Probability Models for Complex Systems

by
Zhiyi Chi

Sc.M. in Mathematics, Brown University, Providence, Rhode Island, 1994

Thesis

Submitted in partial fulfillment of the requirements for
the Degree of Doctor of Philosophy
in Division of Applied Mathematics at Brown University

Brown University
Providence, Rhode Island
May 1998

Abstract of “Probability Models for Complex Systems,” by Zhiyi Chi, Ph.D., Brown University, May 1998

This thesis is a collection of essays on probability models for complex systems.

Chapter 1 is an introduction to the thesis. The main point made here is the importance of probabilistic modeling to complex problems of machine perception.

Chapter 2 studies minimum complexity regression. The results include: (1) weak consistency of the regression, (2) divergence of estimates in L^2-norm with an arbitrary complexity assignment, and (3) condition on complexity measure to ensure strong consistency.

Chapter 3 proposes compositionality as a general principle for probabilistic modeling. The main issues covered here are: (1) existence of general compositional probability measures, (2) subsystems of compositional systems, and (3) Gibbs representation of compositional probabilities.

Chapter 4 and 5 establish some useful properties of probabilistic context-free grammars (PCFGs). The following problems are discussed: (1) consistency of estimated PCFGs, (2) finiteness of entropy, momentum, etc, of estimated PCFGs, (3) branching rates and re-normalization of inconsistent PCFGs, and (4) identifiability of parameters of PCFGs.

Chapter 6 proposes a probabilistic feature based model for languages. Issues dealt with in the chapter include: (1) formulation of such grammars using maximum entropy principle, (2) modified maximum-likelihood type scheme for parameter estimation, (3) a novel pseudo-likelihood type estimation which is more efficient for sentence analysis.

Chapter 7 develops a novel model on the origin of scale invariance of natural images. After presenting the evidence of scale invariance, the chapter goes on to: (1) argue for a $1/r^3$ law of size of object, (2) establish a 2D Poisson model on the origin of scale invariance, and (3) show numerical simulation results for this model.

Chapter 8 is a theoretical extension of Chapter 7. A general approach to construct scale and translation invariant distributions using wavelet expansion is formulated and applied to construct scale and translation invariant distributions on the spaces of generalized functions and functions defined on the whole integer lattice.

Preface

Probabilistic modeling, often called statistical modeling, is becoming increasingly important to the study of many areas of science. The reason for this is twofold. On the one hand, many problems in modern science and technology are so complicated that they cannot be solved accurately by using simple and deterministic rules. However, by introducing stochastic mechanism into the solution, it is possible to find good approximate answers to these problems. For instance, stochastic annealing processes have been used to attack a wide range of hard optimization problems. The performance of the stochastic approach depends largely on how well it incorporates the stochastic mechanism with the elements of the problems. On the other hand, many natural and social phenomena are characterized by a variety of randomness. As an example, in medicine, people observe that the number of the cases of a disease often varies from region to region and from time to time. Usually statistical methods are the main tools to study such phenomena and the effectiveness of these methods relies on how well they model the phenomena and their randomness. It is fair to say that probabilistic modeling is of fundamental importance to the implementation of statistical methods.

This thesis is a collection of essays which have a common theme: the study of complex systems by probabilistic modeling. Under this theme, the essays cover a range of problems which can be roughly divided into five categories: (1) statistical estimation, (2) methodology of probabilistic modeling, (3) probabilistic language model, (4) probabilistic vision model, and finally, (5) probability theory.

Chapter 1 is an introduction to the thesis. From the principle of Grenander's pattern theory, we give further arguments for the importance of probabilistic modeling to vision, speech recognition, or all of machine perception. We also point out the contribution of the results in this thesis to probabilistic modeling.

Chapter 2 is concerned with non-parametric estimation, which is a classical problem in statistics. We study regression based on the minimum complexity principle and establish several consistency results on this estimation method. The results demonstrate that in order to get strong consistency, complexity measures of functions should be tied with the actual behaviors of functions.

Chapter 3 proposes a general theory and methodology, the compositionality principle, for probabilistic modeling of patterns. We introduce the notion of compositionality and formulate composition systems mathematically. The main theoretical result in this chapter is the existence of general probabilistic composition systems. We also introduce the notion of subsystems and represent the compositional probability distributions in the form of Gibbs distribution.

Chapters 4, 5, and 6 study probability models for languages. The first two chapters are devoted to probabilistic context-free grammars, which are among the simplest grammars for languages. In Chapter 4, we demonstrate that estimated production probabilities of a probabilistic context-free grammar always impose a proper distribution on the set of finite parse trees. In Chapter 5, we generalize the results in last chapter and develop an array of other useful statistical results on probabilistic context-free grammars.

As is well known in linguistics, context-free grammars are too weak to well approximate natural languages. In Chapter 6, we introduce a much more general probabilistic language model called probabilistic feature based grammar, which incorporates the theory of unification grammars and the theory of Gibbs distributions. We introduce a pseudo-likelihood type scheme for parameter estimation, which is efficient for language analysis. We also study the more classical maximum-likelihood type estimation scheme and prove the consistency of both schemes.

Chapter 7 applies probabilistic modeling to another complex system — the space of natural images. As is widely believed, statistics of natural images are of fundamental importance to vision as well as image processing. One of the most distinguishing and intriguing statistical properties of natural images is scale invariance of many marginal distributions of images. We establish a model on the origin of scale invariance of natural images. Briefly speaking, the model is a combination of the Poisson point process and projective geometry. We also conduct numerical simulations for the model, and the results show satisfactory scale invariance.

Chapter 8 is an extensive theoretical study of scale invariance. Motivated by the model established in Chapter 7, we develop a general mathematical approach to construct scale and translation invariant distributions on the space of functions defined on the whole integer lattice as well as on the space of generalized functions.

Contents

List of Figures

Chapter 1

Introduction

The center issue for many areas of machine learning and machine perception, including computer vision, speech recognition, language analysis, and medical expert systems, is how to extract useful information efficiently from signals produced by the world. The problem has been studied by two quite different approaches. In the first approach, for each class of signals, logic-like languages are devised to incorporate various knowledge about the world and the features of the signals. Inference about a signal is made by combining the features of the signal in a deterministic way offered by the languages to reach a high-level description of the signal. To combat the combinatorial explosion of the number of the possible combinations, various heuristics are developed for efficiently searching and pruning the trees of combinations.

The second approach is based on the general principle of the pattern theory pioneered by Grenander (Grenander [2] [3]). From the perspective of this theory, learning is equivalent to the accumulation from extensive experience the knowledge of the statistics of the signals and the entities represented in them. Perception is modeled as statistical estimation of random variables not directly observed from the data. Essential to this estimation problem is the Bayesian approach, which combines learned priors on the random variables and a model on the signals (Mumford [4]). The work described in this thesis mainly concerns the problem of building models for different classes of signals, which form different complex systems. Some part of the work also studies the learning of priors.

Now that models are needed for the pattern theory approach, to which extent is modeling important? Is it possible that without any specification for the models, a good model for any given class of signals can be learned from examples? This is in essence a non-parametric estimation problem. In theory, at least for simple problems, this can be done. One of the ideas for doing this is to associate with each model a complexity measure. Given a set of examples, or data, the "learned", or estimated model is the one which among all the possible models achieves the minimum sum of the complexity and the error made by fitting the data. It will be shown (Chapter 2) that the "minimum complexity" estimation is consistent. Naturally, in order to get stronger consistency of the estimation, the complexity measures should be more related to the actual behaviors of the models.

Conceptually appealing as it is, in reality, the minimum complexity estimation method, and indeed all current non-parametric estimation methods, are inadequate for difficult problems in machine perception. The inadequacy is illustrated by the bias-variance dilemma (Geman, et al [6]). On the one hand, because virtually infinite number of parameters are involved, non-parametric estimation methods produce estimates with high variance, hence for difficult perception problems, requiring prohibitively large training sets to reduce the variance contribution to the estimation error. On the other hand, the effort to control the variance in complex inference problems by using model-based estimation usually brings high bias, because proper models are hard to find for such problems, and estimation based on incorrect models can be highly biased.

The dilemma is due to the requirement for generality of the learning. This requirement, however, seems far from being met by the human brain. The human brain is a highly wired machine each part of which performs a specific function. Given the complexity of the input to the brain, the substantial human perception abilities, and the fact that a human being can obtain such abilities so quickly after being born, it is overwhelmingly more reasonable to regard the human brain as a model-based machine than a non-parametric, or "universal" one. This suggests a way to circumvent the bias-variance dilemma. That is, to give up generality and purposefully introduce bias. Putting into details, for each class of complex inference problems, specific machines must be devised with important properties of *the* class of problems being built into the machines' architecture. To introduce beneficial bias toward the problem at hand, the following three aspects are crucial: (1) the selection of the properties to build in, (2) the representation of the properties in the machines' architecture, and (3) the mechanism to tune the properties and the representation. Furthermore, the random nature of signals determines that the properties wired into the machines' architecture must include the description of the randomness of the signals. In short, probabilistic modeling is fundamental to machine learning and perception.

Although there are so many different classes of signals, each of which having properties remarkably different from the others, are there in any case general principles for probabilistic modeling for them? In spite of their differences, many classes of signals coming from the real world — natural images, languages, speech signals, in particular — share a common characteristic. That is, each signal is not a random collection of basic units — pixels, phonemes, and so on — but composition of these units in some specific manner. Human beings seem to exploit the compositional nature of signals in an efficient way by possessing the evident ability to represent in their minds entities as hierarchies of parts, with these parts themselves being meaningful entities, and being reusable in a near-infinite assortment of meaningful combination. As a way to mimic human perception, we therefore propose compositionality as a principle for probabilistic modeling (Chapter 3).

There are various ways to address compositionality which lead to different models. These models conveniently take the forms of grammars. Among the simplest grammars are context-free grammars (CFGs). Their probabilistic versions, called probabilistic context-free grammars (PCFGs), are equivalent to stochastic branching processes. Chapters 4 and 5 will cover PCFGs in detail. Standard grammars are feature based grammars, and their probabilistic versions are equipped with Gibbs distributions, with the potential function of each Gibbs distribution being a weighted sum of "features" that are supposed to capture the most important properties of images or languages (Chapter 6).

The probability model that so far most expressively incorporates compositionality and probability distribution is, in our opinion, the probabilistic composition system model proposed in Chapter 3. It explains compositionality from the point of view of the minimum description length (MDL) principle. This principle states that a “good” understanding of the signals is achieved when a compact description of the signals is gotten. By “compact” we mean the average length of the descriptions of the signals is small. The more compact the description is, the better the understanding. The underlying philosophy for the principle is that by capturing the mechanism to produce the signals, or the “core” of the signals, the description can greatly reduce its redundancy, hence achieving compactness. According to the information theory, the optimal description length of a signal is equivalent to the probability of the signal. By representing the probability of an object as the product of the *conditional* adjoint probability of its components, given its own features, and the prior of the features, the model assigns more probability to a single composite object than to the collection of the object’s components, taken as independent from each other. In this way, the model expresses the idea that composition produces more compact, hence better description, for entities.

So far we have discussed how to model properties of signals in a hierarchical way. The modeling is mainly for high-level understanding of signals. There is another aspect of probabilistic modeling, which is the modeling of the basic statistical properties of the “raw data”, i.e., the signals before they are combined to form higher-level entities. Studying the statistical properties of the raw data not only benefits modeling for higher-level machine perception, but also, as is equally important, helps to better understand why the structures of human sensors are as they are. As the probabilistic modeling for raw data does not need to address higher-level understanding, it is considered relatively more approachable.

We will study a distinguished property of visual signals, which is scale invariance of many statistics of natural images. In Chapter 7, we will build a model on the origin of scale invariance. This model is of physics flavor and can be generalized to a mathematical approach to construct scale and translation invariant distributions (Chapter 8).

Bibliography

[1] S. Geman, E. Bienenstock, and R. Doursat. Neural Networks and the Bias/Variance Dilemma. *Neural Computation* 4, 1-58. 1992.

[2] U. Grenander. *Lectures in Patter Theory.* Springer-Verlag. 1976.

[3] U. Grenander. *General Patter Theory.* Oxford University Press. 1994.

[4] D. B. Mumford. The Statistical Description of Visual Signals. K. Kirchgassner, O. Mahrenholtz, and R. Mennicken (editors). *ICIAM 95.* Akademie Verlag. 1995.

Chapter 2

On the Consistency of Minimum Complexity Nonparametric Estimation

Nonparametric estimation is usually inconsistent without some form of regularization. One way to impose regularity is through a prior measure. Barron and Cover [1][2], have shown that complexity-based prior measures can insure consistency, at least when restricted to countable dense subsets of the infinite-dimensional parameter (i.e. function) space. Strangely, however, these results are independent of the actual complexity assignment: the same results hold under an arbitrary permutation of the match up of complexities to functions. We will show that this phenomenon is related to the weakness of the convergence measures used. Stronger convergence can only be achieved through complexity measures that relate to the actual behavior of the functions.

2.1 Introduction

Maximum likelihood, least squares, and other estimation techniques are generally inconsistent for nonparametric (infinite-dimensional) problems. Some variety of regularization is needed. An appealing and principled approach is to base regularization on complexity: Define an encoding of the (infinite-dimensional) parameter, and adopt code length as a penalty. Barron and Cover ([1], [2]) have shown how to make this work. They get consistent estimation for densities and regressions, as well as some convergence-rate bounds, by constructing complexity-based penalty terms for maximum-likelihood and least-squares estimators.

Can we cite the results of Barron and Cover as an argument for complexity-based regularization (or, equivalently, for complexity-based priors)? Apparently not: The results are independent of the particular assignment of complexities. Specifically, the results are unchanged by an arbitrary permutation of the matching of complexities to parameters.

[1] Supported by the Army Research Office (DAAL03-92-G-0115), the National Science Foundation (DMS-9217655), and the Office of Naval Research (N00014-96-1-0647).

Of course there are many ways to define convergence of functions. We will show here that the surprising indifference of convergence results to complexity assignments is in fact related to the convergence measures used. Stronger convergence requires a stronger tie between the parameters (functions) and their complexity measures.

§2.2 is a review of some Barron and Cover results. Then some new results about consistency for nonparametric regression are presented in §2.3. (Proofs are in the Appendix.) Taken together, the results of §2.3 establish the principle that stronger types of convergence are sensitive to the particulars of the complexity assignment. We work here with regression, but the situation is analogous in density estimation.

Our results are about consistency only. The important practical issue of relating complexity measures to *rates* of convergence remains open.

2.2 Complexity-based Priors

Barron and Cover [1] have shown that the problem of estimating a density nonparametrically can be solved using a complexity-based prior by limiting the prior to a countably-dense subset of the space of densities. More specifically, given a sequence of countable sets of densities, Γ_n, and numbers $L_n(q)$ for densities q in Γ_n, let $\Gamma = \cup_n \Gamma_n$. Set $L_n(q) = \infty$ for q not in Γ_n. For independent random variables $X_1, X_2, \cdots, X_n$ drawn from an unknown probability density function p, a minimum complexity density estimator $\hat{p}_n$ is defined as a density achieving the following minimization

$$\min_{q \in \Gamma_n} \left(L_n(q) - \sum_{i=1}^{n} \log q(X_i) \right).$$

If we think of $L_n(q)$ as the description length of the density q, then the minimization is over total description length—accounting for both the density and the data. Barron and Cover showed that if L_n satisfies the summability condition

$$\sup_n \sum_{q \in \Gamma_n} 2^{-L_n(q)} < +\infty$$

and the growth restriction

$$\limsup_n \frac{L_n(q)}{n} = 0, \quad \text{for every } q \in \Gamma, \tag{2.1}$$

then for each measurable set S,

$$\lim_{n \to \infty} \hat{P}_n(S) = P(S) \quad \text{with probability one,}$$

provided that p is in the information closure $\bar{\Gamma}$ of Γ. Here, $\hat{P}_n$ and P are the probability measures associated with the densities $\hat{p}_n$ and p, respectively, and "p is in the information closure $\bar{\Gamma}$ of Γ" means that $\inf_{q \in \Gamma} D(p \| q) = 0$, where $D(p\|q)$ is the relative entropy of p to q.

Barron and Cover also showed that if L_n satisfies a "light tail condition," i.e. if for some $0 < \alpha < 1$ and b,

$$\sum_{q \in \Gamma_n} 2^{-\alpha L_n(q)} \le b, \qquad \text{for all } n, \tag{2.2}$$

and if L_n also satisfies the growth restriction (2.1), then for $p \in \bar{\Gamma}$, with probability one,

$$\lim_{n\to\infty} \int |p - \hat{p}_n| = 0.$$

A second paper by Barron [2] offers a minimum-complexity solution to the regression problem. Let $(X_i, Y_i)_{i=1}^n$ be independent observations drawn from the unknown joint distribution of random variables X, Y, where the support of X is in $\mathbf{R}^d$. Here X is the vector of explanatory variables and Y is the response variable. Functions $f(X)$ are used to predict the response. The error incurred by a prediction is measured by a distortion function $d(Y, f(X))$, the most common form being $(Y - f(X))^2$. Let h be a function which minimizes $E(d(Y, f(X)))$, which is to say that $h = E(Y|X = x)$ in the squared error case. When a function f is used in place of the optimum function h the "regret" is measured by the difference between the expected distortions

$$r(f, h) = E(d(Y, f(X))) - E(d(Y, h(X))).$$

Barron defines statistical risk for a given estimator $\hat{h}_n$ to be $E(r(\hat{h}_n, h))$. Given a sequence of countable collections of functions, Γ_n, and numbers $L_n(f)$, $f \in \Gamma_n$, satisfying the summability condition

$$\sup_n \sum_{f \in \Gamma_n} 2^{-L_n(f)} < \infty,$$

the index of resolvability is defined as

$$R_n(h) = \min_{f \in \Gamma_n} (r(f, h) + \lambda \frac{1}{n} L_n(f))$$

and a minimum complexity estimator is a function $\hat{h}_n \in \Gamma_n$ which achieves

$$\min_{f \in \Gamma_n} (\frac{1}{n} \sum_{i=1}^n d(Y_i, f(X_i)) + \lambda \frac{1}{n} L_n(f)).$$

Again there is a coding interpretation: if $d(Y, f(X))$ is log probability of Y given X, then $\hat{h}_n$ minimizes total description length for the model, f, plus the data $Y_1, ...Y_n$ given $X_1, ...X_n$. Barron showed that if the support of Y and the range of each function $f(X)$ is in a known interval of length b, then with $\lambda \geq 5b^2/3 \log e$, the mean squared error converges to zero at rate bounded by $R_n(h)$, i.e.,

$$E(r(\hat{h}_n, h)) \leq O(R_n(h)). \tag{2.3}$$

Taken together, these results offer a general prescription for nonparametric estimation of densities and regressions. Furthermore, the connection to complexity is appealing: It is not hard to invent suitable functions $L_n(\cdot)$ by counting the bits involved in a natural encoding of Γ_n (cf. [1]). There is, however, a disturbing indifference of the results to the details of the complexity measure. For any set of permutations σ_n on Γ_n, define $L'_n(\xi) = L_n(\sigma(\xi))$ and observe that L'_n satisfies whatever conditions L_n does, and hence the same results are obtained (with the same bound on rate in (2.3)) using L'_n in place of L_n! In general L'_n will have no meaningful interpretation as a complexity measure.

2.3 What Ties Consistency to Complexity?

Suppose that X is a random variable from a probability space $(\Omega, \mathcal{F}, P)$ to $([0,1], \mathcal{B})$. X introduces a measure P_X on $[0,1]$ through the relation $P_X(B) = P(X^{-1}(B))$, for $B \in \mathcal{B}$. Choose a countable dense subset Γ in $L^2([0,1], P_X)$, and define a "complexity function" $L : \Gamma \to \mathbf{N}$. For any random variable Y from $(\Omega, \mathcal{F}, P)$ to $(R, \mathcal{B})$ with $h(x) = E(Y|X = x) \in L^2([0,1], P_X)$, define the estimator $\hat{h}_n$ to be a function in Γ which achieves

$$\min_{f \in \Gamma} \left\{ \frac{L(f)}{n} + \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 \right\}.$$

We will always assume that L satisfies a much stronger tail condition than (2.2):

$$\sum_{f \in \Gamma} e^{-\epsilon L(f)} < \infty \qquad \text{for any } \epsilon > 0. \tag{2.4}$$

The first proposition demonstrates that for a weak form of convergence consistency is essentially independent of the complexity measure:

Proposition 1. If $EY^4 < \infty$, then

$$\hat{h}_n \xrightarrow{P_X} h, \qquad \text{a.s.}$$

Obviously, the proposition remains true for any permutation σ of Γ and resulting complexity function $L'(f) = L(\sigma(f))$. But, suppose we were to ask for consistency in L^2 (a.s.) in place of consistency in probability (a.s.)? Then, despite the strength of the tail condition (2.4), we would evidently need to pay closer attention to the complexity measure:

Proposition 2. There exists a random variable X, a countable dense subset Γ in $L^2([0,1], P_X)$, and a function $L : \Gamma \to \mathbf{N}$ satisfying (2.4) such that for any Y with $h(x) \notin \Gamma$, the L^2 norm of $\hat{h}_n$ (in $L^2([0,1], P_X)$) goes to $+\infty$ with probability one.

(We are focusing on the regression problem, but analogous arguments apply to probability density estimation. For example, by a construction similar to the one used for Proposition 2, the minimum complexity density estimator discussed in Barron and Cover [1] may not converge to the actual density p in the sense of Kullback-Liebler:

$$\int p \log \frac{p}{\hat{p}_n} \not\to 0,$$

even though the coding L satisfies the strong condition (2.4).)

One way to rescue consistency is to tie the complexity measure $L(f)$ more closely to f:

Proposition 3. Suppose that for every $f \in \Gamma$, $Ef^4(X) < \infty$. Assume EY^4 is finite (and hence so is $Eh^4(X)$). Construct a complexity function as follows: First, define $C_1(f) = (Ef^4(X) + e)e^{2Ef^2(X)}$ and $C(f) = C_1(f) \log C_1(f)$. Then, given any $L_1 : \Gamma \to N$ which satisfies (2.4), let $L(f) = C(f)L_1(f)$. Then

$$\hat{h}_n \xrightarrow{L^2} h \quad \text{a.s.}$$

Proofs for the propositions are in the Appendix.

Bibliography

[1] A. R. Barron, T. M. Cover. Minimum Complexity Density Estimation. *IEEE Trans. Inform. Theory*, vol. 37, pp. 1034-1054. July, 1991.

[2] A. R. Barron. Complexity Regularization with Application to Artificial Neural Networks. *Non-parametric Functional Estimation and Related Topics.* G. Roussas, Kluwer Academic Publishers.

