \subseteq V$ is called *convex* iff $x, v \in S$ implies $\lambda x + (1-\lambda)v \in S$ for all $\lambda \in [0,1]$.

Definition A.2. Let S be a convex subset of $\mathbb{R}^N$. Let $\mathrm{hull}_{\mathrm{aff}}(S)$ denote the smallest affine-linear manifold containing S. An element $x \in S$ is called a *relative interior point* of S iff there is $\rho > 0$ such that

$$\mathrm{B}_\rho(x) \ \cap \ \mathrm{hull}_{\mathrm{aff}}(S) \ \subseteq \ S.$$

Definition A.3. Let S be a convex subset of V. A function $f : S \to \mathbb{R} \cup \{\infty\}$ is said to be *convex* iff for all $x, v \in S$ and all $\lambda \in [0,1]$

$$f\left(\lambda x + (1-\lambda)v\right) \quad \leq \quad \lambda f(x) \ + \ (1-\lambda) f(v). \tag{A.1}$$

A function $f : S \to \mathbb{R}$ is called *strictly convex* iff f is convex and the inequality in (A.1) is strict for all $x, v \in S$, $x \neq v$, and all $\lambda \in (0,1)$.

Any $\mathbb{R}$-valued convex function defined on a convex open subset or an affine-linear submanifold of $\mathbb{R}^N$ is continuous. This continuity property is not guaranteed if instead of $\mathbb{R}^N$ we have a vector space of infinite dimension.

If a convex function $f : S \to \mathbb{R}$ defined on an affine-linear submanifold S of $\mathbb{R}^N$ grows to infinity along all rays starting at some point in the domain of definition, then f possesses a global minimum.

Lemma A.1. *Let $V \subseteq \mathbb{R}^N$ be a linear subspace and $r \in \mathbb{R}^N$. Set $S := r + V$, and let $f : S \to \mathbb{R}$ be a convex function such that $f(r + t \cdot v) \to \infty$ as $t \to \infty$ for all $v \in V \setminus \{0\}$. Then f possesses a global minimum. If, in addition, f is strictly convex, then it has exactly one extremum, namely an isolated global minimum.*

Proof. Let $\mathrm{S}_1(0)$ be the unit sphere in $\mathbb{R}^N$. Set $K := V \cap \mathrm{S}_1(0)$. For $\rho > 0$ define a mapping $t_\rho \colon K \to [0,\infty)$ by $t_\rho(v) := \inf\{t \geq 0 \mid f(r + t \cdot v) > \rho\}$. Notice that $t_\rho(v)$ is finite since $f(r + t \cdot v) \to \infty$ as $t \to \infty$. Choose $\rho > f(r)$ and let $v \in K$. Then $t_\rho(v) > 0$, and $f(r + t_\rho(v) \cdot v) = \rho$ since $[0,\infty) \ni t \mapsto f(r + t \cdot v)$ is continuous. With $t \geq t_\rho(v)$, convexity of f implies

$$\begin{aligned} f\big(r + t_\rho(v) \cdot v\big) &= f\Big(\big(1 - \tfrac{t_\rho(v)}{t}\big) r + \tfrac{t_\rho(v)}{t}(r + t \cdot v)\Big) \\ &\leq \tfrac{t_\rho(v)}{t} f(r + t \cdot v) \;+\; \big(1 - \tfrac{t_\rho(v)}{t}\big) f(r) \\ \Rightarrow \quad f(r + t \cdot v) &\geq f\big(r + t_\rho(v) \cdot v\big) \;+\; \tfrac{t - t_\rho(v)}{t_\rho(v)} \cdot \Big(f\big(r + t_\rho(v) \cdot v\big) - f(r)\Big). \end{aligned}$$

In particular, we have $f(r + t \cdot v) \geq \rho$ for all $t \geq t_\rho(v)$. From the above inequality and the continuity of f one deduces that t_ρ is a continuous positive function on K provided that $\rho > f(r)$.[1]

Clearly, K is compact. Hence t_ρ attains its maximum and $\tau := \max\{t_\rho(v) \mid v \in K\}$ is finite. Set $B := V \cap \overline{\mathrm{B}_\tau(r)}$. Then $f(x) \geq \rho$ for all $x \in S \setminus B$. On the other hand, B also is compact, whence f attains its minimum $y := \min\{f(x) \mid x \in B\}$ on B. Since $r \in B$ we have $y \leq f(r)$, but $f(r) < \rho$. Therefore, y is really the global minimum value of f.