Appendix

Recall that X is a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$, taking values in $([0,1], \mathcal{B})$. P_X is defined on $[0,1]$ by $P_X(B) = P(X^{-1}(B))$, for $B \in \mathcal{B}$. Γ is then a countable dense subset or $L^2([0,1], P_X)$. (Take, for example, Γ to be a countable dense set in $L^2([0,1], dx)$; this will work for any P_X which is absolutely continuous with respect to Lebesgue measure and has bounded derivative dP_X/dx.) The complexity function $L : \Gamma \to \mathbf{N}$ is always assumed to satisfy the "strong tail condition" (2.4).[1] Finally, we assume that the response variable Y (a random variable on $(\Omega, \mathcal{F}, P)$) has an L^2-valued regression

$$h(x) = E(Y|X = x) \in L^2([0,1], P_X).$$

The regression $h(x)$ is estimated by a function $\hat{h}_n \in \Gamma$ that achieves the minimum in

$$\min_{f \in \Gamma} \left\{ \frac{L(f)}{n} + \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 \right\}.$$

We begin with Proposition Proposition 2.

Proposition Proposition 2 There exists a random variable X, a countable dense subset Γ in $L^2([0,1],\ P_X)$, and a function $L : \Gamma \to \mathbf{N}$ satisfying (2.4) such that for any Y with $h(x) \notin \Gamma$, the L^2 norm of $\hat{h}_n$ (in $L^2([0,1], P_X)$) goes to $+\infty$ with probability one.

Proof. Choose X so that P_X is Lebesgue measure. Fix $\Gamma = \{f_1, \ldots, f_n, \ldots\}$ dense in $L^2([0,1],\ P_X)$. Let $B_1, \ldots, B_n,\ \ldots$ be a sequence of measurable subsets in $[0,1]$, each of which has positive probability, such that

$$P(\exists 1 \leq i \leq n,\ X_i \in B_n,\ \text{ i.o. for } n) = 0.$$

[1] For example: choose $a(\cdot)$ strictly positive such that $\sum_f a(f) < \infty$. If $F(x)$ is any strictly positive function satisfying $F(x)/x \to \infty$ as $x \to \infty$, then $L(f) = F(-\log a(f))$ satisfies (2.4).

This condition can be achieved, for instance, if the B's satisfy

$$\sum_{k=1}^{\infty}[1-(1-P_X(B_k))^k] < \infty.$$

Now for $i = 1, 2, \ldots$, define $g_i(x)$ as

$$g_i(x) = \begin{cases} f_i(x) & \text{if } x \notin B_i \\ A_i & \text{if } x \in B_i. \end{cases}$$

We first select A_1 such that $E(g_1 - f_n)^2 > 0$ for all $n \in \mathbf{N}$. This can be done since there are only countably many f's while there are uncountably many choices of A_1. We then inductively select A_i such that $E(g_i - f_n)^2 > 0$, for all $n \in \mathbf{N}$, and $E(g_i - g_k)^2 > 0$, for $k = 1, \ldots, i-1$. We also require of A_i that $Eg_i^2 \to +\infty$. Then $g_1, g_2, \ldots$ are distinct and none of them are in Γ. Modify Γ to include $g_1, g_2, \ldots$. Define $L : \Gamma \to N$ such that

$$L(f_n) > L(g_n)$$

and

$$\sum_{f \in \Gamma} e^{-\epsilon L(f)} < \infty, \qquad \text{for any } \epsilon > 0.$$

Now given Y, with $h(x) = E(Y|X = x) \in L^2([0,1], P_X)$ and $h(x) \notin \Gamma$, the set of ω which satisfies

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 \to E(h(X) - f(X))^2 + E(Y - h(X))^2,\ \forall f \in \Gamma$$

and

$$X_i(\omega) \notin B_n, \quad \forall 1 \le i \le n, \forall \text{ large } n$$

is of probability one. For any ω in this set, let

$$I_n(\omega) = \arg\min_k \left\{ \frac{L(f_k)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_k(X_i))^2 \right\}.$$

Then since $h \notin \Gamma$, $I_n(\omega) \to \infty$ as $n \to \infty$. For large n, $X_i(\omega) \notin B_{I_n(\omega)}$ for all $1 \le i \le I_n(\omega)$, and hence

$$g_{I_n(\omega)}(X_i(\omega)) = f_{I_n(\omega)}(X_i(\omega)) \ \ \forall 1 \le i \le I_n(\omega).$$

Therefore, for large n,

$$\frac{L(g_{I_n})}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - g_{I_n}(X_i))^2 < \frac{L(f_{I_n})}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{I_n}(X_i))^2.$$

Consequently, with probability one, for large n

$$\hat{h}_n = \arg\min_{f \in \Gamma} \left\{ \frac{L(f)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 \right\} \in \{g_1, g_2, \ldots\}.$$

Since $E(g_i^2) \to \infty$, this completes the proof. □

Remark 1. As mentioned in §3, the same argument can be used to show that the minimum complexity estimator $\hat{p}_n$ in [1] may not converge to the true density p, in the sense that

$$\int p \log \frac{p}{\hat{p}_n} \not\to 0.$$

The proof of Proposition Proposition 1 is based on the following three lemmas.

Lemma 1. Fix $\epsilon > 0$. Let $Z_1, Z_2, \dots, Z_n$ be a sequence of i.i.d. random variables satisfying
a. $Z_1 \geq 0$;
b. $EZ_1^2 < \infty$;

$$K \geq (\mathrm{Var}(Z_1) + \epsilon^2)e^{EZ_1} \quad \text{and} \quad \frac{\epsilon}{K} < 1,$$

then

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - EZ_1) \leq -\epsilon\right) \leq \left(1 - \frac{\epsilon^2}{2K}\right)^n$$

Proof. For any $t \in (0, 1]$,

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - EZ_1) \leq -\epsilon\right) \leq \left(Ee^{t(-Z_1+EZ_1-\epsilon)}\right)^n.$$

Let $\phi(t) = Ee^{t(-Z_1+EZ_1-\epsilon)}$, then

$$\phi(0) = 1, \quad \phi'(0) = -\epsilon$$

and for any $t \in (0, 1]$,

$$\phi''(t) = E((Z_1 - EZ_1 + \epsilon)^2 e^{t(-Z_1+EZ_1-\epsilon)}) \leq E(Z_1 - EZ_1 + \epsilon)^2 e^{tEZ_1} \leq K.$$

Hence

$$\phi'(t) \leq -\epsilon + Kt \quad \text{for } t \in (0, 1],$$

and

$$\phi(t) \leq 1 - \epsilon t + \frac{1}{2}Kt^2 \quad \text{for } t \in (0, 1].$$

Take $t = \epsilon/K < 1$, which is the minimizer of $1 - \epsilon t + Kt^2/2$. Then

$$P\left(\frac{1}{n}\sum_{i=1}^{n} Z_i < EZ_1 - \epsilon\right) \leq \left(1 - \frac{\epsilon^2}{2K}\right)^n.$$

□

Lemma 2. Suppose $EY^4 < \infty$. Let $h(x) = E(Y|X = x) \in L^2([0,1], P_X)$. Assume Γ is a countable dense subset of

$$\{f \in L^2([0,1], P_X) : \ |f(x)| \leq M\}$$

and $L : \Gamma \to \mathbf{N}$ satisfies condition (2.4). Then given $0 < \epsilon < 1$, with probability one, for sufficiently large n and all $f \in \Gamma$ with $E(f - h_M)^2 \geq 3\epsilon$,

$$\frac{1}{n}\sum_{i=1}^{n}(h_M(X_i) - Y_i)^2 + \epsilon < \frac{L(f)}{n} + \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2, \tag{A2.1}$$

where for any function f,

$$f_M(x) = \begin{cases} f(x) & \text{if } |f(x)| \leq M \\ \text{sign}(f(x)) \cdot M & \text{otherwise.} \end{cases} \tag{A2.2}$$

Proof. We shall first give the idea of the proof. Assume $|h| < M$. With probability one, when n is sufficiently large, $\sum_{i=1}^{n}(h(X_i) - Y_i)^2/n + \epsilon$ is bounded by $E(h(X) - Y)^2 + 2\epsilon$. We then get a stronger inequality

$$E(h(X) - Y)^2 + 2\epsilon \leq \frac{L(f)}{n} + \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2.$$

The left hand side equals

$$E(f(X) - Y)^2 - E(f(X) - h(X))^2 + 2\epsilon \leq E(f(X) - Y)^2 - \epsilon.$$

Hence we can prove the lemma by showing

$$\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2 - E(f(X) - Y)^2 > -\epsilon - \frac{L(f)}{n}$$

is true with probability one, for sufficiently large n and all $f \in \Gamma$. By Lemma 1, for each fixed n and $f \in \Gamma$, the probability that this inequality does not hold is bounded by

$$\left(1 - \frac{(\epsilon + L(f)/n)}{K}\right)^n \leq \left(1 - \frac{\epsilon^2}{K}\right)^n \left(1 - \frac{\epsilon L(f)/n}{K}\right)^n,$$

where K is a large number independent of n. Because $1 - x < e^{-x}$ for all $x > 0$, the above probability is then bounded by

$$\left(1 - \frac{\epsilon^2}{K}\right)^n e^{-\epsilon L(f)/K}.$$

Sum over all $f \in \Gamma$, we see that the probability that (A2.1) is not true is exponentially small. A Borel-Cantelli argument then finishes the proof.

We turn now to the details of the proof. Define

$$B(h_M) = \{f \in \Gamma : \ E(f - h_M)^2 \geq 3\epsilon\}. \tag{A2.3}$$

For $f \in \Gamma$, define

$$\begin{aligned} T_{f,n}(h_M) &= \left\{ \frac{1}{n}\sum_{i=1}^{n}(f(X_i)-Y_i)^2 + \frac{L(f)}{n} \leq \frac{1}{n}\sum_{i=1}^{n}(h_M(X_i)-Y_i)^2 + \epsilon \right\}, && \text{(A2.4)} \\ V_n(h_M) &= \bigcup_{f \in B} T_{f,n}. && \text{(A2.5)} \end{aligned}$$

Write

$$\begin{aligned} R_n(h_M) &= \left\{ \left| \frac{1}{n}\sum_{i=1}^{n}(h_M(X_i)-Y_i)^2 - E(h_M(X)-Y)^2 \right| < \epsilon \right\}, && \text{(A2.6)} \\ R(h_M) &= \liminf_{n\to\infty} R_n(h_M). && \text{(A2.7)} \end{aligned}$$

Henceforth, we will simplify the notation by writing B instead of $B(h_M)$, $T_{f,n}$ instead of $T_{f,n}(h_M)$, and so on. By the strong law of large numbers, $P(R) = 1$. Next show that $\sum_n P(V_n \cap R_n) < \infty$. If this is true, then by the Borel-Cantelli lemma,

$$P(\limsup_{n\to\infty} V_n) = P(\limsup_{n\to\infty} V_n \cap R) \leq P(\limsup_{n\to\infty}(V_n \cap R_n)) = 0,$$

which is what needs to be proved.

For $\omega \in R_n$ and $f \in B$,

$$\frac{1}{n}\sum_{i=1}^{n}(h_M(X_i)-Y_i)^2 + \epsilon - E(f(X)-Y)^2 \leq 2\epsilon + E(h_M(X)-Y)^2 - E(f(X)-Y)^2.$$

Clearly,

$$E(Y-f(X))^2 = E(Y-h(X))^2 + E(h(X)-f(X))^2.$$

Since $|f| \leq M$, $|h-f| = |h-h_M| + |h_M - f|$,

$$\begin{aligned} E(Y-f(X))^2 &\geq E(Y-h(X))^2 + E(h(X)-h_M(X))^2 + E(h_M(X)-f(X))^2 \\ &= E(Y-h_M(X))^2 + E(h_M(X)-f(X))^2 \\ &\geq E(Y-h_M(X))^2 + 3\epsilon. \end{aligned}$$

Hence

$$\frac{1}{n}\sum_{i=1}^{n}(h_M(X_i)-Y_i)^2 + \epsilon - E(f(X)-Y)^2 \leq -\epsilon.$$

Suppose $f \in B$ and $R_n \cap T_{f,n} \neq \emptyset$. For any $\omega \in R_n \cap T_{f,n}$, by the above inequality,

$$\frac{1}{n}\sum_{i=1}^{n}(f(X_i)-Y_i)^2 - E(f(X)-Y)^2 \leq -\epsilon - \frac{L(f)}{n} = -\delta_{f,n}.$$

Furthermore,

$$\frac{L(f)}{n} \leq \frac{1}{n}\sum_{i=1}^{n}(h_M(X_i)-Y_i)^2 + \epsilon$$

and hence

$$\begin{aligned}\delta_{f,n} &\leq 2\epsilon + \frac{1}{n}\sum_{i=1}^{n}(h_M(X_i) - Y_i)^2 \\ &\leq 3 + E(h_M(X) - Y)^2 = H. \qquad \text{(A2.8)}\end{aligned}$$

Fix K such that

$$K \geq (E(M + |Y|)^4 + H^2)e^{E(M+|Y|)^2}.$$

Now for any $f \in B$ with $R_n \cap T_{f,n} \neq \emptyset$, it is easy to check

$$(\text{Var}((f(X) - Y)^2) + \delta_{f,n}^2)e^{E(f(X)-Y)^2} \leq K \quad \text{and} \quad \delta_{f,n} < K.$$

Then by Lemma 1, for any $f \in B$ with $R_n \cap T_{f,n} \neq \emptyset$,

$$\begin{aligned}P(R_n \cap T_{f,n}) &\leq P\left(\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2 - E(f(X) - Y)^2 \leq -\delta_{f,n}\right) \\ &\leq \left(1 - \frac{(L(f)/n + \epsilon)^2}{2K}\right)^n \\ &\leq \left(1 - \frac{\epsilon^2}{2K}\right)^n \left(1 - \frac{\epsilon L(f)/n}{K}\right)^n\end{aligned}$$

Since

$$\frac{\epsilon L(f)/n}{K} < \frac{\epsilon \delta_{f,n}}{K} < 1,$$

and $1 - x < e^{-x}$, for all $0 < x < 1$, we get $P(R_n \cap T_{f,n})$ is bounded by

$$\left(1 - \frac{\epsilon^2}{2K}\right)^n \exp\left(-\frac{\epsilon L(f)}{K}\right).$$

Therefore

$$P(R_n \cap V_n) \leq \sum_{f \in B} P(R_n \cap T_{f,n}) \leq \left(1 - \frac{\epsilon^2}{2K}\right)^n \sum_{f \in \mathcal{A}} \exp\left(-\frac{\epsilon L(f)}{K}\right),$$

and by the strong tail condition (2.4), $\sum \exp(-\epsilon L(f)/K) < \infty$. Since K is independent of n, $P(R_n \cap V_n)$ is exponentially small and $\sum P(R_n \cap V_n)$ converges. □

Lemma 3. Let μ be a finite measure, and let f and f_n, $n = 1, 2, ...$, be measurable functions. If $f < \infty$, μ-*a.s*, and if

$$\liminf_{M\to\infty} \limsup_{n\to\infty} E(f_{n,M} - f_M)^2 = 0,$$

then $f_n \xrightarrow{\mu} f$.

Proof. Suppose $M_n \to \infty$ is a sequence such that

$$\lim_{k\to\infty} \limsup_{n\to\infty} E(f_{n,M_k} - f_{M_k})^2 = 0.$$

Fix $\epsilon > 0$ and $M > 0$. Then

$$\begin{aligned} \mu(\{|f_n - f| > \epsilon\}) &\leq \mu(\{|f| \geq M_k - \epsilon\}) + \mu(\{|f| < M_k - \epsilon, \, |f_{n,M_k} - f_{M_k}| > \epsilon\}) \\ &\leq \mu(\{|f| \geq M_k - \epsilon\}) + \frac{1}{\epsilon^2} E(f_{n,M_k} - f_{M_k})^2. \end{aligned}$$

Let $n \to \infty$ and then $k \to \infty$ to complete the proof. □

Proposition Proposition 1 If $EY^4 < \infty$, then

$$\hat{h}_n \xrightarrow{P_X} h, \qquad \text{a.s.}$$

Proof. The idea is to choose $M_k \to \infty$ and then truncate the functions in Γ as in (A2.2). Then by Lemma 2, we will get $E(\hat{h}_{n,M_k} - h_{M_k})^2 \to 0$, where h_{n,M_k} is the truncated h_n, and h_{M_k} is the truncated h. We then use Lemma 3 to get $\hat{h}_n \overset{P_X}{\to} h$.

Filling in the details, given $\epsilon > 0$, there is $M = M(\epsilon) > 0$ such that $E(h - h_M)^2 < \epsilon$ and

$$\int_{|Y|>M} (|Y| + M)^2 \leq 4 \int_{|Y|>M} |Y|^2 < \epsilon.$$

With probability one, when n is sufficiently large,

$$\frac{L(\hat{h}_n)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{h}_n(X_i))^2 < \frac{1}{n}\sum_{i=1}^{n}(Y_i - h_M(X_i))^2 + \epsilon.$$

Consider

$$\frac{L(\hat{h}_n)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{h}_{n,M}(X_i))^2.$$

Observe that $|Y_i - \hat{h}_{n,M}(X_i)| > |Y_i - \hat{h}_n(X_i)|$ implies $|Y_i| > M$. Hence

$$\begin{aligned} & \frac{L(\hat{h}_n)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{h}_{n,M}(X_i))^2 \\ \leq \quad & \frac{L(\hat{h}_n)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{h}_n(X_i))^2 + \frac{1}{n}\sum_{i=1}^{n}(|Y_i| + M)^2 \cdot I_{|Y_i|>M}. \end{aligned}$$

With probability one, for sufficiently large n,

$$\frac{1}{n}\sum_{i=1}^{n}(|Y_i| + M)^2 \cdot I_{|Y_i|>M} \leq \int_{|Y|>M} (|Y| + M)^2 + \epsilon < 2\epsilon,$$

and therefore for large n,

$$\frac{L(\hat{h}_n)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{h}_{n,M}(X_i))^2 \leq \frac{1}{n}\sum_{i=1}^{n}(Y_i - h_M(X_i))^2 + 3\epsilon.$$

Let $\Gamma_M = \{f_M : \ f \in \Gamma\} \cup \{h_M\}$, which is dense in $L^2([0,1], P_X) \cap \{\|f\|_\infty \leq M\}$. Define $L' : \Gamma_M \to \mathbf{N}$ as

$$L'(\pi) = \min\{L(f) : \ f_M = \pi, f \in \Gamma\}.$$

Then with probability one, for large n,

$$\frac{L'(\hat{h}_{n,M})}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{h}_{n,M}(X_i))^2 \leq \frac{1}{n}\sum_{i=1}^{n}(Y_i - h_M(X_i))^2 + 3\epsilon.$$

L' satisfies the strong tail condition (2.4). According to Lemma 2, with probability one, for sufficiently large n

$$E(\hat{h}_{n,M} - h_M)^2 \leq 9\epsilon.$$

Let $S(\epsilon)$ be the subset of points in Ω such that the above relation holds, i.e.

$$S(\epsilon) = \liminf_{n\to\infty} \left\{\omega : \ E(\hat{h}_{n,M} - h_M)^2 \leq 9\epsilon\right\}.$$

Choose a sequence $\epsilon_n \to 0$, and let $M_n = M(\epsilon_n)$ and $S_n = S(\epsilon_n)$. Then on $S = \cap S_n$, which has probability one,

$$\lim_{k\to\infty} \sup \lim_{m\to\infty} E(\hat{h}_{n,M_k} - h_{M_k})^2 = 0.$$

By lemma 3, for any $\omega \in S$, $\hat{h}_n \xrightarrow{P_X} h$, which completes the proof. □

Proposition Proposition 3 Suppose that for every $f \in \Gamma$, $Ef^4(X) < \infty$. Assume EY^4 is finite (and hence so is $Eh^4(X)$). Construct a complexity function as follows: First, define $C_1(f) = (Ef^4(X) + e)e^{2Ef^2(X)}$ and $C(f) = C_1(f)\log C_1(f)$. Then, given any $L_1 : \Gamma \to N$ which satisfies (2.4), let $L(f) = C(f)L_1(f)$. Then

$$\hat{h}_n \xrightarrow{L^2} h \quad \text{a.s.}$$

Proof. We will follow closely the proof and the notation of Lemma 2. As in Lemma 2, we need to show that $P(\limsup V_n) = 0$. Fixing a number $D = D(Y, h, \epsilon)$, which will be determined later, we first decompose V_n as

$$V_n = \bigcup_{f\in B} T_{f,n} = \bigcup_{f\in B, L_1(f)\geq D} T_{f,n} \cup \bigcup_{f\in B, L_1(f)<D} T_{f,n} = V_n' \cup V_n''.$$

Since there are only finitely many f with $L_1(f) < D$, by the strong law of large numbers,

$$P(\limsup V_n'') = 0.$$

Thus in order to get $P(\limsup V_n) = 0$, we need only show that $P(\limsup V_n') = 0$. Similar to Lemma 2, it is enough to check

$$\sum_n P(V_n' \cap R_n) < \infty.$$

Derive again the constant H, as in (A2.8). Then for each $f \in \Gamma$, define

$$K(f) = (\mathrm{Var}((f(X) - Y)^2) + H^2)e^{E(f(X)-Y)^2} > e.$$

Then for any $f \in B$ with $R_n \cap T_{f,n} \neq \emptyset$, as in the proof of Lemma 2,

$$P(R_n \cap T_{f,n}) \leq \left(1 - \frac{\epsilon^2}{2K(f)}\right)^n \exp\left(-\frac{\epsilon C(f) L_1(f)}{K(f)}\right).$$

Hence

$$\begin{aligned}\sum_{m=1}^{\infty} P(R_n \cap V_n') &\leq \sum_{L_1(f)\geq D} \sum_{m=1}^{\infty} P(R_n \cap T_{f,n}) \\ &\leq \sum_{L_1(f)\geq D} \frac{2K(f)}{\epsilon^2} \exp\left(-\frac{\epsilon C(f) L_1(f)}{K(f)}\right) \\ &= \frac{2}{\epsilon^2} \sum_{L_1(f)\geq D} \exp(L_1(f) J(f,\epsilon))\end{aligned}$$

where

$$J(f,\epsilon) = -\frac{\epsilon C(f)}{K(f)} + \frac{\log K(f)}{L_1(f)}.$$

It is easy to see that there is a constant $c = c(Y,h) > 0$, such that $C(f) \geq cK(f)\log K(f) > 0$. Now choose $D = D(Y,h,\epsilon)$ such that $\epsilon c D \geq 2$. Then for $L_1(f) \geq D$,

$$\frac{\log K(f)}{L_1(f)} \leq \frac{\epsilon C(f)}{2K(f)}.$$

Since $K(f) > e$,

$$J(f,\epsilon) \leq -\frac{\epsilon C(f)}{2K(f)} \leq -\frac{\epsilon C(f)}{2K(f)\log K(f)} \leq -\frac{\epsilon c}{2}.$$

So

$$\sum_{m=1}^{\infty} P(R_n \cap V_n') \leq \frac{2}{\epsilon^2} \sum_{f\in\Gamma} e^{-\epsilon c L_1(f)/2} < \infty.$$

Similar to Lemma 3, we can now conclude that for any $0 < \epsilon < 1$, the set

$$S(\epsilon) = \left\{\omega : \ E(\hat{h}_n - h)^2 < 3\epsilon, \text{ for sufficiently large } n\right\}$$

has probability one. Finally, then, for $\omega \in \cap_{k=1}^{\infty} S(k^{-1})$, $E(\hat{h}_n - h)^2 \to 0$ as $n \to \infty$. □

Chapter 3

Composition Systems

3.1 Introduction

Compositionality is a mechanism to represent entities in a hierarchical way. Each entity is composed of several parts, which themselves are meaningful entities. Each entity is also reusable in a near-infinite assortment of meaningful combinations to form other entities. Such hierarchical representation of meaningful entities is widely believed to be fundamental to language (Chomsky [2]) as well as to vision, or any other kind of cognition (Bienenstock [1]). On one hand, entities that convey information, such as sentences and scenes, decompose naturally into a hierarchy of meaningful and generic parts, with all the possible meanings of each part being examined. On the other hand, compositions of parts remove ambiguities, because interpretations of parts that do not fit the contextual constraints offered by the composition are removed from further consideration, making parts correctly interpreted at the top level of the hierarchy.