For the rest of the proof let us assume that f is strictly convex. Suppose f had a local maximum at position $\overline{x} \in S$. Then we could choose $x, \tilde{x} \in S$, $x \neq \tilde{x}$, and $\lambda \in (0,1)$ such that both $x \leq f(\overline{x})$ and $\tilde{x} \leq f(\overline{x})$ and $\overline{x} = \lambda x + (1{-}\lambda)\tilde{x}$. The strict version of (A.1) would imply $f(\overline{x}) < f(\overline{x})$ – a contradiction.

We already know that f has a global minimum. Set $y := \min\{f(x) \,|\, x \in S\}$, and let $\hat{x}, \tilde{x}$ be elements of S such that $f(\hat{x}) = y = f(\tilde{x})$. Then $f(\lambda\hat{x} + (1{-}\lambda)\tilde{x}) \geq y$ for all $\lambda \in (0,1)$, and strict convexity implies $\tilde{x} = \hat{x}$. Therefore $f(\hat{x}) < f(x)$ for all $x \in S$.

Finally, notice that f cannot have a local minimum at $\tilde{x} \neq \hat{x}$. Else we could choose $x \in S \setminus \{\tilde{x}, \hat{x}\}$ and $\lambda \in (0,1)$ such that $f(x) \geq f(\tilde{x})$ and $x = \lambda\tilde{x} + (1{-}\lambda)\hat{x}$. But $f(\hat{x}) \leq f(\tilde{x})$, so the strict version of inequality (A.1) would again lead to a contradiction. □

Lemma A.2. *Let $f \colon \mathbb{R}^N \to \mathbb{R}$ be a convex and continuously differentiable function, and suppose that f has a global minimum at $\hat{x} \in \mathbb{R}^N$. For $\rho \geq 0$ set $\varepsilon_\rho := \min\{f(x) - f(\hat{x}) \mid \|x - \hat{x}\| = \rho\}$. Then ε_ρ as a function of $\rho \in [0,\infty)$ is non negative, non decreasing, and $\varepsilon_{t\rho} \geq t \cdot \varepsilon_\rho$ for all $t \geq 1$, $\rho \geq 0$. In particular, either $\varepsilon_\rho = 0$ for all $\rho \geq 0$, or ε_ρ tends to infinity as $\rho \to \infty$.*

Moreover, for all $\rho > 0$ and all $x \in \mathbb{R}^N$ such that $\|x - \hat{x}\| \geq \rho$ it holds that

$$(\star) \qquad \big\|\nabla f(x)\big\| \;\geq\; \left\langle \nabla f(x), \frac{x - \hat{x}}{\|x - \hat{x}\|} \right\rangle \;\geq\; \frac{f(x) - f(\hat{x})}{\|x - \hat{x}\|} \;\geq\; \frac{\varepsilon_\rho}{\rho}.$$

[1] Observe that t_ρ need not be continuous if f satisfies only the growth condition of lemma A.1 and is continuous, but not convex. Consider, for example, a smooth function having "bumps" of height greater than ρ.