The fundamental importance of compositionality entails addressing the mechanism in a more principled way, and composition systems are devised for this purpose. A composition system includes four components: (1) a set of categories, or "labels", for the meaningful entities; (2) for each category, a set of parameters, or "attributes", that are used to describe entities falling into this category; (3) a set of constraints on compositions, or "composition rules"; and (4) a set of primitive entities, or "terminals", which can not be further decomposed, and which have definite interpretations and serve as the building blocks for other entities. Any entity that is built per the composition rules from the terminals is called an object generated by the composition system.

Even after a composition system is established, one still faces the following question: Why is it the case that the interpretation of a collection of objects as a single composite object, when possible, is generally favored over the interpretation of these same objects as independent entities? The answer is that the description length of a composite object is on average smaller than the total description length of its components. This answer clearly depends on how the objects are encoded, or, from the probability point of view, depends on the probability measure on objects. Any reasonable probability measure on objects generated by a composition system should of course address compositionality, which is the reason why it is called a "compositional probability measure". However, how the measure accommodates compositionality can be explained in many different ways which lead to different formalisms. The formalism that this chapter is devoted to is the one given by Geman, *et. al* [6].

The chapter proceeds as follows. In §3.2, we review the formalism of composition systems given in [6]. §§3.3-3.5 study probability distributions for discrete composition systems. For more general treatment, we refer the readers to [6].

3.2 Definitions and Notations

In this section, we collect the conventions and definitions for composition systems.

Convention. (***-Notation**) For a set S, we will use S^* to represent the set of finite *non-empty* strings of elements of S, i.e.,

$$S^* = \bigcup_{n=1}^{\infty} \{s_1 s_2 \cdots s_n : \ s_i \in S, i = 1, \ldots, n\}.$$

This is nonstandard — usually S^* includes the empty string. For any $\alpha^* \in S^*$, its length is defined as the total number of elements in the string and is denoted as $|\alpha^*|$,

If P is a measure on S, then P^* is a measure on S^*, such that for any (measurable) subset $C \subset S^*$,

$$P^*(C) = \sum_{n=1}^{\infty} P^n(C \cap S^n).$$

If f is a numerical function on S, then f^* is a numerical function on S^* such that for any $\alpha^* = \alpha_1 \cdots \alpha_n \in S^*$, $f^*(\alpha^*) = f(\alpha_1) \cdots f(\alpha_n)$. If g is a function on S which takes values in a general set V, then g^* is a function on S^* which takes values in the set V^*, such that $g^*(\alpha^*)$ is the string $g(\alpha_1) \cdots g(\alpha_n) \in V^*$. Without specification, a set is always assumed to be a general set, even if all its elements are numbers.

Definition 1. Given a *label set* N, *which is always assumed to be countable*, a *terminal set* T, the set of *labeled trees*, Θ, is the set of finite trees with nonterminal nodes labeled by elements of N and terminal (leaf) nodes labeled by elements of T.

Remark 2.

1. $T \subseteq \Theta$;
2. By the *label* of the tree $\omega \in \Theta$ we will mean the label of its root node. We use $\mathrm{L}(\omega)$ ($\mathrm{L} : \Theta \to T \cup N$) to represent the label of ω;
3. $\omega = l(\alpha^*)$, $\alpha^* = \alpha_1 \cdots \alpha_n$, means $\mathrm{L}(\omega) = l$ and the left-to-right daughter subtrees of ω are $\alpha_1, \ldots \alpha_n$;
4. The ordering of daughter nodes is distinguished. So, for example, $l(\alpha, \beta) \neq l(\beta, \alpha)$ unless $\alpha = \beta$;

5. For any $\omega \in \Theta$, define $|\omega|$ as the total number of nodes (including terminals) in ω and $h(\omega)$ as the height (including terminals) of ω;

6. The *yield* of any tree $\omega \in \Theta$, denoted $Y(\omega)$, is the left-to-right string of terminals of ω.

Definition 2. A *composition rule* for the label $l \in N$ is a pair $(B_l, \mathcal{S}_l)$ where B_l, the *binding function*, maps $\Theta^* = \cup_{n=1}^{\infty} \Theta^n$ into an arbitrary *range space*, $\mathcal{R}_l$:

$$B_l : \Theta^* \to \mathcal{R}_l,$$

and $\mathcal{S}_l$, the *binding support*, is a distinguished non-empty subset of $\mathcal{R}_l$, $\emptyset \neq \mathcal{S}_l \subseteq \mathcal{R}_l$. The triple

$$\mathcal{C} = (T, N, \{B_l, \mathcal{S}_l\}_{l \in N})$$

is called a *composition system.*

Remark 3.

1. The *attribute value* of any $\omega = l(\alpha^*) \in \Theta^*$ is the value of $B_l(\alpha^*)$ and is denoted as $A(\omega)$;

2. The *type* of any $\omega \in \Theta$, denoted $\mathrm{T}(\omega)$, is defined as as follows. If $\omega \in T$, then $\mathrm{T}(\omega)$ is ω itself. If $\omega \in \Theta_l$, then $\mathrm{T}(\omega)$ is the pair $(l, A(\omega))$.

3. For any type t, define Θ_t as the set $\{\omega \in \Theta : \ \mathrm{T}(\omega) = t\}$. If $t = (l, b)$, also write Θ_t as $\Theta_{l,b}$.

Definition 3. Given a composition system $\mathcal{C} = (T, N, \{B_l, \mathcal{S}_l\}_{l \in N}\})$, the set of *objects* Ω is the closure of T under $\{(B_l, \mathcal{S}_l)\}_{l \in N}$ in Θ. That is, $\omega \in \Theta$ is an object ($\omega \in \Omega$) if and only if either $\omega \in T$ or $\omega = l(\alpha^*)$, where $\alpha^* \in \Omega^*$ and $B_l(\alpha^*) \in \mathcal{S}_l$. The set of yields of all objects in Ω, i.e.,

$$\{Y(\omega) : \omega \in \Omega\},$$

is called the *language* generated by $\mathcal{C}$.

Remark 4.

1. $\mathcal{S}_l$ is required to be minimal. In other words, for any $b \in \mathcal{S}_l$, there is an $\omega \in \Omega$ such that $\mathrm{L}(\omega) = l$ and $A(\omega) = b$;

2. We use $\mathcal{T}$ to represent the set of all types of objects, i.e.

$$\mathcal{T} = T \cup \{(l, b) : \ \exists\ \omega \text{ with } \mathrm{L}(\omega) = l \text{ and } A(\omega) = b\}.$$

   Because $\mathcal{S}_l$, $l \in N$, are minimal,

$$\mathcal{T} = T \cup \{(l, b) : l \in N, \ b \in \mathcal{S}_l\}.$$

3. For any type t, define $\Omega_t = \Omega \cap \Theta_t$.

Definition 4. The *observable measures* are

1. Q, a probability measure on $T \cup N$ with its support being the whole $T \cup N$;
2. Q_l, a probability measure on $\mathcal{R}_l$ with support $\mathcal{S}_l$, for any $l \in N$.

Remark 5. Q and Q_l induce a probability measure on $\mathcal{T}$ which is identical to Q on T, and equals $Q(l)Q_l$ on $\mathcal{S}_l$ for each $l \in N$. The induced measure is still written as Q.

3.3 Compositional Probability Distribution and Its Existence

We only consider the case where T is countable. Because N is always countable, therefore Ω is also countable. Since by definition, $\mathcal{S}_l$ is minimal, then $\mathcal{S}_l$ must be countable. Because for any $l \in N$, the support of Q_l is $\mathcal{S}_l$, for each $b \in \mathcal{S}_l$, $Q_l(b) > 0$.

Definition 5. A *compositional probability measure P on Ω with observable probability measures Q and Q_l* is a probability measure such that

$$P(\omega) = \begin{cases} Q(\omega), & \text{for any } \omega \in T \\ Q(l)Q_l(b)\dfrac{P^*(\alpha^*)}{P^*\Big(\{\beta^* \in \Omega^* :\ B_l(\beta^*) = b\}\Big)}, & \text{for any } \omega = l(\alpha^*) \in \Omega_{l,b}. \end{cases} \tag{3.1}$$

For explanations of this formulation, see [6].

We now address the issue of existence of compositional probability distributions. Obviously, existence depends not only on the composition rules, but also on the observable measures Q and Q_l. However, we are more interested in results on existence which only depend on composition rules. Firstly, as the term "observable" suggests, Q and Q_l are determined by data and cannot be alternated artificially to accommodate the existence of solution for (3.1). Secondly, results only depending on composition rules are more informative about the structures of composition systems, hence offering more insight into the criteria for "good" composition systems.

Our basic result on existence is the following proposition.

Proposition 4. If for any $l \in N$ and any $b \in \mathcal{S}_l$,

$$\max\{h(\omega) :\ \omega \in \Omega_{l,b}\} < \infty \tag{3.2}$$

and

$$\max\{|\alpha^*| :\ l(\alpha^*) \in \Omega_{l,b}\} < \infty, \tag{3.3}$$

then for any observable probabilities Q and Q_l, there exists a compositional probability measure satisfying (3.1).

The proof of Proposition Proposition 4 is quite complicated. We put it in Appendix at the end of the chapter.

Suppose (3.1) has a solution P. For $l \in N$ and $b \in \mathcal{S}_l$, write

$$Z_{l,b} = \frac{Q(l)Q_l(b)}{P^*\Big(\{\beta^* \in \Omega^* :\ B_l(\beta^*) = b\}\Big)}. \tag{3.4}$$

Then for $t = (l, b) \in \mathcal{T}$ and $\omega = l(\alpha^*) \in \Omega_t$, (3.1) can be written as

$$P(\omega) = P^*(\alpha^*)Z_t,$$

Let $f(t;\omega)$ be the number of subtrees of ω with type t. By induction, it is easy to see that

$$P(\omega) = \prod_{t\in T} Q(t)^{f(t;\omega)} \prod_{t\in \mathcal{T}\backslash T} Z_t^{f(t;\omega)}.$$

Because $\prod_{t\in T} Q(t)^{f(t;\omega)} = Q^*(Y(\omega))$,

$$P(\omega) = Q^*(Y(\omega)) \prod_{t\in \mathcal{T}\backslash T} Z_t^{f(t;\omega)} = Q^*(Y(\omega))Z^{f(\omega)}, \tag{3.5}$$

where $Z = \{Z_t\}_{t\in\mathcal{T}\backslash T}$, $f(\omega) = \{f(t;\omega)\}_{t\in\mathcal{T}\backslash T}$ and $Z^{f(\omega)}$ is the product of all $Z_t^{f(t;\omega)}$. Because P is a compositional probability distribution, for any $t \in \mathcal{T}\backslash T$,

$$\sum_{\omega\in\Omega_t} Q^*(Y(\omega))Z^{f(\omega)} = \sum_{\omega\in\Omega_t} P(\omega) = Q(t).$$

Recall that for $t = (l, b)$, $Q(t) = Q(l)Q_l(t)$.

Therefore, we have proved that if (3.1) has a solution, then the equation system induced by the composition system with Z as the unknowns,

$$\sum_{\omega\in\Omega_t} Q^*(Y(\omega))Z^{f(\omega)} = Q(t), \quad \text{for all } t \in \mathcal{T}\backslash T, \tag{3.6}$$

has a solution given by (3.4). Conversely, if (3.6) has a solution Z, then P given by (3.5) satisfies (3.1). Therefore, the existence of solution for (3.1) is equivalent to the existence of solution for (3.6).

Based on Proposition Proposition 4, we can prove another result on existence without assuming (3.2).

Proposition 5. Assume the set $\mathcal{T}\backslash T$ is finite. Also assume for each $(l, b) \in \mathcal{T}$, (3.3) is satisfied. If for every $t \in \mathcal{T}\backslash T$, the domain of convergence of the series

$$\sum_{\omega\in\Omega_t} Q^*(Y(\omega))Z^{f(\omega)} \tag{3.7}$$

is open inside the region $Z > 0$, then there is a solution for (3.6).

Proof. Because of (3.3),

$$h(l,b) = |\{|\alpha^*| : \ l(\alpha^*) \in \Omega_{l,b} \ \}| < \infty.$$

For $n \in \mathbf{N}$, let $\Omega_n = \{\omega \in \Omega, \ h(\omega) \leq n\}$. Because $\mathcal{T}\backslash T = \{(l,b) : \Omega_{l,b} \neq \emptyset\}$ is finite, when n is large enough, Ω_n intersects with each Ω_t, $t \in \mathcal{T}$. For such Ω_n, both (3.2) and (3.3) are satisfied, i.e.,

$$\max\{h(\omega) : \ \omega \in \Omega_{l,b} \cap \Omega_n\} < \infty$$

and

$$\max\{|\alpha^*| : \ l(\alpha^*) \in \Omega_{l,b} \cap \Omega_n\} < \infty.$$

Then by Proposition Proposition 4, there is a compositional probability distribution P_n on Ω_n, such that for any $\omega = l(\alpha^*) \in \Omega_{l,b} \cap \Omega_n$,

$$P_n(\omega) = Q(l)Q_l(b)\frac{P_n^*(\alpha^*)}{P_n^*\Big(\{\beta^* \in \Omega_n^* : \ B_l(\beta^*) = b\}\Big)}.$$

Note that if $\omega = l(\alpha^*) \in \Omega_n$, then $\alpha^* \in \Omega_n^*$, and therefore $P_n^*(\alpha^*)$ in the above formula makes sense.

Define $Z_n = \{Z_{l,b,n}\}$ as in (3.4), i.e.,

$$Z_{l,b,n} = \frac{Q(l)Q_l(b)}{P_n^*\Big(\{\beta^* \in \Omega^* : \ B_l(\beta^*) = b\}\Big)}. \tag{3.8}$$

Then as in (3.6),

$$\sum_{\substack{\omega \in \Omega_{l,b} \\ h(\omega) \leq n}} Q^*(Y(\omega))Z_n^{f(\omega)} = Q(l)Q_l(b).$$

Since for each $(l,b) \in \mathcal{T}\backslash T$,

$$\frac{Q(l)Q_l(b)}{h(l,b)} \leq Z_{l,b,n} = \frac{Q(l)Q_l(b)}{\sum\limits_{\substack{B_l(\beta^*)=b \\ l(\beta^*) \in \Omega_n}} P_n^*(\beta^*)} \leq \frac{Q(l)Q_l(b)}{\sum\limits_{\substack{B_l(\beta^*)=b \\ l(\beta^*) \in \Omega_n}} D^*(\beta^*)},$$

Z_n are bounded. The definition of D is given by (A3.3) in Appendix.

Because $\mathcal{T}\backslash T$ is finite, there is a subsequence Z_{n_i} of Z_n which is uniformly convergent to, say, $\xi = \{\xi_{l,b}\}$. Given any $\epsilon = \{\epsilon_{l,b}\}$, with $0 < \epsilon_{l,b} < \xi_{l,b}$, for large enough i, $Z_{n_i} > \xi - \epsilon$, that is, for each (l,b), $Z_{l,b,n_i} > \xi_{l,b} - \epsilon_{l,b}$ Therefore

$$\sum_{\substack{\omega \in \Omega_{l,b} \\ h(\omega) \leq n_i}} Q^*(Y(\omega))(\xi - \epsilon)^{f(\omega)} \leq Q(l)Q_l(b).$$

Letting $i \to \infty$ and then $\epsilon \to 0$, we get

$$\sum_{\omega \in \Omega_{l,b}} Q^*(Y(\omega))\xi^{f(\omega)} \le Q(l)Q_l(b).$$

By the assumption that the domain of convergence of the series of (3.7) is open for each $t \in \mathcal{T} \backslash T$, for any $\beta = \{\beta_{l,b}\}$ with $\beta_{l,b} > 0$ being small enough,

$$\sum_{\omega \in \Omega_{l,b}} Q^*(Y(\omega))(\xi + \beta)^{f(\omega)} < \infty.$$

When i is large enough, $Z_{n_i} \le \xi + \beta$. Therefore,

$$\sum_{\omega \in \Omega_{l,b}} Q^*(Y(\omega))(\xi + \beta)^{f(\omega)} \ge Q(l)Q_l(b).$$

Letting $\beta \to 0$, we get

$$\sum_{\omega \in \Omega_{l,b}} Q^*(Y(\omega))\xi^{f(\omega)} \ge Q(l)Q_l(b).$$

Therefore, ξ is a solution of (3.6). □

Example 1. We consider the following composition system (also see §4.3, [6],). Let $T = \{t\}$, and $N = \{S\}$. If

$$B_S(\alpha^*) = \begin{cases} 1 & \text{when } \alpha^* = (\beta_1, \beta_2), |Y(\beta_1)| = |Y(\beta_2)| \\ 0 & \text{otherwise} \end{cases}$$

and $\mathcal{S}_S = \{1\}$, then Ω is the set of balanced binary trees. The associated language is the set of strings of t of length 2^n, $n \ge 0$. Let $Q(S) = p$ and $Q(t) = q = 1 - p$, with $p \in (0, 1)$. Then the corresponding equation system is

$$\sum_{n=1}^{\infty} q^{2^n} Z^{2^n - 1} = p.$$

The convergence interval of the series on the left hand side of the equation is $(-1/q, 1/q)$, which is open. Therefore, there is a solution of the equation on $\{Z > 0\}$.

If in the above system, we change the binding function B_S to

$$B_S(\alpha^*) = \begin{cases} 1 & \text{when } \alpha^* = (\beta_1, \beta_2), |Y(\beta_1)| = |Y(\beta_2)| \text{ or } |Y(\beta_2)| + 1 \\ 0 & \text{otherwise} \end{cases}$$

while keeping everything else unchanged, then the generated language is the set of strings t^n, $t \ge 1$. The corresponding equation is

$$\sum_{n=2}^{\infty} q^n Z^{n-1} = p.$$

Again, the convergence interval of the series on the left hand side is $(-1/q, 1/q)$, which implies there is a solution for the equation on $\{Z > 0\}$. □

The following example shows the optimality of Proposition Proposition 5.

Example 2. Take $T = \{t\}$, $N = \{S\}$, and

$$B_S(\alpha^*) = \begin{cases} 1 & \text{if } \alpha^* = t \\ 2 & \text{if } \alpha^* = (\alpha, \beta) \text{ and } \mathrm{L}(\alpha) = \mathrm{L}(\beta) = S \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mathcal{S}_S = \{1, 2\}$ corresponds to the context-free grammar

$$S \to SS, \ S \to t.$$

Take $Q(t) = u$ and $Q(S) = v = 1 - u$. The probabilities

$$Q_S(b) = \begin{cases} p & \text{if } b = 2 \\ q = 1 - p & \text{if } b = 1 \end{cases}$$

correspond to the production probabilities $P(S \to SS) = p$ and $P(S \to t) = q$. The string that the only tree in $\Omega_{S,1}$ generates is t, and the set of strings that trees in $\Omega_{S,2}$ generate is $\{t^n\}_{n \geq 2}$. For each $n \geq 2$, there are $\Gamma(2n-1)/\Gamma(n)\Gamma(n+1)$ trees with the same yield t^n. For each such tree ω, $f(S, 1; \omega) = n$, and $f(S, 2; \omega) = n - 1$. Hence the corresponding equation system is

$$\begin{cases} uZ_{S,1} = vq \Rightarrow Z_{S,1} = \dfrac{vq}{u} \\ \displaystyle\sum_{n=2}^{\infty} \frac{(2n-2)!}{(n-1)!n!} u^n Z_{S,1}^n Z^{n-1} = vp. \end{cases}$$

Substitute $Z_{S,1} = vqu^{-1}$ into the second equation. The convergence domain of the resulting power series

$$F(Z) = \sum_{n=2}^{\infty} \frac{(2n-2)!}{(n-1)!n!} (vq)^n Z^{n-1}$$

is the closed interval $[-1/4vq, 1/4vq]$. We know that if $p > 1/2$, then there is no compositional probability distribution for the grammar (see §4.3, [6]). When $Z = 1/4vq$, the value of $F(Z)$ is vq. In order that there is a solution, it is necessary and sufficient that $F(1/4vq) \geq vp$, i.e., $q \geq p$, or $p \leq 1/2$. □

Example 3. The composition systems in Example 1 share the following properties.

1. The set $\mathcal{T} = T \cup \{(l, b) : l \in N, b \in \mathcal{S}_l\}$ is finite;
2. The arity of each B_l is 2;
3. For each $t \in \mathcal{T} \backslash T$,

$$\limsup_{n \to \infty} |\{Y(\omega) : \ \omega \in \Omega_t, \ |Y(\omega)| = n\}|^{1/n} = 1.$$

and

$$\limsup_{n \to \infty} \max_{\substack{Y \in T^* \\ |Y| = n}} |\{\omega \in \Omega_t : \ |Y(\omega)| = Y\}|^{1/n} = 1.$$

4. For each $t \in \mathcal{T}\backslash\mathcal{T}$, there are constants $0 \leq \beta_{s,t} \leq 1$, and $K_{s,t} \geq 0$ for all $s \in \mathcal{T}$, and $\omega_n \in \Omega_t$ with $|\omega_n| \to \infty$, such that
$$\beta_{s;t}|\omega_n| - K_{s;t} \leq f(s;\omega_n) \leq \beta_{s;t}|\omega_n| + K_{s;t}.$$

We show that if a composition system satisfies the above conditions, then for each $t \in \mathcal{T}$, the domain of convergence of the induced series
$$F_t(Z) = \sum_{\omega\in\Omega_t} Q^*(Y(\omega))Z^{f(\omega)}$$
is open inside the region $Z > 0$ for each $t \in \mathcal{T}\backslash T$, and hence compositional probability measures always exist.

Suppose F_t converges at some $Z > 0$. We want to show Z is an inner point in the domain of convergence of F_t. The series in x
$$F_t(Zx) = \sum_{\omega\in\Omega_t} Q^*(Y(\omega))Z^{f(\omega)}x^{|f(\omega)|}$$
is a univariate power series, where $|f(\omega)| = \sum f(t;\omega)$. By condition 2,
$$|f(\omega)| = \frac{|\omega| - 1}{2}.$$
Define power series
$$f_t(x) = \sum_{\omega\in\Omega_t} Q^*(Y(\omega))Z^{f(\omega)}x^{|\omega|},$$
Let ρ be the radius of convergence of f_t. We will show $\rho > 1$. Once this is proved, it is easy to see every $Z' < \rho Z$ is in the domain of convergence of F_t, implying Z is an inner point of the domain of convergence of F_t.

By conditions 2 and 3,
$$\limsup_{n\to\infty} |\{\omega \in \Omega_t : \ |\omega| = n\}|^{1/n} = 1.$$
Therefore, by condition 4,
$$\frac{1}{\rho} = \limsup_{\substack{|\omega|\to\infty \\ \omega\in\Omega_t}} \left|Q^*(Y(\omega))Z^{f(\omega)}\right|^{1/|\omega|} = \prod_{s\in T} Q(s)^{\beta_{s;t}} \prod_{s\in\mathcal{T}\backslash T} Z_s^{\beta_{s;t}}.$$

There are infinitely many $\omega \in \Omega_t$, such that
$$\begin{aligned}
& Q^*(Y(\omega))Z^{f(\omega)} \\
= \ & \prod_{s\in T} Q(s)^{f(s;\omega)} \prod_{s\in\mathcal{T}\backslash T} Z_s^{f(s;\omega)} \\
\geq \ & \prod_{s\in T} Q(s)^{\beta_{s;t}|\omega|+K_{s;t}} \prod_{\substack{s\in\mathcal{T}\backslash T \\ Z_s\geq 1}} Z_s^{\beta_{s;t}|\omega|-K_{s;t}} \prod_{\substack{s\in\mathcal{T}\backslash T \\ Z_s\leq 1}} Z_s^{\beta_{s;t}|\omega|+K_{s;t}} \\
\geq \ & \rho^{-|\omega|} \prod_{s\in T} Q(s)^{K_{s;t}} \prod_{\substack{s\in\mathcal{T}\backslash T \\ Z_s\geq 1}} Z_s^{-K_{s;t}} \prod_{\substack{s\in\mathcal{T}\backslash T \\ Z_s\leq 1}} Z_s^{K_{s;t}}.
\end{aligned}$$
Because $F_{l,b}(Z)$ converges, $1/\rho < 1$. Thus $\rho > 1$. □

3.4 Subsystems

Suppose we have a composition system $\mathcal{C}' = (T', N', \{B_l, \mathcal{S}_l\}_{l\in N'}$ with Ω' as the set of trees. We can build a new composition system in the following way. First, take Ω' as part of a new terminal set T''. Suppose $T'' = \Omega' \cup A$, where $A \cap \Omega' = \emptyset$. Then we define a label set N'' disjoint from N', and for each label $l \in N''$, a composition rule $(B_l, \mathcal{S}_l)$. The new composition system $\mathcal{C}'' = (T'', N'', \{B_l, \mathcal{S}_l\}_{l\in N''})$ is not a super-system of $\mathcal{C}'$, because $\mathcal{C}'$ and $\mathcal{C}''$ have disjoint label sets and composition rules and their terminal sets are different. On the other hand, the composition system $\mathcal{C}$ with terminal set $T' \cup A = T' \cup (T''\backslash\Omega')$, label set $N \cup N'$, and composition rules $\{B_l, \mathcal{S}_l\}_{l\in N\cup N'}$ is a super-sytem of $\mathcal{C}'$.

The above construction can be formulated into the following definition.

Definition 6. Suppose $\mathcal{C} = (N, \{B_l, \mathcal{S}_l\}_{l\in N}, T)$ is a composition system with Ω being the set of objects. Suppose T' and N' are non-empty subsets of T and N, respectively. For each $l \in N'$, assume $\mathcal{S}'_l$ is a non-empty subset of $\mathcal{S}_l$. Let $\mathcal{C}'$ be the composition system formed by T', N', and $\{B_l, \mathcal{S}'_l\}_{l\in N'}$. Let Ω' be the set of objects generated by $\mathcal{C}'$. Since $T' \neq \emptyset$, Ω' is not empty.

$\mathcal{C}'$ is said to be a subsystem of $\mathcal{C}$, denoted as $\mathcal{C}' \subset \mathcal{C}$, if Ω' contains all $\omega \in \Omega$ with $\mathrm{L}(\omega) = l \in N'$ and $A(\omega) \in \mathcal{S}'_l$. The composition system with terminal set $T \cup \Omega'$, label set $N_1 \cup N_2$, where $N_1 = N\backslash N'$, and $N_2 = \{l \in N' :\ \mathcal{S}_l\backslash\mathcal{S}'_l \neq \emptyset\}$, and composition rules $\{B_l, \mathcal{S}_l\}_{l\in N_1} \cup \{B_l, \mathcal{S}_l\backslash\mathcal{S}'_l\}_{l\in N_2}$, is called the quotient system of $\mathcal{C}$ over $\mathcal{C}'$ and is denoted as $\mathcal{C}/\mathcal{C}'$. The set of objects generated by $\mathcal{C}/\mathcal{C}'$ is denoted as Ω/Ω'. □

Intuitively speaking, $\mathcal{C}/\mathcal{C}'$ is an abstraction of $\mathcal{C}$. It takes objects in $\mathcal{C}'$ as terminals, which, by definition, are not decomposable, thus losing the details about them. On the other hand, $\mathcal{C}$ can be thought of as being more detailed than $\mathcal{C}/\mathcal{C}'$. The information about $\mathcal{C}$ is determined by that about both $\mathcal{C}/\mathcal{C}'$ and $\mathcal{C}'$.

Subsystems can be used to construct of compositional probability distributions. For example, if both $\mathcal{C}'$ and $\mathcal{C}/\mathcal{C}'$ satisfy the conditions of Proposition Proposition 4, then for any Q on $T \cup N$ and Q_l on $\mathcal{S}_l$, $l \in N$, both

$$P_1(\omega) = \begin{cases} Q(\omega), & \text{if } \omega \in T',\\ Q(l)Q_l(B_l(\alpha^*))\dfrac{P_1^*(\alpha^*)}{P_1^*\Big(\{\beta^* \in \Omega^* : B_l(\beta^*) = B_l(\alpha^*)\}\Big)}, & \text{if } \omega = l(\alpha^*) \in \Omega', \end{cases}$$

and

$$P_2(\omega) = \begin{cases} Q(\omega), & \text{if } \omega \in T\backslash T',\\ P_1(\omega), & \text{if } \omega \in \Omega',\\ Q(l)Q_l(B_l(\alpha^*))\dfrac{P_2^*(\alpha^*)}{P_2^*\Big(\{\beta^* \in \Omega^* : B_l(\beta^*) = B_l(\alpha^*)\}\Big)}, & \text{if } \omega = l(\alpha^*) \in \Omega/\Omega', \end{cases}$$

have solutions. Note that neither P_1 nor P_2 is a probability distribution, because each of the sums of P_1 and P_2 is less than 1. The existence of the solutions is guaranteed by Proposition

Proposition 6 in Appendix. On the other hand, the measure

$$P(\omega) = \begin{cases} P_1(\omega), & \text{if } \omega \in \Omega' \\ P_2(\omega), & \text{if } \omega \in \Omega/\Omega' \end{cases}$$

is a compositional probability distribution on Ω.

There is another application of subsystems. Assume we have a compositional probability distribution P_1 on a system $\mathcal{C}' = (T', N', \{B'_l, \mathcal{S}'_l\}_{l \in N'})$, and the observable measures are $Q'(l)$ and $Q'_l(b)$. Suppose $\mathcal{C}'$ is expanded into a larger system $\mathcal{C}$. Assume the expansion does not change $\mathcal{S}_l$ for any $l \in N'$. It just adds more terminals to T', and more labels to N', and sets up rules for the new labels.

$Q'(l)$ now becomes the conditional probability measure on N'. Thus in $\mathcal{C}$, the probability of each $l \in N'$ is changed to $\lambda Q'(l)$, for some constant λ. However, for any $l \in N'$ and any $b \in \mathcal{S}'_l$, $Q'_l(b)$ is not changed. If all the binding functions B_l, $l \in N'$, have the same arity, then the probability of $\omega \in \Omega'$ is simply changed to $\lambda P_1(\omega)$ when ω is considered as an element in Ω. This makes enlarging a system and adjusting the probability distribution easy.

3.5 The Gibbs Form of Compositional Probability Distributions

We now discuss the Gibbs form of compositional probability distributions. Suppose P is a compositional probability distribution on Ω. Then P can be formulated as in (3.5). Extend $Z = \{Z_t\}_{t \in \mathcal{T} \setminus T}$ to $\{Z_t\}_{t \in \mathcal{T}}$, where for $t \in T$, $Z_t = Q(t)$. Also extend $f(\omega) = \{f(t;\omega)\}_{t \in \mathcal{T} \setminus T}$ to $\{f(t;\omega)\}_{t \in \mathcal{T}}$. Finally, let $\lambda = \{\log Z_t\}_{t \in \mathcal{T}}$. Then $P(\omega)$ takes the form of Gibbs distribution,

$$P(\omega) = P_\lambda(\omega) = \exp\Big(\lambda \cdot f(\omega)\Big). \tag{3.9}$$

A special property of the Gibbs distribution (3.9) is that its normalization constant is 1.

For an arbitrary λ, P_λ is a positive measure on Ω, but not necessarily a probability measure. Among all the λ's which make P_λ a probability measure on Ω, $\lambda = \{\log Z_t\}_{t \in \mathcal{T}}$ has the following minimization property,

$$\lambda = \mathop{\arg\min}_{\substack{\lambda':\ P_{\lambda'} \text{ is} \\ \text{a prob.}}} \sum_{t \in \mathcal{T}} Q(t) \log \frac{Q(t)}{P_{\lambda'}(\Omega_t)}. \tag{3.10}$$

Indeed, the sum on the right hand side of (3.10) is always non-negative. If an compositional distribution exists, then the sum achieves 0 at $\lambda = \{\log Z_t\}_{t \in \mathcal{T}}$. Therefore λ is a minimizer.

Bibliography

[1] E. Bienenstock. Composition. A. Aertsen and V. Braitenberg (editors). *Brain Theory: Biological Basis and Computational Theory of Vision.* In preparation.

[2] N. Chomsky. *Syntactic Structures.* Mouton. 1976.

[3] S. Geman, D. Potter, and Z. Chi. *Compositional Systems.* Technical Report, Division of Applied Mathematics, Brown University. February, 1998.

[4] D. F. Potter. *Compositional Pattern Recognition.* PhD Thesis, Division of Applied Mathematics, Brown University. 1998. In preparation.

Appendix

In this Appendix we will prove Proposition Proposition 4. First, we need to introduce some notations.

Definition 7.

1. The mapping $l : \alpha^* \mapsto l(\alpha^*)$ can be thought of as a function from Θ^* to Θ_l, which is one-to-one and onto. We write its inverse as l^{-1};

2. The *graph* of any tree $\omega \in \Theta$ is a tree with the same topology as ω but with all nodes being unlabeled (Figure 3.1);

3. That a tree ω is *compatible* with a tree graph g, denoted as $\omega \sim g$ means the following. If g is a tree with a single node, then $\omega \sim g$. If g is a tree with daughter subtrees $g_1, \ldots, g_n$, then $\omega = l(\alpha_1, \ldots, \alpha_m)$ is compatible with g if and only if $m = n$ and each α_i, $1 \le i \le m$, is compatible with g_i. If $\omega \sim g$, then for each node $v \in g$, let $\omega(v)$ represent the subtree of ω with v as the root;

4. The *arrangement* of any $\omega \in \Omega$, denoted $E(\omega)$, is a tree with the same topology as ω but with each node being annotated by its type (Figure 3.1);

5. For any $\omega \in \Theta$, the depth of a subtree ω' is the depth of the root of ω' in the tree ω and is denoted as $d(\omega', \omega)$. By this definition, $d(\omega, \omega) = 1$. □

Proposition Proposition 4 can be expressed in a little more general form, where Q is a finite positive measure instead of a probability measure on $\mathcal{T}$.

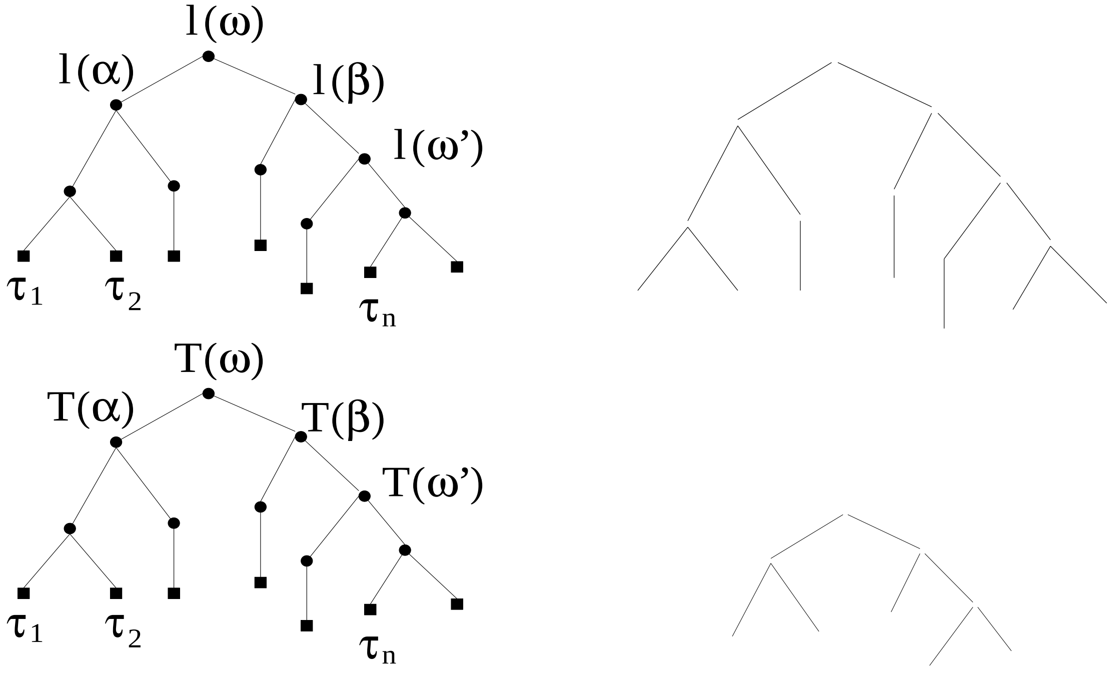


Figure 3.1: A tree $\omega = l(\alpha, \beta)$ (upper left), its graph (upper right), its arrangement (lower left) and a compatible graph. Circles are nonterminals and squares are terminals

Proposition 6. Suppose for any $l \in N$ and any $b \in \mathcal{S}_l$,

$$\max\{h(\omega) : \ \omega \in \Omega_{l,b}\} < \infty \tag{A3.1}$$

and

$$\max\{|l^{-1}(\omega)| : \ \omega \in \Omega_{l,b}\} < \infty. \tag{A3.2}$$

Assume Q is a positive measure on $\mathcal{T}$, with $Q(t) > 0$ for each $t \in \mathcal{T}$. If $Q(\mathcal{T}) < \infty$, then there exists a compositional probability measure satisfying (3.1).

Our proof of Proposition Proposition 6 is based on the following fixed point theorem, which is due to Schauder.

Theorem Suppose X is a Banach space, $C \subset X$ is closed and convex. If $F : C \to C$ is continuous and $F(C)$ is sequentially compact, then F has a fixed point in C. □

Proof of Proposition Proposition 6: For any type $t = (l, b)$, let

$$\begin{aligned} h(t) &= \max_{\omega \in \Omega_t} h(\omega), \\ m(t) &= |\{l^{-1}(\omega) : \ \omega \in \Omega_t\}|, \\ n(t) &= \max_{\omega \in \Omega_t} |l^{-1}(\omega)|. \end{aligned}$$

Then $h(t)$, $m(t)$, and $n(t)$ are finite. Also write $h(l, b)$, $m(l, b)$ and $n(l, b)$ for $h(t)$, $m(t)$, and $n(t)$, respectively. For consistency, define, for $\tau \in T$, $h(\tau) = 1$, $m(\tau) = 1$ and $n(\tau) = 0$.

Let X be the l^1 space on Ω, i.e.

$$X = \left\{ x : \Omega \to \mathbf{R} : \ x(\tau) = Q(\tau), \ \forall \ \tau \in T, \ \sum_{\omega \in \Omega} |x(\omega)| < \infty \right\}.$$

Let $M = \max\{Q(\mathcal{T}), 1\}$. Define a positive measure D on Ω inductively as follows. For $\tau \in T$, $D(\tau) = Q(\tau) > 0$. For any $t = (l, b) \in \mathcal{T} \backslash T$ and $\omega = l(\alpha^*) \in \Omega_t$,

$$D(\omega) = \frac{Q(t)}{m(t) M^{n(t)}} D^*(\alpha^*). \tag{A3.3}$$

For consistency, we define $D^*(\emptyset) = 1$. Then $D(\tau)$ can also be written in the form of (A3.3).

For ease of typing, we introduce a new notation. If $\beta^* \in \Omega^*$ satisfies $B_l(\beta^*) = b \in \mathcal{S}_l$, then we say β^* is compatible with type $t = (l, b)$ and use $\beta^* \sim t$ to represent this. For consistency, we define $\emptyset$ to be the only string that is compatible with a type t if $t \in T$.

Lemma 4. D has the following properties,

$$D(\Omega) \le M \tag{A3.4}$$

$$0 < \sum_{\beta^* \sim t} D^*(\beta^*) \le m(t) M^{n(t)}. \tag{A3.5}$$

Proof. We will get (A3.4) by showing for all $n \ge 1$,

$$\sum_{h(\omega) \le n} D(\omega) \le M, \tag{A3.6}$$

When $n = 1$, the sum equals $\sum_T Q(\tau) \le M$. Assume (A3.6) is true for $n \le k$. Then

$$\sum_{h(\omega) \le k+1} D(\omega) = \sum_{t \in \mathcal{T}} \frac{Q(t)}{m(t) M^{n(t)}} \sum_{\substack{\alpha^* \sim t \\ h(\alpha^*) \le k}} D^*(\alpha^*),$$

where $h(\alpha^*) = \max_{\alpha \in \alpha^*} h(\alpha)$. By induction hypothesis,

$$\sum_{\substack{\alpha^* \sim t \\ h(\alpha^*) \le k}} D^*(\alpha^*) = \sum_{j=1}^{\infty} \sum_{\substack{\alpha^* \sim t \\ h(\alpha^*) \le k \\ |\alpha^*| = j}} D^*(\alpha^*) \le \sum_{j=1}^{\infty} \mathbf{1}_{\{\exists \omega \in \Omega_t, \ |l^{-1}(\omega)| = j\}} M^{n(t)} = m(t) M^{n(t)}, \tag{A3.7}$$

which, together with last equation, implies (A3.6). Letting $n \to \infty$ in (A3.6), we then prove (A3.4). Letting $k \to \infty$ in (A3.7), we get (A3.5). □

Lemma 5. Let g be a tree graph. To each node $v \in g$, assign a type $t(v)$, such that $t(v) \in \mathcal{T} \backslash T$ unless v is a leaf of g. Then

$$\sum_{\substack{\omega \sim g, \forall v \in g \\ \mathrm{T}(\omega(v)) = t(v)}} D(\omega) \le E(g, t), \tag{A3.8}$$

where

$$E(g,t) = {\prod_{v\in g}}' \frac{Q(t(v))}{\sum\limits_{\beta^*\sim t(v)} D^*(\beta^*)} \cdot {\prod_{v\in g}}'' Q(t(v)), \tag{A3.9}$$

where the production $\prod'$ runs over all non-terminal nodes of g and $\prod''$ over all terminal nodes of g.

Proof. When $h(g) = 1$, the right hand side of (A3.8) is $Q(t)$, where t is the type assigned to the only node in g. The left hand side of (A3.8) is the sum of

$$\sum_{\omega\in\Omega_t} D(\omega) = \sum_{\alpha^*\sim t} \frac{Q(t)}{m(t)M^{n(t)}} D^*(\alpha^*).$$

By (A3.5), the sum is less than $Q(t)$.

Suppose (A3.8) is true for all finite graphs g with $h(g) \le k$. Given a tree graph g with height $k+1$ and daughter subtrees $g_1, \ldots, g_n$, by (A3.3) and (A3.5), for any $\omega \sim g$ with $\mathrm{T}(\omega) = t(v_0)$, where v_0 is the root of g,

$$D(\omega) \le \frac{Q(t(v_0))}{\sum\limits_{\beta^*\sim t(v_0)} D^*(\beta^*)} D^*(\alpha^*),$$

which leads to

$$\begin{aligned}
\sum_{\substack{\omega\sim g, \forall v\in g\\ \mathrm{T}(\omega(v))=t(v)}} D(\omega) \quad &\le \quad \frac{Q(t(v_0))}{\sum\limits_{\beta^*\sim t(v_0)} D^*(\beta^*)} \sum_{\substack{\alpha^*\sim t(v_0)\\ \text{for } i=1,\ldots n, \alpha_i\sim g_i,\\ \forall v\in g_i, \mathrm{T}(\alpha_i(v))=t(v)}} D^*(\alpha^*) \\
&\le \quad \frac{Q(t(v_0))}{\sum\limits_{\beta^*\sim t(v_0)} D^*(\beta^*)} \prod_{i=1}^{n} \sum_{\substack{\alpha_i\sim g_i, \forall v\in g_i\\ \mathrm{T}(\alpha_i(v))=t(v)}} D(\alpha_i).
\end{aligned}$$

Every $h(g_i) \le k$. Then by induction, we prove (A3.8). □

Now define C as the set of all $x \in X$ which satisfy the following conditions,

- C1. For any $\tau \in T$, $x(\tau) = Q(\tau)$;
- C2. For any $\omega \in \Omega$, $x(\omega) \ge D(\omega)$;
- C3. For any tree graph g, any assignment $t : \{v \in g\} \to \mathcal{T}$ with $t(v) \notin T$ unless v is a terminal of g,

$$\sum_{\substack{\omega\sim g, \forall v\in g\\ \mathrm{T}(\omega(v))=t(v)}} x(\omega) \le E(g,t). \tag{A3.10}$$

C is not empty, because $D \in C$. We want to use Schauder's fixed point theorem to prove there is a solution for (3.1) in C. To this end, define a mapping $F : C \to R^{\Omega}$, such that

$$\begin{cases} (Fx)(\tau) = x(\tau), & \forall\ \tau \in T \\ (Fx)(\omega) = Q(l)Q_l(b)\dfrac{x^*(\alpha^*)}{\sum\limits_{\substack{\beta^* \in \Omega^* \\ B_l(\beta)=b}} x^*(\beta^*)}, & \forall\ \omega = l(\alpha^*) \in \Omega, B_l(\alpha^*) = b \end{cases} \tag{A3.11}$$

The definition (A3.11) makes sense because

$$0 < \sum_{\beta^* \sim t} x^*(\beta^*) \leq m(t)M^{n(t)}. \tag{A3.12}$$

The second half of (A3.12) can be proved in the same way as (A3.5).