Proof. Notice that the minimum in the definition of ε_ρ is justified in that the infimum of a continuous function over a compact set, here the sphere of radius ρ centered at $\hat{x}$, is really a minimum, and that $\varepsilon_\rho \geq 0$, because f attains a global minimum at $\hat{x}$. Let $\rho > 0$ and $v \in \mathbb{R}^N$ such that $\|v - \hat{x}\| = \rho$. By convexity of f we have for all $t \geq 0$, $\lambda \in [0,1]$

$$f\big(\lambda(t(v-\hat{x})+v) + (1-\lambda)\hat{x}\big) \quad \leq \quad \lambda f(t(v-\hat{x})+v) \; + \; (1-\lambda)f(\hat{x}).$$

Choosing $\lambda := \frac{1}{t+1}$ we obtain for all $v \in \mathbb{R}^N$ with $\|v - \hat{x}\| = \rho$ and all $t \geq 0$

$$f(t(v-\hat{x})+v) \; - \; f(\hat{x}) \quad \geq \quad (t+1)\big(f(v) - f(\hat{x})\big) \quad \geq \quad (t+1)\varepsilon_\rho.$$

The first inequality in ($\star$) is clear, since $\langle v, v\rangle \geq \langle v, w\rangle$ for all $v, w \in \mathbb{R}^N$ such that $\|w\| = \|v\|$. As a consequence of inequality (A.1) we have with $h \in (0,1)$

$$f\big(x + h(\hat{x}-x)\big) \quad = \quad f\big((1-h)x + h\hat{x}\big) \quad \leq \quad (1-h)f(x) \; + \; hf(\hat{x})$$

$$\Rightarrow \quad \frac{1}{h}\Big(f\big(x + h(\hat{x}-x)\big) - f(x)\Big) \quad \leq \quad f(\hat{x}) \; - \; f(x).$$

Observing that $\langle \nabla f(x), x - \hat{x}\rangle$ is the directional derivative along $x - \hat{x}$ we obtain the middle part of ($\star$). A further application of inequality (A.1) yields

$$f\big(\hat{x} + \tfrac{\rho}{\|x-\hat{x}\|}(x-\hat{x})\big) \quad \leq \quad \big(1 - \tfrac{\rho}{\|x-\hat{x}\|}\big)f(\hat{x}) \; + \; \tfrac{\rho}{\|x-\hat{x}\|} \cdot f(x)$$

$$\Rightarrow \quad f(x) \; - \; f(\hat{x}) \quad \geq \quad \tfrac{1}{\rho} \cdot \big(f\big(\hat{x} + \tfrac{\rho}{\|x-\hat{x}\|}(x-\hat{x})\big) - f(\hat{x})\big) \cdot \|x - \hat{x}\|.$$

Clearly, $\|\frac{\rho}{\|x-\hat{x}\|}(x-\hat{x})\| = \rho$, and the last part of ($\star$) follows. □

We observe that the function $\rho \mapsto \varepsilon_\rho$ as defined in lemma A.2 is increasing if f has an isolated global minimum at $\underline{x}$.

Appendix B

Optimization and Lagrange's method

The following ancient and simple observation is of great importance for solving problems of constrained optimization.

Proposition B.1 (Lagrange's lemma). *Let S be a non-empty set and $f\colon S \to \mathbb{R}$, $g\colon S \to \mathbb{R}^N$ be arbitrary functions. Set $S_0 := \{x \in S \mid g(x) = 0\}$. If there are $\lambda \in \mathbb{R}^N$ and $x_0 \in S_0$ such that the Lagrange function $L\colon S \to \mathbb{R}$ defined by*

$$S \ni x \mapsto f(x) \ + \ \big\langle \lambda, g(x) \big\rangle$$

is minimal at x_0, then the restriction of f to S_0 has a minimum at x_0.

Proof. With λ, x_0 as hypothesized and $x \in S_0$ it holds that

$$f(x_0) \ = \ f(x_0) + \big\langle \lambda, g(x_0) \big\rangle \ \leq \ f(x) + \big\langle \lambda, g(x) \big\rangle \ = \ f(x).$$

□

The parameter λ appearing in proposition B.1 is called a *Lagrange multiplier*. When the sets S, S_0 have additional structure, a version of Lagrange's theorem might be found which guarantees the existence of a Lagrange multiplier provided a minimum solution to the restricted problem exists and certain regularity conditions are met. The next proposition covers the special case of affine-linear constraints.

Proposition B.2. *Let V be a real vector space, $S \subseteq V$ be a convex subset and $f : S \to \mathbb{R}$ be a convex function. Let $\Gamma : V \to \mathbb{R}^N$ be a linear mapping and $y \in \mathbb{R}^N$ such that $S_0 := \{x \in S \mid \Gamma(x) = y\}$ is non-empty. Suppose $\inf\{f(x) \mid x \in S\}$ is finite and y is a relative interior point of $\Gamma(S)$. Then a Lagrange multiplier $\lambda \in \mathbb{R}^N$ exists such that*

$$\inf\{f(x) \mid x \in S_0\} \ = \ \inf\{f(x) + \langle \lambda, \Gamma(x) - y\rangle \mid x \in S\}.$$

If, in addition, an element $x_* \in S_0$ exists such that f restricted to S_0 attains its minimum at x_*, i.e. $f(x_*) = \inf\{f(x) \mid x \in S_0\}$, then it also holds that

$$f(x_*) + \langle \lambda, G(x_*) - y\rangle \ = \ \inf\{f(x) + \langle \lambda, G(x) - y\rangle \mid x \in S\}.$$

We state without proof the fundamental theorem of convex optimization, which provides an optimality criterion in terms of the right-sided Gâteaux derivative.

Theorem B.1 (Convex optimization). *Let V be a real vector space, $S \subseteq V$ be a convex subset and $f : S \to \mathbb{R}$ be a convex function. Let x_* be any element of S. Then f attains a minimum at x_* if and only if it holds that*

$$f'_+(x_*, x - x_*) \geq 0 \quad \text{for all } x \in S. \tag{B.1}$$

Bibliography

Emile Aarts and Jan Korst. *Simulated Annealing and Boltzmann Machines*. John Wiley & Sons, Chichester, 1989.

Heinz Bauer. *Wahrscheinlichkeitstheorie*. Walter de Gruyter, Berlin, 4th edition, 1991.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

Reinhard Blutner. Pragmatics and the lexicon. In Laurence R. Horn and Gregory Ward, editors, *Handbook of Pragmatics*, chapter 22. Blackwell, Oxford, 2004.

Paul Boersma. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences*, 21:43–58, 1997.

Paul Boersma. *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. PhD thesis, University of Amsterdam, 1998.

Paul Boersma and Bruce Hayes. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32(1):45–86, 2001.

Pierre Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. Springer-Verlag, New York, 1999.

Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 1986.

Sharon Goldwater and Mark Johnson. Learning OT constraint rankings using a maximum entropy model. In J. Spenader, A. Eriksson, and Ö. Dahl, editors, *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pages 111–120. Stockholm University, 2003.

Peter Harremoës and Flemming Topsøe. Maximum entropy fundamentals. *Entropy*, 3: 191–226, 2001. Online at http://www.mdpi.org/entropy.

Gerhard Jäger. Maximum entropy models and stochastic optimality theory. 2003.

Gerhard Jäger and Anette Rosenbach. The winner takes it all – almost. Cumulativity in grammatical variation. 2005.

E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106(4):620–630, 1957a.

E. T. Jaynes. Information theory and statistical mechanics II. *Phys. Rev.*, 108(2):171–190,

1957b.

Frank Keller and Ash Asudeh. Probabilistic learning algorithms and optimality theory. *Linguistic Inquiry*, 33(2):225–244, 2002.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35 of *Applications of Mathematics*. Springer-Verlag, New York, 2nd edition, 2003.

E. L. Lehmann. *Theory of Point Estimation*. Springer-Verlag, New York, 1983.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge (MA), 1999.

Alan Prince and Paul Smolensky. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell Publishing, Malden (MA), 2004.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Stat.*, 22:400–407, 1951.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

Daniel W. Stroock and S. R. Srinivasa Varadhan. *Multidimensional Diffusion Processes*, volume 233 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, 1979.