It is clear that C is convex and closed. In order to show that $F(C)$ is sequentially compact, it is enough to show that $F(C) \subset C$ and C is tight. First we shall show that C is tight.

Lemma 6. For any $\epsilon > 0$, $n \geq 2$, and finite $I \subset \mathcal{T}$, there is a finite $J \subset \mathcal{T}$ with $J \supset I$, such that

$$\sum_{(g,t) \in G} E(g,t) < \epsilon, \tag{A3.13}$$

where $G = G_n(I, J)$ is the set of pairs (g, t) satisfying the following conditions,

G1 For each $(g,t) \in G$, $h(g) = n$, and $t : \{v \in g\} \to \mathcal{T}$ is a mapping such that $t(v) \notin T$ unless v is a terminal of g;

G2 For any $v \in g$ with $d(v, g) \leq n - 1$, $t(v) \in I$;

G3 There is a $v \in g$ with $d(v, g) = n$ such that $t(v) \notin J$;

G4 The set $\{\omega \in \Omega : \omega \sim g$, and for every $v \in g, \mathrm{T}(\omega(v)) = t(v)\}$ is not empty;

G5 Every $(g,t) \in G$ is maximum. That is, there are no (g,t) and (g', t'), such that $g \subset g'$ and for any $v \in g$, $t(v) = t'(v)$.

Proof. Let $N = \max_{t \in I} n(t)$. Here t represents an element in $\mathcal{T}$ instead of a mapping to $\mathcal{T}$. Then N is the maximum number of daughter subtrees a tree ω whose type is in I can have. By (A3.2), N is finite. Fix $J \supset I$ and let $G = G_n(I, J)$. If g is a tree graph with $(g,t) \in G$ for some mapping $t : \{v \in g\} \to \mathcal{T}$, then by condition G4, $h(g)$ has to be n. For any $v \in g$ with $d(v, g) \leq n - 1$, since $t(v) \in I$, the number of daughter subtrees of v must be less or equal to N, otherwise there would not be an $\omega \in \Omega$ with $\omega \sim g$ and $\mathrm{T}(\omega(v)) \in I$, contradicting to G4. Therefore, the set of all g with $(g,t) \in G$ for some t is finite. In addition, this set is independent of the selection of $J \supset I$.

Given $(g,t) \in G$,

$$E(g,t) \leq \prod_{\substack{v \text{ non-} \\ \text{terminal}}} \frac{Q(t(v))}{\sum\limits_{\beta^* \sim t(v)} D^*(\beta^*)} \prod_{\substack{v \text{ terminal} \\ d(v,g)<n}} Q(t(v)) \prod_{\substack{v \text{ terminal} \\ d(v,g)=n}} Q(t(v)).$$

Let

$$R_0 = \max_{t \in I} \frac{Q(t)}{\sum\limits_{\beta^* \sim t} D^*(\beta^*)},$$

and $R = \max\{R_0, M\}$ (recall $M = \max\{Q(\mathcal{T}), 1\}$). Since a non-terminal of g necessarily has depth less than n, then

$$E(g,t) \le R^{|g|} \prod_{\substack{v \text{ terminal} \\ d(v,g)=n}} Q(t(v)).$$

Because there are only finite number of g with $(g,t) \in G$ for some t, $|g|$ is bounded by a constant, say, A. So we get

$$\sum_{(g,t)\in G} E(g,t) \le R^A \sum_{(g,t)\in G} \prod_{\substack{v \text{ terminal} \\ d(v,g)=n}} Q(t(v)).$$

Notice that A is independent of the selection of J.

G is the union of disjoint sets G_α which have the following two properties,

1. For any (g,t), and $(g',t') \in G_\alpha$, $g = g'$, and for any $v \in g$ with $d(v,g) < n$, $t(v) = t'(v)$;
2. If $\alpha \ne \beta$, then for $(g,t) \in G_\alpha$ and $(g',t') \in G_\beta$, either $g \ne g'$ or there is a $v \in g$ with $d(v,g) < n$, such that $t(v) \ne t'(v)$.

It is easy to check that the number of G_α's is finite. In addition, the number is independent of the selection of J. Let the number be K_0. For any G_α, consider

$$\sum_{(g,t)\in G_\alpha} \prod_{\substack{v \text{ terminal} \\ d(v,g)=n}} Q(t(v)).$$

Since at least one of the $t(v)$ is not in J, then the sum is bounded $M^a - Q(J)^a \le M^{|g|} - Q(J)^{|g|} \le M^A - Q(J)^A$, where a is the number of $v \in g$ with $d(v,g) = n$. We then get

$$\sum_{(g,t)\in G} E(g,t) \le K_0 R^A (M^A - Q(J)^A).$$

Again, the bound is independent of the selection of $J \supset I$. Therefore, we can choose $J \supset I$ large enough to make the right hand side less than ϵ. This proves the lemma. □

Lemma 7. C is tight.

Proof. Fix $\epsilon > 0$. Then there is a finite set $I_1 \subset \mathcal{T}$ such that

$$\sum_{t \in I_1} Q(t) < \frac{\epsilon}{2}.$$

Define

$$H = \max_{t \in I_1} h(t).$$

By Lemma Lemma 6, there is a nested sequence of finite sets $I_2 \subset I_3 \ldots \subset I_H$ with $I_2 \supset I_1$, such that

$$\sum_{(g,t) \in G_k(I_{k-1}, I_k)} E(g,t) < \frac{\epsilon}{2H}, \quad \text{for } 2 \leq k \leq H,$$

where $G_k(I_{k-1}, I_k)$ are defined as in Lemma Lemma 6.

Define $\tilde{S}_1 = \{\omega : \ \mathrm{T}(\omega) \notin I_1\}$. For $2 \leq k \leq H$, define $\tilde{S}_k$ as the set of ω which satisfy the following conditions

1. For any i, $1 \leq i < k$, for any $\omega' \subset \omega$ with $d(\omega', \omega) = i$, $\mathrm{T}(\omega) \in I_i$;
2. There is an $\omega' \subset \omega$ with $d(\omega', \omega) = k$ such that $\mathrm{T}(\omega') \notin I_k$.

Then $\tilde{S}_i$ are disjoint and

$$\bigcup_{i=1}^{H} \tilde{S}_i = \{\omega : \ \text{ there is an } i, 1 \leq i \leq k, \text{ and } \omega' \subset \omega \text{ with } d(\omega', \omega) = i, \ \mathrm{T}(\omega') \notin I_i\}$$

Because I_k are increasing, for k, $2 \leq k \leq H$, $\tilde{S}_k \subset S_k$, where S_k is the set of ω satisfying

1. For any $\omega' \subset \omega$ with $d(\omega', \omega) < k$, $\mathrm{T}(\omega) \in I_{k-1}$;
2. There is an $\omega' \subset \omega$ with $d(\omega', \omega) = k$ such that $\mathrm{T}(\omega') \notin I_k$.

It is easy to see that for k, $2 \leq k \leq H$,

$$S_k = \bigcup_{(g,t) \in G_k(I_{k-1}, I_k)} \{\omega : \ \omega \sim g, \ \mathrm{T}(\omega(v)) = t(v), \text{ for any } v \in g\}.$$

Therefore, by (A3.10), for k, $2 \leq k \leq H$,

$$\sum_{\omega \in \tilde{S}_k} x(\omega) \leq \sum_{\omega \in S_k} x(\omega) \leq \sum_{(g,t) \in G_k} E(g,t) \leq \frac{\epsilon}{2H}.$$

We also have

$$\sum_{\omega \in \tilde{S}_1} x(\omega) = \sum_{\mathrm{T}(\omega) \notin I_1} x(\omega) \leq \frac{\epsilon}{2}.$$

Thus we get

$$x\left(\bigcup_{i=1}^{H} \tilde{S}_i\right) \leq \epsilon.$$

Because every ω with $\mathrm{T}(\omega) \in I_1$ has height less or equal to H, therefore if $\omega \in A$, where

$$A = \left(\bigcup_{i=1}^{H} \tilde{S}_i \right)^c = \{\omega : \text{ for any } i, 1 \leq i \leq H, \text{ and } \omega' \subset \omega, \text{ with } d(\omega', \omega) = i, \mathrm{T}(\omega') \in I_i\},$$

then $h(\omega) \leq H$, and for each $\omega' \subset \omega$, $\mathrm{T}(\omega') \in I_i \subset I_H$. Therefore, each label of $E(\omega)$ is in I_H. Since the correspondence between objects and their arrangements is one-to-one, then A is a finite set. Thus we get $x(A^c) < \epsilon$. This completes the proof that C is tight. □

Now we prove $F(C) \subset C$. For any $x \in C$, condition C1 is clearly satisfied. By (A3.3), (A3.11), and (A3.12), for any type $t = (l, b) \in \mathcal{T} \backslash T$, for any $\omega = l(\alpha^*) \in \Omega_t$,

$$(Fx)(\omega) = Q(t) \frac{x^*(\alpha^*)}{\sum\limits_{\beta^* \sim t} x^*(\beta^*)} \geq Q(t) \frac{D^*(\alpha^*)}{m(t) M^{n(t)}} = D(\omega).$$

As for C3, if a tree graph g is of height 1, then for any t assigned to the single node in g,

$$\sum_{\substack{\omega \sim g, \forall v \in g \\ \mathrm{T}(\omega(v)) = t(v)}} (Fx)(\omega) = Q(t) = E(g, t).$$

The case where $h(g) \geq 2$ can then be proved following the proof of (A3.8).

The only thing that remains to show is the continuity of F. For this purpose, we shall use the following version of dominance convergence theorem without giving its proof.

Lemma 8. Let ν be a positive measure on a measurable space X. Suppose $\{f_n\}$, $\{g_n\}$ are sequences of measurable functions on X such that $|f_n| \leq g_n$, $\forall\ n \geq 1$, $f_n \to f$, ν-a.s. and $g_n \to g$, ν-a.s. If

$$\lim_{n \to \infty} \int g_n \, d\nu = \int g \, d\nu < \infty,$$

then

$$\lim_{n \to \infty} \int f_n \, d\nu = \int f \, d\nu.$$

□

Continuing the proof, suppose $x_n \to x$ in C, i.e., $\sum_{\omega \in \Omega} \|x_n(\omega) - x(\omega)\| \to 0$. Let $y_n = F(x_n)$ and $y = F(x)$. We want to show $\|y_n - y\| = \sum_{\omega \in \Omega} \|y_n(\omega) - y(\omega)\| \to 0$. The sum is dominated by $\sum_{\omega \in \Omega} g_n(\omega)$, where $g_n = y_n + y$.

Our plan is to show that for each ω, $y_n(\omega) \to y(\omega)$. Then $g_n(\omega) \to 2y(\omega)$. Since $\sum_{\omega \in \Omega} g_n(\omega) \equiv 2 \sum_{\omega \in \Omega} y(\omega) = 2M$, then by the above dominance convergence result, $\sum_{\omega \in \Omega} \|y_n(\omega) - y(\omega)\| \to 0$.

Now we show $y_n(\omega) \to y(\omega)$. Given $t = (l, b) \in \mathcal{T} \backslash T$, for any $\omega = l(\alpha^*) \in \Omega_t$,

$$\left| \sum_{B_l(\alpha^*) = b} x_n^*(\alpha^*) - \sum_{B_l(\alpha^*) = b} x^*(\alpha^*) \right| \leq \sum_{k=1}^{h(t)} \left| \sum_{\substack{B_l(\alpha^*) = b \\ |\alpha^*| = k}} x_n^*(\alpha^*) - \sum_{\substack{B_l(\alpha^*) = b \\ |\alpha^*| = k}} x^*(\alpha^*) \right|$$

For each k, $1 \leq k \leq h(t)$,

$$\begin{aligned}
& \sum_{\substack{B_l(\alpha^*)=b \\ |\alpha^*|=k}} |x_n^*(\alpha^*) - x^*(\alpha^*)| \\
\leq \quad & \sum_{\substack{B_l(\alpha^*)=b \\ |\alpha^*|=k}} \sum_{i=1}^{k} x^*(\alpha_1 \cdots \alpha_{i-1}) \, |x_n(\alpha_i) - x(\alpha_i)| \, x_n^*(\alpha_{i+1} \cdots \alpha_k) \\
\leq \quad & kM^{k-1} \|x_n - x\| \to 0,
\end{aligned}$$

leading to

$$\sum_{B_l(\alpha^*)=b} x_n^*(\alpha^*) \to \sum_{B_l(\alpha^*)=b} x^*(\alpha^*) > 0.$$

Therefore,

$$Q(t) \frac{x_n^*(\alpha^*)}{\displaystyle\sum_{B_{\mathrm{L}(\omega)}(\beta^*)=b} x_n(\beta^*)} \to Q(t) \frac{x^*(\alpha^*)}{\displaystyle\sum_{B_{\mathrm{L}(\omega)}(\beta^*)=b} x(\beta^*)},$$

i.e., $y_n(\omega) \to y(\omega)$, completing the proof. □

Chapter 4

Estimation of Probabilistic Context-Free Grammars

The assignment of probabilities to the productions of a context-free grammar may generate an improper distribution: the probability of all finite parse trees is less than one. The condition for proper assignment is rather subtle. Production probabilities can be estimated from parsed or unparsed sentences, and the question arises as to whether or not an estimated system is *automatically* proper. We show here that estimated production probabilities always yield proper distributions.

4.1 Introduction

Context-free grammars (CFG's) are useful because of their relatively broad coverage and because of the availability of efficient parsing algorithms. Furthermore, CFG's are readily fit with a probability distribution (to make *probabilistic* CFG's—or PCFG's), rendering them suitable for ambiguous languages through the maximum *a posteriori* rule of choosing the most probable parse.

For each non-terminal symbol, a (normalized) probability is placed on the set of all productions from that symbol. Unfortunately, this simple procedure runs into an unexpected complication: The language generated by the grammar may have probability less than one. The reason is that the derivation tree may have probability greater than zero of never terminating—some mass can be lost to infinity. This phenomenon is well known and well understood, and there are tests for "tightness" (by which we mean total probability mass equal to one) involving a matrix derived from the expected growth in numbers of symbols generated by the probabilistic rules (see for example Booth & Thompson [3], Grenander [5], and Harris [6]).

What if the production probabilities are estimated from data? Suppose, for example, that we have a parsed corpus that we treat as a collection of (independent) samples from a gram-

[1] Supported by the Army Research Office (DAAL03-92-G-0115), the National Science Foundation (DMS-9217655), and the Office of Naval Research (N00014-96-1-0647).

mar. It is reasonable to hope that if the trees in the sample are finite, then an estimate of production probabilities based upon the sample will produce a system that assigns probability zero to the set of infinite trees. For example, there is a simple maximum-likelihood prescription for estimating the production probabilities from a corpus of trees (see §2), resulting in a PCFG— is it tight? If the corpus is unparsed then there is an iterative approach to maximum likelihood estimation (the "EM" or "Baum-Welsh" algorithm—again, see §2) and the same question arises: do we get actual probabilities or do the estimated PCFG's assign some mass to infinite trees?

We will show that in both cases the estimated probability is tight. [1]

Wetherell [9] has asked a similar question: a scheme (different from maximum likelihood) is introduced for estimating production probabilities from an unparsed corpus, and it is conjectured that the resulting system is tight. (Wetherell and others use the designation "consistent" instead of "tight," but in statistics consistency refers to the asymptotic correctness of an estimator.)

A trivial example is the CFG with one nonterminal and one terminal symbol, in Chomsky normal form:

$$\begin{aligned} A &\rightarrow AA \\ A &\rightarrow a \end{aligned}$$

where 'a' is the only terminal symbol. Assign probability p to the first production ($A \rightarrow AA$) and $q = 1 - p$ to the second ($A \rightarrow a$). Let S_h be the total probability of all trees with depth less than or equal to h. For example, $S_2 = q$ corresponding to $A \rightarrow a$, and $S_3 = q + pq^2$ corresponding to $\{A \rightarrow a\} \cup \{A \rightarrow AA, A \rightarrow a, A \rightarrow a\}$. In general, $S_{h+1} = q + pS_h^2$. (Condition on the first production: with probability q the tree terminates and with probability p it produces two nonterminal symbols, each of which must now terminate with depth less than or equal to h.) It is not hard to show that S_h is nondecreasing and converges to $\min(1, \frac{q}{p})$, meaning that a proper probability is obtained if and only if $p \leq \frac{1}{2}$.

What if p is estimated from data? Given a set of finite parse trees $\omega_1, \omega_2, ...\omega_n$, the maximum likelihood estimator for p (see §2) is, sensibly enough, the "relative frequency" estimator

$$\hat{p} = \frac{\sum_{i=1}^{n} f(A \rightarrow AA; \omega_i)}{\sum_{i=1}^{n} [f(A \rightarrow AA; \omega_i) + f(A \rightarrow a; \omega_i)]}$$

where $f(\cdot; \omega)$ is the number of occurrences of the production "$\cdot$" in the tree ω. The sentence a^m, although ambiguous (there are multiple parses when $m > 2$), always involves $m - 1$ of the $A \rightarrow AA$ productions and m of the $A \rightarrow a$ productions. Hence $f(A \rightarrow AA; \omega_i) < f(A \rightarrow a; \omega_i)$ for each ω_i. Consequently

$$f(A \rightarrow AA; \omega_i) < \frac{1}{2}[f(A \rightarrow AA; \omega_i) + f(A \rightarrow a; \omega_i)]$$

for each ω_i, and $\hat{p} < \frac{1}{2}$. The maximum likelihood probability is tight.

[1] When estimating from an unparsed corpus, we shall assume a model without null or unit productions—see §2.

If only the *yields* (left-to-right sequence of terminals) $Y(\omega_1), Y(\omega_2), ...Y(\omega_n)$ are available, the EM algorithm can be used to iteratively "climb" the likelihood surface (see §2). In the simple example here, the estimator converges in one step and is the same $\hat{p}$ as if we had observed the entire parse tree for each ω_i. Thus, $\hat{p}$ is again less than $\frac{1}{2}$ and the distribution is again tight.

4.2 Maximum Likelihood Estimation

More generally, let $G = (V, T, R, S)$ denote a context-free grammar with finite variable set V, start symbol $S \in V$, finite terminal set T, and finite production (or rule) set R. (We use "R" in place of the more typical "P" to avoid confusion with probabilities.) Each production in R has the form $A \to \alpha$, where $A \in V$ and $\alpha \in (V \cup T)^*$. In the usual way, probabilities are introduced through the productions: $P : R \to [0,1]$ such that $\forall A \in V$

$$\sum_{\substack{\alpha \in (V \cup T)^* \\ \text{s.t. } (A \to \alpha) \in R}} p(A \to \alpha) = 1. \tag{4.1}$$

Given a set of finite parse trees $\omega_1, \omega_2, ...\omega_n$, drawn independently according to the distribution imposed by p, we wish to estimate p.

In terms of the frequency function f, introduced in §1, the likelihood of the data is

$$\begin{aligned} L &= L(p; \omega_1, \omega_2, ...\omega_n) \\ &= \prod_{i=1}^{n} \prod_{(A \to \alpha) \in R} p(A \to \alpha)^{f(A \to \alpha; \omega_i)}. \end{aligned}$$

Recall the derivation of the *maximum likelihood* estimator of p: The log of the likelihood is

$$\sum_{A \in V} \sum_{\substack{\alpha \text{ s.t.} \\ (A \to \alpha) \in R}} \sum_{i=1}^{n} f(A \to \alpha; \omega_i) \log p(A \to \alpha). \tag{4.2}$$

The function $p : R \to [0,1]$ subject to (4.1) that maximizes (4.2) satisfies

$$\frac{\delta}{\delta p(B \to \beta)} \sum_{A \in V} \sum_{\substack{\alpha \text{ s.t.} \\ (A \to \alpha) \in R}} \{\lambda_A p(A \to \alpha) + \sum_{i=1}^{n} f(A \to \alpha; \omega_i) \log p(A \to \alpha)\} = 0$$

$\forall (B \to \beta) \in R$ where $\{\lambda_A\}_{A \in V}$ are Lagrange multipliers. Denote the maximum likelihood estimator by $\hat{p}$:

$$\begin{aligned} \lambda_B + \frac{\sum_{i=1}^{n} f(B \to \beta; \omega_i)}{\hat{p}(B \to \beta)} &= 0 \quad \forall (B \to \beta) \in R \\ &\Longrightarrow (\text{Since } \sum_{\substack{\beta \text{ s.t.} \\ (B \to \beta) \in R}} \hat{p}(B \to \beta) = 1) \end{aligned}$$

$$\hat{p}(B \to \beta) = \frac{\sum_{i=1}^{n} f(B \to \beta; \omega_i)}{\sum_{\substack{\alpha \text{ s.t.} \\ (B \to \alpha) \in R}} \sum_{i=1}^{n} f(B \to \alpha; \omega_i)}. \tag{4.3}$$

The maximum likelihood estimator is the natural, "relative frequency," estimator.

Suppose $B \in V$ is unobserved among the parse trees $\omega_1, \omega_2, ... \omega_n$. Then we can assign $\hat{p}(B \to \beta)$ arbitrarily, requiring only that (4.1) be respected. Evidently the likelihood is unaffected by the particular assignment of $\hat{p}(B \to \beta)$. Furthermore, it is not hard to see that any such B has probability zero of arising in any derivation that is based upon the maximum-likelihood probabilities[2]—hence the issue of tightness is independent of this assignment.

We will show that if Ω is the set of all (finite) parse trees generated by G, and if $\hat{p}(\omega)$ is the probability of $\omega \in \Omega$ under the maximum-likelihood production probabilities, then $\hat{p}(\Omega) = 1$.

The E-M Algorithm. Usually the derivation trees are unobserved—the sample, or corpus, contains only the yields $Y(\omega_1), Y(\omega_2), ... Y(\omega_n)$ ($Y(\omega_i) \in T^*$ for each $1 \leq i \leq n$). The likelihood is substantially more complex, since $p(Y(\omega))$ is now a marginal probability; we need to sum over the set of $\omega \in \Omega$ that yield $Y(\omega)$:

$$p(Y(\omega)) = \sum_{\substack{\omega' \in \Omega \text{ s.t.} \\ Y(\omega') = Y(\omega)}} p(Y(\omega')).$$

In the case where only yields are observed, the treatment is complicated considerably by the possibility of null productions ($A \to \emptyset$) and unit productions ($A \to B \in V$). If, however, the language of the grammar does not include the null string, then there is an equivalent grammar (one with the same language) that has no null productions and no unit productions (cf. Hopcroft & Ullman [7], Theorem 4.4). It is, then, perhaps best to simplify the treatment by *assuming that there are no null or unit productions.* Therefore, when the corpus consists of yields only, we shall assume *a priori* a model free of null and unit productions, and study tightness for probabilities estimated under such a model. Based upon the results of Stolcke [8] it is likely that this restriction can be relaxed, but we have not pursued this.

Letting Ω_Y denote $\{\omega \in \Omega : Y(\omega) = Y\}$, the likelihood of the corpus becomes

$$\prod_{i=1}^{n} \sum_{\omega \in \Omega_{Y(\omega_i)}} \prod_{(A \to \alpha) \in R} p(A \to \alpha)^{f(A \to \alpha; \omega)}.$$

And the maximum-likelihood equation becomes

$$\lambda_B + \frac{1}{\hat{p}(B \to \beta)} \sum_{i=1}^{n} \frac{\sum_{\omega \in \Omega_{Y(\omega_i)}} f(B \to \beta; \omega) \prod_{(A \to \alpha) \in R} \hat{p}(A \to \alpha)^{f(A \to \alpha; \omega)}}{\sum_{\omega \in \Omega_{Y(\omega_i)}} \prod_{(A \to \alpha) \in R} \hat{p}(A \to \alpha)^{f(A \to \alpha; \omega)}} = 0$$

[2] Consider any sequence of productions that leads from S to B. If the parent (antecedent) of B arose in the sample, then the last production has $\hat{p}$ probability zero and hence the sequence has probability zero. Otherwise, move "up" through the ancestors of B until finding the first variable in the S-to-B sequence represented in the sample (certainly S is represented). Apply the same reasoning to the production from that variable, and conclude that the given sequence has $\hat{p}$ probability zero.

$\Longrightarrow$

$$\hat{p}(B \to \beta) = \frac{\sum_{i=1}^{n} E_{\hat{p}}[f(B \to \beta; \omega) \mid \omega \in \Omega_{Y(\omega_i)}]}{\sum_{\substack{\alpha \text{ s.t.} \\ (B \to \alpha) \in R}} \sum_{i=1}^{n} E_{\hat{p}}[f(B \to \alpha; \omega) \mid \omega \in \Omega_{Y(\omega_i)}]} \tag{4.4}$$

where $E_{\hat{p}}$ is expectation under $\hat{p}$ and where "$\mid \omega \in \Omega_{Y(\omega_i)}$" means "conditioned on $\omega \in \Omega_{Y(\omega_i)}$."

There is no hope for a closed form solution, but (4.4) does suggest an iteration scheme which, as it turns out, "climbs" the likelihood surface (though there are no guarantees about approaching a *global* maximum): Let $\hat{p}_0$ be an arbitrary assignment respecting (4.1). Define a sequence of probabilities, $\hat{p}_n$, by the iteration

$$\hat{p}_{n+1}(B \to \beta) = \frac{\displaystyle\sum_{i=1}^{n} E_{\hat{p}_n}[f(B \to \beta; \omega) \mid \omega \in \Omega_{Y(\omega_i)}]}{\displaystyle\sum_{\substack{\alpha \text{ s.t.} \\ (B \to \alpha) \in R}} \sum_{i=1}^{n} E_{\hat{p}_n}[f(B \to \alpha; \omega) \mid \omega \in \Omega_{Y(\omega_i)}]} \tag{4.5}$$

The right hand side is manageable, as long as we can manageably compute all possible parses of a sentence (yield) $Y(\omega)$. (More efficient approaches exist—see [1].) This iteration procedure is an instance of the "EM Algorithm." Baum ([2]) first introduced it for hidden Markov models (regular grammars) and Baker ([1]) extended it to the problem addressed here (estimation for context-free grammars). Dempster, Laird, and Rubin ([4]) put the idea into a much more general setting and coined the term EM for "Expectation-Maximization." (The right hand side of (4.5) is computed using the *expected* frequencies under $\hat{p}_n$; $\hat{p}_{n+1}$ is then the *maximum*-likelihood estimator, treating the expected frequencies as though they were observed frequencies.)

The issue of tightness comes up again. We will show that $\hat{p}_n(\Omega) = 1$ for each $n > 0$.

4.3 Tightness of the Maximum-Likelihood Estimator

Given a context-free grammar $G = (V, T, R, S)$, let Ω be the set of finite parse trees, let $p : R \to [0, 1]$ be a system of production probabilities satisfying (4.1), and let $\omega_1, \omega_2, ...\omega_n$ be a set (sample) of finite parse trees $\omega_k \in \Omega$. For now, null and unit productions are permitted. Finally, let $\hat{p}$ be the maximum-likelihood estimator of p, as defined by (4.3). (See also the remarks following (4.3) concerning variables unobserved in $\omega_1, \omega_2, ...\omega_n$.) More generally, $\hat{p}$ will refer to the probability distribution on (possibly infinite) parse trees induced by the maximum-likelihood estimator.

Theorem 1. $\hat{p}(\Omega) = 1$.

Proof. Let $q_A = \hat{p}$(derivation tree rooted with A fails to terminate). We will show that $q_S = 0$ (i.e. derivation trees rooted with S always terminate).

For each $A \in V$, let $F(A; \omega)$ be the number of instances of A in ω and let $\tilde{F}(A; \omega)$ be the number of non-root instances of A in ω. Given $\alpha \in (V \cup T)^*$, let $n_A(\alpha)$ be the number of instances of A in the string α, and, finally, let α_i be the $i'th$ component of the string α. For any $A \in V$

$$\begin{aligned}
q_A &= \hat{p}\left(\bigcup_{B\in V} \bigcup_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} \bigcup_{\substack{i \text{ s.t.} \\ \alpha_i = B}} \{\alpha_i \text{ fails to terminate}\} \right) \\
&\leq \sum_{B\in V} \hat{p}\left(\bigcup_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} \bigcup_{\substack{i \text{ s.t.} \\ \alpha_i = B}} \{\alpha_i \text{ fails to terminate}\} \right) \\
&= \sum_{B\in V} \sum_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} \hat{p}(A \to \alpha) \hat{p}\left(\bigcup_{\substack{i \text{ s.t.} \\ \alpha_i = B}} \{\alpha_i \text{ fails to terminate}\} \right) \\
&\leq \sum_{B\in V} \sum_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} \hat{p}(A \to \alpha) n_B(\alpha) q_B \\
&= \sum_{B\in V} q_B \left\{ \frac{\sum_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} n_B(\alpha) \sum_{i=1}^n f(A \to \alpha; \omega_i)}{\sum_{\substack{\alpha \text{ s.t.} \\ (A\to\alpha)\in R}} \sum_{i=1}^n f(A \to \alpha; \omega_i)} \right\} \\
&= \sum_{B\in V} q_B \left\{ \frac{\sum_{i=1}^n \sum_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} n_B(\alpha) f(A \to \alpha; \omega_i)}{\sum_{i=1}^n \sum_{\substack{\alpha \text{ s.t.} \\ (A\to\alpha)\in R}} f(A \to \alpha; \omega_i)} \right\} \\
&= \sum_{B\in V} q_B \left\{ \frac{\sum_{i=1}^n \sum_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} n_B(\alpha) f(A \to \alpha; \omega_i)}{\sum_{i=1}^n F(A; \omega_i)} \right\}
\end{aligned}$$

$$\Longrightarrow q_A \sum_{i=1}^n F(A; \omega_i) \leq \sum_{B\in V} q_B \sum_{i=1}^n \sum_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} n_B(\alpha) f(A \to \alpha; \omega_i)$$

Sum over $A \in V$:

$$\begin{aligned}
\sum_{A\in V} q_A \sum_{i=1}^n F(A; \omega_i) \quad &\leq \sum_{B\in V} q_B \sum_{i=1}^n \sum_{A\in V} \sum_{\substack{\alpha \text{ s.t. } B\in\alpha \\ (A\to\alpha)\in R}} n_B(\alpha) f(A \to \alpha; \omega_i) \\
&= \sum_{B\in V} q_B \sum_{i=1}^n \tilde{F}(B; \omega_i)
\end{aligned}$$

i.e.

$$\sum_{A\in V} q_A \sum_{i=1}^n (\tilde{F}(A; \omega_i) - F(A; \omega_i)) \geq 0$$

Clearly, for every $i = 1, 2, ...n$ $\tilde{F}(A; \omega_i) = F(A; \omega_i)$ whenever $A \neq S$ and $\tilde{F}(S; \omega_i) < F(S; \omega_i)$. Hence $q_S = 0$, completing the proof of the theorem.

Now let $\hat{p}_n$ be the system of probabilities produced by the $n'th$ iteration of the EM Algorithm (4.5):

Corollary 1. If R contains no null productions and no unit productions, then $\hat{p}_n(\Omega) = 1\ \forall n \geq 1$.

Proof. Almost identical, except that we use (4.5) in place of (4.3) and end up with

$$\sum_{A \in V} q_A \sum_{i=1}^{n} E_{\hat{p}_{n-1}}[\tilde{F}(A;\omega_i) - F(A;\omega_i) \mid \omega \in \Omega_{Y(\omega_i)}] \geq 0. \tag{4.6}$$

In the absence of unit productions and null productions, $F(A;\omega) < 2|\omega|$ (twice the length of the string ω). Hence the expectations in (4.6) are finite. Furthermore, $\tilde{F}(A;\omega)$ and $F(A;\omega)$ satisfy the same conditions as before: $\tilde{F}(A;\omega) = F(A;\omega)$ except when $A = S$, in which case $\tilde{F}(A;\omega) < F(A;\omega)$. Again, we conclude that $q_S = 0$.

Bibliography

[1] J. K. Baker. Trainable Grammars for Speech Recognition. In *Speech Communications Papers of 97'th Meeting of Acoustical Society of America*, pages 547–550, Cambridge, Massachusetts. 1979.

[2] L. E. Baum. An Inequality and Associated Maximization Techniques in Statistical Estimation of Probabilistic Functions of Markov processes. *Inequalities*, 3:1–8. 1972.

[3] T. L. Booth and R. A. Thompson. Applying Probability Measures to Abstract Languages. *IEEE Trans. on Computers*, C-22:442–450. 1973.

[4] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *J. Royal Stat. Society, Ser. B*, 39:1–38. 1977.

[5] U. Grenander. *Lectures in Pattern Theory Volume 1, Pattern Synthesis.* Springer-Verlag, New York. 1976.

[6] T .E. Harris. *The Theory of Branching Processes*. Springer-Verlag, Berlin. 1963.

[7] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison Wesley, Reading, Massachusetts. 1979.

[8] A. Stolcke. An Efficient Probabilistic Context-Free Parsing Algorithm that Computes Prefix Probabilities. *Association for Computational Linguistics*, 21:165–201. 1995.

[9] C. S. Wetherell. Probabilistic Languages: A Review and Some Open Questions. *Computing Surveys*, 12:361–379. 1980.

Chapter 5

Statistical Properties of Probabilistic Context-Free Grammars

In this chapter we collect and prove an array of useful results about probabilistic context-free grammars (PCFGs) and their Gibbs representations. We present a method to generate production probabilities for context-free grammars (CFGs) which always impose consistent probability distributions on the finite parse trees. In addition, we demonstrate that these probability distributions have finite entropy, and under the distributions, the expected value of size of parse tree to any order is finite. We establish connections between PCFGs and Gibbs distributions on CFGs and prove the equivalence of the maximum-likelihood (ML) estimation methods for these two categories of probability distributions. We show how to "renormalize" an inconsistent PCFG so that it becomes a consistent PCFG. Finally, some minor issues, including the identifiability of parameters for PCFGs as well as for Gibbs distributions on CFGs are discussed.

5.1 Introduction

Finite parse trees, or parses, generated by a context-free grammar (CFG), can be equipped with a variety of probability distributions. The simplest way to do this is by production probabilities. Firstly, for each non-terminal symbol in the CFG, a probability distribution is placed on the set of all productions from that symbol. Then each finite parse tree is allocated a probability equal to the product of the probabilities of all productions in the tree. Denote a finite parse tree by ω. For any production rule $A \to \alpha$ of the CFG, let $f(A \to \alpha; \omega)$ be the number of times that this production rule is applied in ω. Let D be the set of all production rules. Then

$$p(\omega) = \prod_{(A\to\alpha)\in D} p(A \to \alpha)^{f(A\to\alpha;\omega)}.$$

[1] Supported by the Army Research Office (DAAL03-92-G-0115), the National Science Foundation (DMS-9217655), and the Office of Naval Research (N00014-96-1-0647).

A CFG with a probability distribution on all its parses assigned by this procedure is called a probabilistic context-free grammar (PCFG) (Booth & Thompson [6], Grenander [10]). It is well known that a PCFG may be inconsistent, i.e., the total probability of all finite parse trees is less than one[1]. It has been shown, however, that production probabilities estimated by the maximum-likelihood (ML) estimation procedure (or the relative frequency estimation procedure, as called in computational linguistics) always impose consistent probability distributions on finite parse trees (Chi & Geman [7]). In this chapter we generalize this result to a simple procedure, called "relative weighted frequency" method, which always generates production probabilities that impose consistent probability distributions on parses.

In addition to consistency, there are several important aspects of PCFGs. One of them is the entropy of probability distributions on parses (Mark *et al* [13], Mark [15], Miller *et al* [16]). We will demonstrate that if a probability distribution on parses is imposed by production probabilities estimated by the ML estimation procedure, or more generally, generated via the relative weighted frequency method, then it has finite entropy. Our proof for this result also derives the closed from of the entropy. Furthermore, we will show that, under such distributions, size of parse tree has finite momentum of any order.

PCFGs are indeed random branching processes, therefore their asymptotic behavior can be characterized by their branching rates. Using the notion of branching rate, Sánchez & Benedi [17] proved the consistency of distributions imposed by estimated production probabilities around the same time Chi & Geman got the same result. In this chapter, we will explore further the properties of branching rate.

Besides distributions imposed by production probabilities, parses of can be equipped with many types of probability distributions. Among the widely studied are Gibbs distributions (Abney [1], Mark *et al.* [13], [14], Mark [15]) . Gibbs distributions are deemed more useful than PCFG distributions in the sense that they provide better approximation to the actual distributions of languages, because they incorporate more features of parse trees than PCFGs, whose features only include frequencies of production rules. One the other hand, a Gibbs distribution degenerates into a PCFG if it only takes into account the same things as PCFGs, i.e., the frequencies of production rules in parses. More precisely, we will show that any CFG with a Gibbs distribution of the form

$$P_\lambda(\omega) = \frac{1}{Z_\lambda} \prod_{(A\to\alpha)\in D} e^{\lambda_{A\to\alpha} f(A\to\alpha;\omega)}. \tag{5.1}$$

is indeed a PCFG, where Z_λ is the partition number of the Gibbs distribution.

On the other hand, although PCFG distributions are Gibbs distributions, as can be easily seen, one still can not put PCFGs into the category of CFGs equipped with Gibbs distributions if the ML estimation procedures for these two types of distributions are different. Indeed, as will be seen, numerically these two estimation procedures are different. However, we will show that they are *equivalent* in the sense that the estimates by the two procedures impose the same distributions. For this reason, we can assure ourselves that Gibbs distributions can be considered as a generalization of PCFGs.

The fact that Gibbs distributions of the form (5.1) are imposed by production probabilities

[1] In probability theory, such probability distributions on finite parse trees are called being not tight, or improper

has a useful consequence. We often come up with inconsistent PCFGs, i.e.,

$$\sum_{\omega} p(\omega) < 1.$$

Writing the sum of the left hand side as Z, we "renormalize" the inconsistent distribution p by assigning to each parse tree a new probability equal to $p(\omega)/Z$. What (5.1) implies is, this renormalization procedure gives rise to a consistent distribution $\tilde{p}$ on parses which is imposed by a set of production probabilities $\{\tilde{p}(A \to \alpha)\}_{(A\to\alpha)\in D}$, i.e.,

$$\tilde{p}(\omega) = \prod_{(A\to\alpha)\in D} \tilde{p}(A \to \alpha)^{f(A\to\alpha;\omega)}. \tag{5.2}$$

In addition, the production probabilities can be written out explicitly. Moreover, we will show that, under a certain condition, $\tilde{p}$ has has finite entropy.

Finally, we will discuss the identifiability of production probabilities of PCFGs and their counterparts in Gibbs distributions. Briefly speaking, in PCFGs, two different systems of production probabilities always impose different distributions on parses. However, in Gibbs distributions, there can be infinitely many different sets of parameters which impose the same distribution. Besides the results on identifiability, we will establish a relation between the production probabilities and the expected values of frequencies of productions.

This chapter proceeds as follows. In §5.2, we gather the notations for PCFGs that will be used later on in the chapter. In §5.3, the ML estimation schemes for PCFGs are briefly reviewed. After that, the "relative weighted frequency" method is established. In §5.4, we study the entropy and the statistical properties of size of parse tree of PCFGs. In §5.5, we make connections between CFGs equipped with Gibbs distributions and PCFGs. Renormalization of inconsistent PCFGs is also discussed here. In §5.6, PCFGs are studied from the random branching process point of view. Finally, in §5.7, some minor issues, including consistency of the ML estimators, and identifiability of production probabilities are addressed.

5.2 Notations and Definitions

A context-free grammar G is a quadruple (V, T, D, S), where S is the start symbol, V the set of variables, T the set of terminals, and D the set of production rules. S and elements of V are also called non-terminal symbols. V, T and D are always assumed to be finite. Let $V^+ = V \cup \{S\}$. Let Ω denote the set of finite parse trees of G. ω will always denote a finite parse tree. For each $\omega \in \Omega$ and each production rule $(A \to \alpha) \in D$, define $f(A \to \alpha; \omega)$ as the number of occurrences of the rule in ω. Define $h(\omega)$ as the "height" of ω, i.e. the total number of non-terminal nodes on the longest route from ω's root to its terminals. Define $|\omega|$ as the "size" of ω, i.e., the total number of non-terminal nodes in ω. For any $A \in V$ and any string $\gamma \in (V \cup T)^*$, define $n(B;\ \gamma)$ as the number of instances of B in γ and define $|\gamma|$ as the length of the string.

For any two symbols $A, B \in V^+$, not necessarily different, B is said to be reachable from A in G, if there are symbols $A_0 = A$, A_1, …, $A_n = B$ in V^+ and strings α_0, …, α_{n-1} in

$(V \cup T)^*$ such that $(A_i \to \alpha_i) \in D$ and $A_{i+1} \in \alpha$ for each i. G is called connected if any symbol can be reached from any non-terminal symbol.

A system of production probabilities of G is a function $p : D \to [0, 1]$ such that for any $A \in V^+$,

$$\sum_{\substack{\alpha \in (V \cup T)^* \\ \text{s.t. } (A \to \alpha) \in D}} p(A \to \alpha) = 1. \tag{5.3}$$

We will use the same notation p to represent the probability distribution on parse trees imposed by p. Similarly, for any estimated system of production probabilities $\hat{p}$, we will use the same notation $\hat{p}$ to represent the probability distribution on parse trees imposed by $\hat{p}$. We will write $p(\Omega)$ as the total probability of all finite parse trees in Ω.

Besides probability distributions imposed by production probabilities, Ω can have other kinds of probability distributions. Let p be an arbitrary distribution on Ω. If $g(\omega)$ is a function of ω, then $E_p g(\omega)$ is defined as the expectation of $g(\omega)$ under the distribution p, i.e.

$$E_p g(\omega) = \sum_{\omega \in \Omega} p(\omega) g(\omega).$$

We just defined reachability of one symbol from another one in a CFG. We now define reachability in a CFG whose language is equipped with a distribution p. For any two symbols $A, B \in V^+$, not necessarily different, B is said to be reachable from A in G under p, if there is an $\omega \in \Omega$, such that ω contains a subtree with A as its root and this subtree contains an instance of B. G is called connected under p if any symbol can be reached from any non-terminal symbol under p.

So far we have talked about parse trees in the language and their probabilities. All these parse trees have S as their root. It is often useful to examine the subtrees of parse trees. Therefore it is necessary to consider trees with roots other than S. We call a tree with root $A \in V$ a parse tree with root A if it is generated by the production rules in D. Let Ω_A be the set of all finite parse trees with root A. Define p_A as the probability distribution on Ω_A imposed by the system of production probabilities p. Also extend the definition of "height" and "size" to trees in Ω_A.

When we write $p_A(\omega)$, we always assume that ω is a parse tree with A as its root. When $p = p_A$, $E_p g(\omega)$ means

$$E_p g(\omega) = \sum_{\omega \in \Omega_A} p_A(\omega) g(\omega).$$

We will use $p(\Omega_A)$ instead of $p_A(\Omega_A)$ to represent the total probability of finite parse trees in Ω_A.

When Ω and p appear without subscripts, they always mean Ω_S and p_S.

For convenience, we also extend the definition of trees to terminals. For each terminal $\tau \in T$, define Ω_τ as the set of a single "tree" $\{\tau\}$. Define $p_\tau(\tau) = 1$, $|\tau| = 0$ and $h(\tau) = 0$.

For this paper we make the following assumptions:

1. For any $A \in V$, there is an $\omega \in \Omega$ such that A appears in ω. This is reasonable. Because if no finite parse tree contains A, then A is useless in the grammar and can be removed from V.

This assumption implies two things. Firstly, any $A \in V$ is reachable by S. Secondly, for each $A \in V^+$, the total probability of finite parse trees with root A is positive.

2. When a system of production probabilities p appears in the context but not assigned, we always assume $p(A \to \alpha) > 0$ for each production rule $(A \to \alpha) \in D$. This is also reasonable. Because if $p(A \to \alpha) = 0$, then the rule $A \to \alpha$ can never appear in the PCFG and hence can be removed from D.

5.3 Maximum-likelihood Estimations for PCFGs

We consider two cases of the ML estimation. In the first case, we assume the data are *fully observed*. This means that all the parse trees are observed. Let $\omega_1, \omega_2, \ldots, \omega_n$ be a set of observed finite parse trees. Then the ML estimator of p, $\hat{p}$ can be shown to have the form

$$\hat{p}(B \to \beta) = \frac{\sum_{i=1}^{n} f(B \to \beta; \omega_i)}{\sum_{\substack{\alpha \text{ s.t.} \\ (B \to \alpha) \in D}} \sum_{i=1}^{n} f(B \to \alpha; \omega_i)}. \tag{5.4}$$

Because of (5.4), the ML estimator in the full observation case is also called the relative frequency estimator in computational linguistics. This simple estimator, as shown in [7], produces consistent probability distributions on the language.

In the second case, the parse trees are unobserved. Instead, the yields $Y(\omega_1), Y(\omega_2), \ldots$ $Y(\omega_n)$, which are the left-to-right sequences of terminals of $\omega_1, \ldots, \omega_n$, form the data. It can be proved that the ML estimator $\hat{p}$ should satisfy

$$\hat{p}(B \to \beta) = \frac{\sum_{i=1}^{n} E_{\hat{p}}[f(B \to \beta; \omega) | \omega \in \Omega_{Y(\omega_i)}]}{\sum_{\substack{\alpha \text{ s.t.} \\ (B \to \alpha) \in D}} \sum_{i=1}^{n} E_{\hat{p}}[f(B \to \beta; \omega) | \omega \in \Omega_{Y(\omega_i)}]}, \tag{5.5}$$

where $\Omega_Y = \{\omega \in \Omega : \; Y(\omega) = Y\}$.

Equation (5.5) can not be solved in closed form. Usually, the solution is computed by EM algorithm with the following iteration (Baum [3], Baker [2], Dempster *et al.* [8])

$$\hat{p}_{n+1}(B \to \beta) = \frac{\displaystyle\sum_{i=1}^{n} E_{\hat{p}_n}[f(B \to \beta; \omega) | \omega \in \Omega_{Y(\omega_i)}]}{\displaystyle\sum_{\substack{\alpha \text{ s.t.} \\ (B \to \alpha) \in D}} \sum_{i=1}^{n} E_{\hat{p}_n}[f(B \to \beta; \omega) | \omega \in \Omega_{Y(\omega_i)}]}. \tag{5.6}$$

Like $\hat{p}$ in (5.4), $\hat{p}_n$ for $n > 0$ imposes a consistent probability distribution on Ω ([7]).

We can write both (5.4) and (5.6) in the same form. Indeed, $\hat{p}$ in (5.4) and $\hat{p}_{n+1}$ in (5.6) can be written as

$$\hat{p}_{n+1}(B \to \beta) = \frac{\displaystyle\sum_{i=1}^{n} E_{\hat{p}_n}[f(B \to \beta; \omega) | \omega \in S(\omega_i)]}{\displaystyle\sum_{\substack{\alpha \text{ s.t.} \\ (B \to \alpha) \in D}} \sum_{i=1}^{n} E_{\hat{p}_n}[f(B \to \beta; \omega) | \omega \in S(\omega_i)]}. \tag{5.7}$$

In the above expression, $S(\omega)$ is a finite subset of Ω which depends on ω. In the case of full observations, $p_1 = p_2 = \ldots = \hat{p}$ and $S(\omega) = \{\omega\}$. In the case of EM algorithm, $S(\omega) = \Omega_{Y(\omega)}$.

An important observation is that all the $S(\omega)$ form a *partition* of Ω, i.e., a family $\{S_i\}$ of finite subsets of Ω, such that (i) $\cup_i S_i = \Omega$ (ii) either $S_i = S_j$ or $S_i \cap S_j = \emptyset$, for any i and j. $S(\omega)$ can then be thought as the unique S_i such that $\omega \in S_i$.

In image processing, the estimation (5.7) has been used for a long time and called *pseudo-likelihood estimation* (Besag [4][5], Geman & Graffigne [9]). Even this estimation is not the most general form for our purpose to generate production probabilities that impose consistent distributions on languages. Note that, in (5.7), the number of involved ω's is finite. The numerator and the denominator are weighted sums of the frequencies of the production rules in ω's. Let us forget about the iterations involved in the formula as well as the fact that all the ω's come from observed data. We just pick an arbitrary finite subset Λ of Ω, as long as every production rule appears in one of the trees in Λ, and an arbitrary weight distribution $\{W(\omega)\}_{\omega \in \Lambda}$ on Λ such that $W(\omega) > 0$ for each $\omega \in \Lambda$ and, for simplicity, $\sum_{\omega \in \Lambda} W(\omega) = 1$. Then we define a system of production probabilities by

$$\hat{p}(B \to \beta) = \frac{\displaystyle\sum_{\omega \in \Lambda} f(B \to \beta; \omega) W(\omega)}{\displaystyle\sum_{\substack{\alpha \text{ s.t.} \\ (B \to \alpha) \in D}} \sum_{\omega \in \Lambda} f(B \to \beta; \omega) W(\omega)}. \tag{5.8}$$

By the similarity between (5.4) and (5.8), we call the method to assign production probabilities by (5.8) the "relative weighted frequency" method.

Proposition 7. Production probabilities given by (5.8) always impose a consistent distribution on Ω.

The direct proof of Proposition Proposition 7 is almost identical to that in [7] and hence is omitted. In §5.6, Proposition Proposition 7 will be a consequence of Proposition Proposition 11.

5.4 Entropy and Expected Size of Parse Tree

In this section, we will show that $\hat{p}$ given by (5.8) impose a probability distribution on Ω such that Ω has finite entropy and the expected value of the size of parse tree in Ω to an arbitrary fixed power is finite.

To make the following proofs more readable, we define, for any given $\Lambda = \{\omega_1, \ldots, \omega_n\}$,

$$\begin{aligned} F(A \to \alpha) &= \sum_{\omega \in \Lambda} f(A \to \alpha; \omega) W(\omega), \quad \text{for any } A \to \alpha \in D \\ F(A) &= \sum_{\substack{\alpha \in (V \cup T)^* \\ \text{s.t. } (A \to \alpha) \in D}} F(A \to \alpha), \quad \text{for any } A \in V^+, \end{aligned}$$

i.e., $F(A \to \alpha)$ is the total weighted number of instances of the production rule $(A \to \alpha)$ in ω's and $F(A)$ is the total weighted number of instances of the symbol A in ω's.

The relative weighted frequency method given by (5.8) then can be written as

$$\hat{p}(B \to \beta) = \frac{F(B \to \beta)}{F(B)} \tag{5.9}$$

Then we have the following simple lemma

Lemma 9. For any $A \in V^+$,

$$\sum_{A \in V^+} F(A) = \sum_{\omega \in \Lambda} |\omega| W(\omega) \tag{5.10}$$

and

$$\sum_{B \in V^+} \sum_{\substack{\gamma \text{ s.t.} \\ A \in \gamma \\ (B \to \gamma) \in D}} F(B \to \gamma) n(A;\ \gamma) = \begin{cases} F(S) - 1 & \text{if } A = S \\ F(A) & \text{if } A \neq S \end{cases} \tag{5.11}$$

Remark 6. If $\sum_{\omega \in \Lambda} W(\omega) \neq 1$, $F(S) - 1$ should be changed to $F(S) - \sum_{\omega \in \Lambda} W(\omega)$.

Proof. For the first equation,

$$\begin{aligned} \sum_{A \in V^+} F(A) &= \sum_{A \in V^+} \sum_{\substack{\alpha \in (V \cup T)^* \\ \text{s.t. } (A \to \alpha) \in D}} \sum_{\omega \in \Lambda} f(A \to \alpha; \omega) W(\omega) \\ &= \sum_{\omega \in \Lambda} \sum_{A \in V^+} \sum_{\substack{\alpha \in (V \cup T)^* \\ \text{s.t. } (A \to \alpha) \in D}} f(A \to \alpha; \omega) W(\omega) \\ &= \sum_{\omega \in \Lambda} |\omega| W(\omega) \end{aligned}$$

For the second one,

$$\begin{aligned}
&\sum_{B\in V^+} \sum_{\substack{\gamma \text{ s.t.} \\ A\in\gamma \\ (B\to\gamma)\in D}} F(B\to\gamma) n(A;\ \gamma) \\
=\ & \sum_{B\in V^+} \sum_{\substack{\gamma \text{ s.t.} \\ A\in\gamma \\ (B\to\gamma)\in D}} \sum_{\omega\in\Lambda} f(B\to\gamma;\omega) W(\omega) n(A;\ \gamma) \\
=\ & \sum_{\omega\in\Lambda} W(\omega) \sum_{B\in V^+} \sum_{\substack{\gamma \text{ s.t.} \\ A\in\gamma \\ (B\to\gamma)\in D}} f(B\to\gamma;\omega) n(A;\ \gamma) \qquad (5.12)
\end{aligned}$$

For each A,

$$\sum_{B\in V^+} \sum_{\substack{\gamma \text{ s.t.} \\ A\in\gamma \\ (B\to\gamma)\in D}} f(B\to\gamma;\omega) n(A;\ \gamma)$$

is the number of non-root instances of A in ω. When $A \neq S$, the number of non-root instances of A in ω is the total number of instances of A in ω, the latter one being equal to

$$\sum_{\substack{\alpha\in(V\cup T)^* \\ \text{s.t. } (A\to\alpha)\in D}} f(A\to\alpha;\omega).$$

Substitute this into (5.12) to prove (5.11) for the case $A \neq S$. The case $A = S$ is similarly proved. □

Proposition 8. Given $\Lambda = \{\omega_1, \omega_2, \ldots, \omega_n\}$, the estimated $\hat{p}$ satisfies

$$E_{\hat{p}} f(A\to\alpha;\omega) = \sum_{\omega\in\Lambda} f(A\to\alpha;\omega) W(\omega) \qquad (5.13)$$

for any $(A\to\alpha)\in D$. Therefore, the expected number of instances of the rule $A\to\alpha$ in all parse trees of Ω equals the sum of weighted numbers of instances of $A\to\alpha$ in $\omega_1, \ldots, \omega_n$.

Proof. Fix $(A\to\alpha)\in D$. For each $C\in V^+$ and $k\in\mathbf{N}$, define

$$E_{k,C} = \sum_{\substack{\omega\in\Omega_C \\ h(\omega)\le k}} \hat{p}_C(\omega) f(A\to\alpha;\omega).$$

Define

$$\chi(C\to\gamma) = \begin{cases} 0 & \text{if } C\to\gamma \neq A\to\alpha, \\ 1 & \text{otherwise.} \end{cases}$$

For each $\omega\in\Omega_C$, let $\gamma\in(V\cup T)^*$ be the string such that $C\to\gamma$ is the first production rule that ω applies. Then for each variable in γ, a subtree of ω with this variable as its root is independently generated. Thus, by simple computation,

$$E_{k+1,C} = \sum_{\substack{\gamma\in(V\cup T)^* \\ \text{s.t. } (C\to\gamma)\in D}} \hat{p}(C\to\gamma)(\chi(C\to\gamma) + \sum_{\substack{B\in V^+ \\ \text{s.t. } B\in\gamma}} n(B;\ \gamma) E_{k,B}).$$

By (5.9),

$$E_{k+1,C} = \frac{1}{F(C)} \sum_{\substack{\gamma \in (V \cup T)^* \\ \text{s.t. } (C \to \gamma) \in D}} F(C \to \gamma)(\chi(C \to \gamma) + \sum_{\substack{B \in V^+ \\ \text{s.t. } B \in \gamma}} n(B;\ \gamma) E_{k,B}),$$

hence

$$F(C) E_{k+1,C} = \sum_{\substack{\gamma \in (V \cup T)^* \\ \text{s.t. } (C \to \gamma) \in D}} F(C \to \gamma)(\chi(C \to \gamma) + \sum_{\substack{B \in V^+ \\ \text{s.t. } B \in \gamma}} n(B;\ \gamma) E_{k,B})$$

Summing over all $C \in V^+$,

$$\begin{aligned} & \sum_{C \in V^+} F(C) E_{k+1,C} \\ = & \sum_{C \in V^+} \sum_{\substack{\gamma \in (V \cup T)^* \\ \text{s.t. } (C \to \gamma) \in D}} F(C \to \gamma) \chi(C \to \gamma) \\ & + \sum_{C \in V^+} \sum_{\substack{\gamma \in (V \cup T)^* \\ \text{s.t. } (C \to \gamma) \in D}} F(C \to \gamma) \sum_{\substack{B \in V^+ \\ \text{s.t. } B \in \gamma}} n(B;\ \gamma) E_{k,B} \\ = & F(A \to \alpha) + \sum_{B \in V^+} E_{k,B} \sum_{C \in V^+} \sum_{\substack{\gamma \text{ s.t.} \\ B \in \gamma \\ (C \to \gamma) \in D}} F(C \to \gamma) n(B;\ \gamma). \end{aligned}$$

By (5.11)

$$\sum_{C \in V^+} F(C) E_{k+1,C} = F(A \to \alpha) + \sum_{B \in V^+} E_{k,B} F(B) - E_{k,S}.$$

Hence

$$\sum_{C \in V^+} F(C)(E_{k+1,C} - E_{k,C}) = F(A \to \alpha) - E_{k,S}. \tag{5.14}$$

Obviously, $E_{k,C}$ is a sequence increasing in k. Then the left hand side of (5.14) is larger or equal to 0, which implies

$$E_{k,S} \leq F(A \to \alpha).$$

This implies that

$$E_{\hat{p}} f(A \to \alpha; \omega) = \lim_{k \to \infty} E_{k,S}$$

exists. Now take limits on both sides of (5.14) to get

$$E_{\hat{p}} f(A \to \alpha; \omega) = F(A \to \alpha),$$

which is just (5.13). □

Corollary 2. $H(\hat{p}) < \infty$, where $H(p)$ is the entropy of a distribution p.

Proof. By changing the order of summations, we first get

$$\begin{aligned} H(\hat{p}) &= -\sum_{\omega \in \Omega} \hat{p}(\omega) \sum_{(A\to\alpha)\in D} f(A \to \alpha) \log \hat{p}(A \to \alpha) \\ &= \sum_{(A\to\alpha)\in D} E_{\hat{p}} f(A \to \alpha; \omega) \log \frac{1}{\hat{p}(A \to \alpha)}. \end{aligned}$$

Then by Proposition Proposition 8 and (5.9), the summation equals

$$\sum_{A \in V^+} F(A) \log F(A) - \sum_{(A\to\alpha)\in D} F(A \to \alpha) \log F(A \to \alpha).$$

□

We will next show that the expected value of $|\omega|^k$ is finite for any $k \geq 0$ if ω is distributed via $\hat{p}$. Before demonstrating this result, we first note that, because of the assumptions we made in §5.2 about G, every variable $A \neq S$ can be reached from S under the distribution $\hat{p}$.

Proposition 9. For each $m \in \mathbf{N} \cup \{0\}$,

$$E_{\hat{p}}|\omega|^m < \infty. \tag{5.15}$$

Proof. In fact, we shall show that for any $A \in V^+$, if $p = \hat{p}_A$, then $E_p|\omega|^m < \infty$. When $m = 0$, this is clearly true. Now suppose the claim is true for $0, \ldots, m-1$. For each $A \in V^+$ and $k \in \mathbf{N}$, define

$$M_{k,A} = \sum_{\substack{\omega \in \Omega_A \\ h(\omega) \leq k}} \hat{p}_A(\omega)|\omega|^m.$$

Clearly, $M_{k,A}$ is increasing in k. It is easy to check

$$M_{k+1,A} = \sum_{\substack{\alpha \in (V \cup T)^* \\ (A\to\alpha)\in D}} \sum_{\substack{\omega_1,\ldots,\omega_L \\ \omega_i \in \Omega_{\alpha_i} \\ h(\omega_i) \leq k}} (1 + \sum_{i=1}^{L} |\omega_i|)^m \hat{p}(A \to \alpha) \hat{p}_{\alpha_1}(\omega_1) \ldots \hat{p}_{\alpha_L}(\omega_L), \tag{5.16}$$

where for clarity, we write L for $|\alpha|$. For fixed α, write

$$(1 + \sum_{i=1}^{L} |\omega_i|)^m = P(|\omega_1|, \ldots, |\omega_L|) + \sum_{i=1}^{L} |\omega_i|^m.$$

P is a polynomial in $|\omega_1|, \ldots, |\omega_L|$, each of whose terms is of the form

$$|\omega_1|^{s_1} |\omega_2|^{s_2} \ldots |\omega_L|^{s_L},\ 0 \leq s_i < m,\ s_1 + s_2 + \ldots s_L \leq m. \tag{5.17}$$

By induction hypothesis, there is a $C > 1$, such that for any $0 \leq s < m$ and $A \in V^+ \cup T$,

$$\sum_{\omega \in \Omega_A} \hat{p}_A(\omega)|\omega|^s = E_{\hat{p}_A}|\omega|^s < C.$$

Then for each term with the form (5.17),

$$\sum_{\substack{\omega_1,\dots,\omega_L\\ \omega_i\in\Omega_{\alpha_i}}} |\omega_1|^{s_1}\dots|\omega_L|^{s_L}\hat{p}_{\alpha_1}(\omega_1)\dots\hat{p}_{\alpha_L}(\omega_L)\le C^L.$$

For each α, there are less than $(L+1)^m = (|\alpha|+1)^m$ terms in $P(|\omega_1|,\dots,|\omega_{|\alpha|}|)$. Hence

$$\begin{aligned} M_{k+1,A} &\le \sum_{\substack{\alpha\in(V\cup T)^*\\ \text{s.t. } (A\to\alpha)\in D}} (|\alpha|+1)^m C^{|\alpha|}\hat{p}(A\to\alpha) \\ &+ \sum_{\substack{\alpha\in(V\cup T)^*\\ \text{s.t. } (A\to\alpha)\in D}} \sum_{\substack{\omega_1,\dots,\omega_{|\alpha|}\\ \omega_i\in\Omega_{\alpha_i}\\ h(\omega_i)\le k}} \sum_{i=1}^{|\alpha|} |\omega_i|^m\hat{p}(A\to\alpha)\hat{p}_{\alpha_1}(\omega_1)\dots\hat{p}_{\alpha_{|\alpha|}}(\omega_{|\alpha|}) \\ &= \sum_{\substack{\alpha\in(V\cup T)^*\\ \text{s.t. } (A\to\alpha)\in D}} (|\alpha|+1)^m C^{|\alpha|}\hat{p}(A\to\alpha) + \sum_{\substack{\alpha\in(V\cup T)^*\\ \text{s.t. } (A\to\alpha)\in D}} \sum_{i=1}^{|\alpha|} M_{k,\alpha_i}\hat{p}(A\to\alpha). \end{aligned}$$

Because the set of production rules is finite,

$$\sup\{|\alpha|:\ \text{ for some } A\in V^+,\ (A\to\alpha)\in D\}<\infty.$$

Therefore we can bound $(|\alpha|+1)^m C^{|\alpha|} = (|\alpha|+1)^m C^{|\alpha|}$ by an even larger number, say K. Then we get

$$M_{k+1,A}\le K + \sum_{\substack{\alpha\in(V\cup T)^*\\ \text{s.t. } (A\to\alpha)\in D}} \sum_{i=1}^{|\alpha|} M_{k,\alpha_i}\hat{p}(A\to\alpha). \tag{5.18}$$

The method in the proof of Proposition Proposition 8 can now be used. Multiply both sides of (5.18) by $F(A)$ and add up over all $A\in V^+$ with $F(A)>0$,

$$\sum_{A\in V^+} F(A)M_{k+1,A}\le K\sum_{\omega\in\Lambda}|\omega|W(\omega) + \sum_{A\in V^+}\sum_{\substack{\alpha\in(V\cup T)^*\\ \text{s.t. } (A\to\alpha)\in D}} \sum_{i=1}^{|\alpha|} M_{k,\alpha_i}F(A\to\alpha)$$

Then, as in the proof of Proposition Proposition 8, we get

$$M_{k,S}\le K\sum_{\omega\in\Lambda}|\omega|W(\omega)<\infty.$$

Letting $k\to\infty$, we see $M_{k,S}\uparrow E_{\hat{p}_S}|\omega|^m$. So we have proved $E_{\hat{p}_S}|\omega|^m < infty$. To complete the induction, we still need to show for every $A\in V\cup T$ other than S, $E_{\hat{p}_A}|\omega|^m<\infty$. For the case $V\in T$ this is obvious. For the case $A\in V$, if there is an $\alpha\in(V\cup T)^*$ such that $A\in\alpha$ and $(S\to\alpha)\in D$, then by (5.16),

$$\hat{p}(S\to\alpha)M_{k,A}\le M_{k+1,S}$$

Hence $E_{\hat{p}_A}|\omega|^m = \lim_k M_{k,A}<\infty$. In general, for any $A\in V$, there are $A_0=S, A_1,\dots A_n = A\in V^+$, and $\alpha_0,\dots,\alpha_{n-1}\in(V\cup T)^*$, such that $(A_i\to\alpha_i)\in D$ and $A_{i+1}\in\alpha_i$. We then get

$$\begin{aligned}
&\hat{p}(A_{n-1} \to \alpha)M_{k,A_n} \leq M_{k+1,A_{n-1}} \\
&\hat{p}(A_{n-2} \to \alpha)M_{k+1,A_{n-1}} \leq M_{k+2,A_{n-2}} \\
&\cdots\cdots \\
&\hat{p}(A_0 \to \alpha)M_{n+k,A_1} \leq M_{n+k+1,A_0}
\end{aligned}$$

Then by induction, $E_{\hat{p}_i}|\omega|^m < \infty$, $i = 0, \ldots, n-1$, where $\hat{p}_i = \hat{p}_{A_i}$. Hence $E_{\hat{p}_A}|\omega|^m < \infty$. Thus Proposition Proposition 9 is proved. □

5.5 CFGs with Gibbs Distributions and PCFGs

A Gibbs distribution on Ω has the form

$$\begin{aligned}
&P_\lambda(\omega) = \frac{e^{\lambda \cdot U(\omega)}}{Z_\lambda}, \\
&Z_\lambda = \sum_{\omega \in \Omega} e^{\lambda \cdot U(\omega)}, \\
&\lambda = \{\lambda_i\}_{i \in I},\ U(\omega) = \{U(\omega)_i\}_{i \in I}.
\end{aligned} \tag{5.19}$$

In the above expression, I is a finite index set, λ_i are constants, and $U_i(\omega)$ are functions on Ω. Z_λ is called the partition number because it makes P_λ a consistent distribution.

The functions $U_i(\omega)$ are usually considered as features of parse trees and the constants λ_i are parameters that weight these features. The index set I and the functions $U_i(\omega)$ can take various forms. The simplest choice for I is D, the set of production rules. Correspondingly, let

$$U(\omega) = f(\omega) = \{f(A \to \alpha; \omega)\}_{(A \to \alpha) \in D}. \tag{5.20}$$

If the constants $\lambda_{A \to \alpha}$ satisfy

$$Z_\lambda = \sum_{\omega \in \Omega} e^{\lambda \cdot f(\omega)} < \infty,$$

then we have a Gibbs distribution on Ω given by

$$P_\lambda(\omega) = \frac{e^{\lambda \cdot U(\omega)}}{Z_\lambda} = \frac{e^{\lambda \cdot f(\omega)}}{Z_\lambda}. \tag{5.21}$$

A consistent PCFG distribution is a Gibbs distribution of the form (5.21). To see this, let $\lambda_{A \to \alpha} = \log p(A \to \alpha)$ for each $(A \to \alpha) \in D$. Then

$$e^{\lambda \cdot U(\omega)} = \prod_{(A \to \alpha) \in D} p(A \to \alpha)^{f(A \to \alpha;\, \omega)}$$

and

$$Z_\lambda = \sum_{\omega \in \Omega} P_\lambda(\omega) = 1.$$

Then the PCFG probability distribution $p(\omega)$ imposed by $p(A \to \alpha)$ can be written as

$$p(\omega) = \prod_{(A\to\alpha)\in D} p(A \to \alpha)^{f(A\to\alpha;\,\omega)} = \frac{1}{Z_\lambda} e^{\lambda\cdot U(\omega)},$$

which is a Gibbs distribution.

We are interested in the inverse problem, i.e., is it true that the Gibbs distribution (5.21) is imposed by some production probabilities? As seen from the next proposition, the answer is positive. In other words, if the language generated by a CFG is equipped with a Gibbs distribution that only has frequencies of productions as its features, then this CFG is merely a PCFG.

Proposition 10. If every symbol in V of a CFG G can be reached from S under distribution (5.21), then G with distribution (5.21) is a probabilistic context-free grammar. That is, there are $p(A \to \alpha)$, such that for any $A \in V^+$,

$$\sum_{(A\to\alpha)\in D} p(A \to \alpha) = 1$$

and for every $\omega \in \Omega$,

$$P(\omega) = \prod_{(A\to\alpha)\in D} p(A \to \alpha)^{f(A\to\alpha;\omega)}.$$

Proof. Note that in (5.20), (5.21) and the above equation, ω's are parse trees in $\Omega = \Omega_S$. By obvious generalization, we can define $f(\omega)$ for $\omega \in \Omega_A$ and then define

$$Z_\lambda(A) = \sum_{\omega\in\Omega_A} e^{\lambda\cdot f(\omega)}$$

and $P_A(\omega)$. For simplicity, also define $Z_\lambda(\tau) = 1$ and $P_\tau(\tau) = 1$ for each $\tau \in T$.

We first need to show that if $Z_\lambda(S) = Z_\lambda < \infty$, then $Z_\lambda(A) < \infty$ for all A. Suppose $(S \to \alpha) \in D$. The sum of $e^{\lambda\cdot f(\omega)}$ over all $\omega \in \Omega$ with the first production applied being $S \to \alpha$ is $e^{\lambda_{S\to\alpha}} Z_\lambda(\alpha_1)\dots Z_\lambda(\alpha_n)$, where $n = |\alpha|$. Hence

$$Z_\lambda(S) \geq e^{\lambda_{S\to\alpha}} Z_\lambda(\alpha_1)\dots Z_\lambda(\alpha_n).$$

Each $Z_\lambda(\alpha_i)$ is positive, hence finite. Then each variable in α has finite Z_λ value. For any variable A, there are variables $A_0 = S, A_1, \dots, A_n = A \in V^+$ and $\alpha^{(0)}, \dots, \alpha^{(n-1)} \in (V \cup T)^*$, such that $(A_i \to \alpha^{(i)}) \in D$ and $A_{i+1} \in \alpha^{(i)}$. With the same argument as above, we get

$$Z_\lambda(A_i) \geq e^{\lambda_{A_i\to\alpha^{(i)}}} \prod_{k=1}^{|\alpha^{(i)}|} Z_\lambda(\alpha_k^{(i)}),$$

where $\alpha_k^{(i)}$ is the kth element in $\alpha^{(i)}$. Then by induction, $Z_\lambda(A) < \infty$.

Now for $A \to \alpha$, $\alpha \in (V \cup T)^*$, $|\alpha| = n$, define

$$p(A \to \alpha) = \frac{1}{Z_\lambda(A)} e^{\lambda_{A\to\alpha}} Z_\lambda(\alpha_1)\dots Z_\lambda(\alpha_n), \qquad (5.22)$$

Since $Z_\lambda(A)$ and all $Z_\lambda(\alpha_i)$ are finite, $p(A \to \alpha)$ is well defined.

Then

$$\sum_{(A\to\alpha)\in D} p(A \to \alpha) = \frac{1}{Z_\lambda(A)} \sum_{(A\to\alpha)\in D} e^{\lambda_{A\to\alpha}} \prod_{k=1}^{|\alpha|} Z_\lambda(\alpha_k) = \frac{1}{Z_\lambda(A)} \sum_{\omega\in\Omega_A} e^{\lambda\cdot f(\omega)} = 1$$

We prove

$$P(\omega) = \prod_{(A\to\alpha)\in D} p(A \to \beta)^{f(A\to\alpha;\omega)},$$

by induction on $h(\omega)$, the height of ω. When $h(\omega) = 0$, ω is just a terminal, therefore this is obvious. Suppose the equation is true for all $\omega \in \Omega_A$, $A \in V^+$, with $h(\omega) < h$. For any $\omega \in \Omega_A$ with $h(\omega) = h$, let $A \to \beta$ be the first production rule ω applies. Then

$$P_A(\omega) = \frac{1}{Z_\lambda(A)} e^{\lambda\cdot f(\omega)} = \frac{1}{Z_\lambda(A)} e^{\lambda_{A\to\beta}} \prod_{k=1}^{|\beta|} e^{\lambda\cdot f(\omega_k)},$$

where ω_k is the kth subtree of ω. Each ω_k has height $< h$. Hence, by induction assumption,

$$\frac{1}{Z_\lambda(\beta_k)} e^{\lambda\cdot f(\omega_k)} = \prod_{(B\to\alpha)\in D} p(B \to \alpha)^{f(B\to\alpha;\omega_k)}.$$

Then

$$\begin{aligned} P_A(\omega) &= \frac{1}{Z_\lambda(A)} e^{\lambda_{A\to\beta}} \prod_{k=1}^{|\beta|} Z_\lambda(\beta_k) \prod_{(B\to\alpha)\in D} p(B \to \alpha)^{f(B\to\alpha;\omega_k)} \\ &= p(A \to \beta) \prod_{k=1}^{|\beta|} \prod_{(B\to\alpha)\in D} p(B \to \alpha)^{f(B\to\alpha;\omega_k)} \\ &= \prod_{(B\to\alpha)\in D} p(B \to \alpha)^{f(B\to\alpha;\omega)}. \end{aligned}$$

Letting $A = S$, we then complete the proof. □

One of the most important issue about the Gibbs distribution (5.19) is the estimation of λ. Because of the easiness to apply various mathematical techniques to it, the ML estimation procedure is among the most widely used estimation procedures (Younes [20]). In the full observation case, the estimator $\hat{\lambda}$ is

$$\hat{\lambda} = \arg\max_{\lambda} \prod_{i=1}^{n} \frac{e^{\lambda\cdot U(\omega_i)}}{Z_\lambda}, \tag{5.23}$$

where $\omega_1, \ldots, \omega_n$ are the observed parse trees. In the partial observation case, the estimator $\hat{\lambda}$ is

$$\hat{\lambda} = \arg\max_{\lambda} \prod_{i=1}^{n} \sum_{\omega\in\Omega_{Y_i}} \frac{e^{\lambda\cdot U(\omega)}}{Z_\lambda}, \tag{5.24}$$

where $Y_1, \ldots, Y_n$ are the observed yields.

We have seen from Proposition Proposition 10 that when $U(\omega)$'s are the numbers of occurrences of production rules, the Gibbs distribution (5.21) is just PCFG distribution. However, if we want to put PCFGs under the framework of Gibbs distributions, we need further to show that the ML estimators for PCFGs are the same as the ML estimators for Gibbs distributions.

The ML estimators for a PCFG in the full observation case and the partial observation case are defined by (5.4) and (5.5), respectively. It is not obvious that the parameters estimated via (5.23) and the production probabilities estimated via (5.4) are the same. Nor it is obvious that the parameters estimated via (5.24) and the production probabilities estimated via (5.5) are the same. Indeed, numerically they are different. For example, the estimators (5.4) and (5.5) always give a single estimated system of production probabilities, while the estimators (5.23) and (5.24) may give us infinitely many solutions. We will discuss this in detail in §5.7.

Despite the numerical differences between the ML estimators for PCFGs and the ML estimators for Gibbs distributions, the following corollary to Proposition Proposition 10 shows the ML estimators of PCFG (5.4) and (5.5) are *equivalent* to (5.23) and (5.24), respectively, in the sense that they yield the same distributions on languages of CFGs. Because of this result, the theory of PCFGs can be entirely put into the framework of Gibbs distributions.

Corollary 3. The ML estimator (5.4) for a PCFG G is the exponential of *a* ML estimator for the Gibbs distribution (5.21) on Ω generated by G. Hence any estimated parameters by the ML estimation procedure for the Gibbs distribution (5.21) yield the same distributions as the estimated production probabilities by the ML estimation procedure (5.4) for the PCFG. The same relation holds for (5.5) and (5.24).

Proof. Suppose $\hat{\lambda}$ is a ML estimator for (5.21). Then by Proposition Proposition 10, the Gibbs distribution $P_{\hat{\lambda}}$ is imposed by some system of production probabilities $\hat{p}$. Clearly, $\hat{p}$ is the ML estimator for the PCFG G. It is also clear that $\tilde{\lambda}$, $\tilde{\lambda}(A \to \alpha) = \log \hat{p}(A \to \alpha)$ is another ML estimator for (5.21). Note that, even if $\hat{\lambda}$ and $\tilde{\lambda}$ are different, the Gibbs distributions they correspond to are the same. □

We have mentioned in several places that a system of production probabilities of a PCFG does not always impose a consistent distribution on the language. Suppose that a system of production probabilities $p(A \to \alpha)$, $(A \to \alpha) \in D$, does impose an inconsistent distribution on $\Omega = \Omega_S$. We then define a new distribution q on Ω, by

$$\tilde{p}(\omega) = \frac{p(\omega)}{p(\Omega)}, \quad \omega \in \Omega.$$

We call $\tilde{p}$ the *renormalized* distribution of p on Ω. More generally, we define renormalized distribution of p_A on Ω_A, for each $A \in V^+$, by

$$\tilde{p}_A(\omega) = \frac{p_A(\omega)}{p(\Omega_A)}, \quad \omega \in \Omega_A. \tag{5.25}$$

By Proposition Proposition 10, we have the following

Corollary 4. The renormalized distributions $\tilde{p}_A$ are induced by a single system of production probabilities

$$\tilde{p}(A \to \alpha) = \frac{1}{p(\Omega_A)} p(A \to \alpha) \prod_{B \in V} p(\Omega_B)^{n(B;\, \alpha)}. \tag{5.26}$$

Therefore, q on Ω is still a PCFG distribution.

Indeed, in (5.22), let $\lambda_{A \to \alpha} = \log p(A \to \alpha)$ for each production rule $A \to \alpha$, then $p(\Omega_A) = Z_\lambda(A)$ and (5.26) is exactly (5.22).

5.6 Branching Rates of PCFGs

In this section we consider the underlying branching processes of PCFGs and introduce the notion of branching rate. Adopting the set-ups in Miller & O'Sullivan [16] , we define the $V \times V$ mean matrix $\mathbf{M}$ of p, with (A, B)th entry $\mathbf{M}(A, B)$ being the expected number of variables B resulting from rewriting A:

$$\mathbf{M}(A, B) = \sum_{\substack{\alpha \in (V \cup T)^* \\ \text{s.t. } (A \to \alpha) \in D}} p(A \to \alpha) n(B; \alpha). \tag{5.27}$$

$\mathbf{M}$ is a non-negative matrix.

We say $B \in V^+$ can be reached from $A \in V^+$, if for some $n > 0$, $\mathbf{M}^{(n)}(A, B) > 0$, where $\mathbf{M}^{(n)}(A, B)$ is the (A, B)-th element of $\mathbf{M}^n$. $\mathbf{M}$ is irreducible if for any pair $A, B \in V^+$, B can be reached from A. The corresponding branching process is called connected if $\mathbf{M}$ is irreducible (Walters [19]).

Put in the context of PCFGs, that B can be reached from A means there are productions $A_0 \to \alpha_0$, $A_1 \to \alpha_1, \ldots$, $A_{n-1} \to \alpha_{n-1}$, such that $A_0 = A$, $A_{i+1} \in \alpha_i$, $A_n = B$, and $p(A_i \to \alpha_i) > 0$. The branching process is connected if any variable can be reached from any non-terminal variable in this way.

We need to use the following result

Theorem 2. (Perron-Frobenius Theorem) Let $\mathbf{M} = [m_{ij}]$ be a non-negative $k \times k$ matrix.

(1) There is a non-negative eigenvalue ρ such that no eigenvalue of A has absolute value greater than ρ.
(2) We have $\min_i(\sum_{j=1}^k m_{ij}) \leq \rho \max_i(\sum_{j=1}^k m_{ij})$.
(3) Corresponding to the eigenvalue ρ there is a non-negative left (row) eigenvector $\nu = (\nu_1, \ldots, \nu_k)$ and a non-negative right (column) eigenvector

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}.$$

(4) If $\mathbf{M}$ is irreducible then ρ is a simple eigenvalue and the corresponding eigenvectors are strictly positive (i.e. $u_i > 0$, $v_i > 0$ all i).

The eigenvalue ρ is called the branching rate of the branching process. A branching process is called sub-critical, critical, and super-critical, if $\rho < 1$, $\rho = 1$, and $\rho > 1$, respectively.

Using the notions for PCFGs, when a PCFGs underlying branching process is sub-critical, then the PCFG is consistent. When the underlying branching process is super-critical, then the PCFG is inconsistent.

The following proposition shows that the production probabilities p makes the underlying branching process of the PCFG sub-critical. This immediately leads to the consistency of the distribution on the language.

Proposition 11. For p given by (5.8) and $\mathbf{M}$ given by (5.27),

$$\rho < 1. \tag{5.28}$$

Proof. We have $\mathbf{M}\mu = \rho\mu$. Then for each variable A,

$$\sum_{B \in V} \mathbf{M}(A, B)\mu(B) = \rho\mu(A).$$

Therefore

$$\sum_{B \in V} \sum_{\substack{\alpha \in (V \cup T)^* \\ \text{s.t. } (A \to \alpha) \in D}} p(A \to \alpha) n(B; \alpha)\mu(B) = \rho\mu(A).$$

By (5.9),

$$\sum_{B \in V} \sum_{\substack{\alpha \in (V \cup T)^* \\ \text{s.t. } (A \to \alpha) \in D}} \frac{F(A \to \alpha)}{F(A)} n(B; \alpha)\mu(B) = \rho\mu(A).$$

Multiply both sides by $F(A)$ and take sum over $A \in V^+$. By (5.11),

$$\sum_{A \in V^+} F(A)\mu(A) - \mu(S) = \rho \sum_{A \in V^+} F(A)\mu(A). \tag{5.29}$$

We need to show that $\mu(S) > 0$. Assume $\mu(S) = 0$. Then for any $n > 0$, since $\mathbf{M}^n\mu = \rho^n\mu$, we have

$$\sum_{A \in V} \mathbf{M}^{(n)}(S, A)\mu(A) = \rho^n\mu(S) = 0.$$

Therefore, for each $A \in V$, $\mathbf{M}^{(n)}(S, A)\mu(A) = 0$. Because each $A \in V$ is reachable from S, there is some $n > 0$ such that $\mathbf{M}^{(n)}(S, A) > 0$ and therefore $\mu(A) = 0$. Hence $\mu = 0$. This is contradicting to that μ is the eigenvalue of $\mathbf{M}$. Hence $\mu(S) > 0$. Then by (5.29),

$$\sum_{A \in V^+} F(A)\mu(A) > \rho \sum_{A \in V^+} F(A)\mu(A) \geq 0$$

This proves $\rho < 1$. □

We will give another proof of Proposition Proposition 9 using the fact that ρ is less than 1. But first we need to introduce the following well-known spectrum theorem for matrices:

Theorem 3. Suppose $\mathbf{M}$ is an $n \times n$ real matrix. Let $\sigma(\mathbf{M})$ be the largest absolute value of $\mathbf{M}$'s eigenvalues. Then

$$\sigma(\mathbf{M}) = \lim_{n \to \infty} \|\mathbf{M}^n\|^{1/n},$$

where $\|\mathbf{M}\|$ is the norm of $\mathbf{M}$ defined by

$$\|\mathbf{M}\| = \sup_{\substack{\vec{v} \in \mathbf{M}^n \\ |\vec{v}|=1}} |\mathbf{M}\vec{v}|.$$

Now we can prove the following result.

Proposition 12. If $\mathbf{M}$ given by (5.27) has branching rate $\rho < 1$, then for each $m \in \mathbf{N} \cup \{0\}$,

$$E_p|\omega|^m < \infty. \tag{5.30}$$

Proof. We repeat the proof of Proposition Proposition 9 in §5.4 up to (5.18). Instead of taking sum over A, we observe that (5.18) can be written as

$$M_{k+1,\,A} \le K + \sum_{B \in V} \mathbf{M}(A, B) M_{k,\,B}.$$

Write $\{M_{k,\,A}\}_{A \in V}$ as $\vec{M}_k$, which is a vector indexed by $A \in V$. We then have

$$\vec{M}_{k+1} \le K\mathbf{1} + \mathbf{M}\vec{M}_k,$$

where $\mathbf{1} = \{1, \ldots, 1\}$ and $\vec{\mu} \le \vec{\nu}$ means each component of μ is less or equal to the component of ν with the same index. All the components in $K\mathbf{1}$, $\mathbf{M}$ and $\vec{M}_k$ are positive. Hence we can get

$$\vec{M}_{k+2} \le K\mathbf{1} + \mathbf{M}\vec{M}_{k+1} \le K\mathbf{1} + K\mathbf{M}1\mathbf{M}^2\vec{M}_{k+1}$$

By induction, we get

$$\vec{M}_k \le K \sum_{j=0}^{k-2} \mathbf{M}^j \mathbf{1} + \mathbf{M}^{k-1}\vec{M}_1,$$

therefore,

$$|\vec{M}_k| \le K \sum_{j=0}^{k-2} \|\mathbf{M}^j\||\mathbf{1}| + \|\mathbf{M}^{k-1}\||\vec{M}_1|. \tag{5.31}$$

By Theorem Theorem 2, for $\mathbf{M}$ given by (5.27), $\sigma(\mathbf{M}) = \rho$. Then by Theorem Theorem 3, for any $\rho < \rho' < 1$, $\|\mathbf{M}^n\| = o(\rho'^n)$. Then (5.31) implies that $|\vec{M}_k|$ is bounded. Since $\vec{M}_k$ are positive and increasing, it follows that $\vec{M}_k$ converge. This completes the proof. Note that the finiteness of entropy is just a special case here. □

In §5.5, we demonstrated that if p is an inconsistent PCFG distribution, then its renormalized version $\tilde{p}$ is a consistent PCFG distribution. We can then talk about entropy of the language under the distribution $\tilde{p}$. Is the entropy finite? We first look at a simple example. Consider a CFG in Chomsky normal form:

$$\begin{aligned} S &\rightarrow SS \\ S &\rightarrow a \end{aligned}$$

where a is the only terminal symbol. Assign probability p to the first production ($S \rightarrow SS$). Then the total probability of finite parse trees is $\min(1, 1/p - 1)$. If $p > 1/2$, then $\min(1, 1/p - 1) = 1/p - 1 < 1$. Therefore for $p > 1/2$, the PCFG is inconsistent. The probability of ($S \rightarrow SS$) after normalization is $1 - p < 1/2$. It is easy to see that the entropy of the renormalized distribution is finite.

More generally, we have the following result, which, put in words, means that a renormalized imconsistent distribution of a connected PCFG is a consistent distribution with finite entropy.

Proposition 13. If p imposes an imconsistent distribution on $\Omega = \Omega_S$ and the underlying branching process is connected, then the renormalized distribution (5.25) on Ω_S has $\rho < 1$, therefore finite entropy.

Proof. Because the branching process is connected and $p(\Omega_S) < 1$, all $p(\Omega_A) < 1$. To see this, consider all the parse trees (not necessarily finite) with A as their root that contain S. Since S is reachable from A, these trees have positive probabilities. Because $p(\Omega_S) < 1$, therefore, with a positive probability, some parse trees with S as their root do not terminate. Then it is seen that some of the parse trees with A as the root that contain S do not terminate, either. Therefore $p(\Omega_A) < 1$.

For each A, define the generating function (Harris [11], §2.2):

$$g_A(s) = \sum_{\substack{\alpha \in (V \cup T)^* \\ \text{s.t. } (A \rightarrow \alpha) \in D}} p(A \rightarrow \alpha) \prod_{B \in V} s_B^{n(B;\alpha)}, \tag{5.32}$$

where $s = \{s_A\}_{A \in V^+}$. Write $g = \{g_A\}_{A \in V^+}$ and $g^{(n)} = \{g_A^{(n)}\}$, where $g_A^{(n)}$ is recursively defined as

$$\left.\begin{aligned} g_A^{(1)}(s) &= g_A(s) \\ g_A^{(n)}(s) &= g_A(g^{(n-1)}(s)) \end{aligned}\right\} \tag{5.33}$$

It is easy to see that $g_A(\mathbf{0})$ is the total probability of trees with root A and height 1, and $g_A^{(n)}(\mathbf{0})$ is the total probability of trees with root A and height $\le n$. Therefore, $g_A^{(n)}(\mathbf{0}) \uparrow p(\Omega_A) < 1$. Write

$$q = \{p(\Omega_A)\}_{A \in V^+}.$$

Then $g(q) = g(\lim g^{(n)}(\mathbf{0})) = \lim g(g^{(n)}(\mathbf{0})) = \lim g^{(n+1)}(\mathbf{0}) = q$. q is the "smallest" non-negative solution of $g(s) = s$. That is, if there is some non-negative $q' \neq q$ with $g(q') = q'$, then $q \le q'$. This is because $\mathbf{0} \le q'$ implies $g^{(n)}(\mathbf{0}) \le g^{(n)}(q') = q'$ for all $n > 0$. Let $n \rightarrow \infty$, we then see $q \le q'$. Clearly $g(\mathbf{1}) = \mathbf{1}$.

We now renormalize p and get $\tilde{p}$ by (5.26). Let $f = \{f_A\}$ be the corresponding generating functions of $\tilde{p}$ defined as in (5.32). Define $f^{(n)}$ recursively as in (5.33). Then

$$\begin{aligned} f_A(s) &= \sum_{\substack{\alpha\in(V\cup T)^* \\ \text{s.t. } (A\to\alpha)\in D}} \tilde{p}(A\to\alpha) \prod_{B\in V} s_B^{n(B;\alpha)} \\ &= \sum_{\substack{\alpha\in(V\cup T)^* \\ \alpha\in(V\cup T)^* \\ \text{s.t. } (A\to\alpha)\in D}} \frac{1}{q_A} p(A\to\alpha) \prod_{B\in V} q_B^{n(B;\alpha)} \prod_{B\in V} s_B^{n(B;\alpha)}, \qquad (5.34) \end{aligned}$$

hence $f_A(s) = g_A(qs)/q_A$, where qs is defined as $\{q_A s_A\}$. Note that each $q_A = p(\Omega_A)$ is positive, hence $f_A(s)$ is well defined. If q' is a solution of $f(s) = s$, then qq' is a solution of $g(s) = s$. Because q is the smallest non-negative solution of $g(x) = x$, $\mathbf{1}$ is the only solution of $f(s) = s$ in the unit cube. Since $g(s) = s$ has a solution $\mathbf{1}$, $f(s) = s$ has a solution $\{1/q_A\}$, which is strictly larger than $\mathbf{1}$. We want to know how f_A change on the line segment connecting $\mathbf{1}$ and $\{1/q_A\}$. So we let $u = \{u_A\}$, where $u_A = 1/q_A - 1$. Then u is strictly positive. Each element on the line segment between $\mathbf{1}$ and $\{1/q_A\}$ can be written as $\mathbf{1} + tu$, $t \in [0, 1]$. Define $h(t) = \{h_A(t)\}$ such that

$$\begin{aligned} h_A(t) &= f_A(\mathbf{1} + tu) - 1 - tu_A \\ &= \sum_{\substack{\alpha\in(V\cup T)^* \\ \text{s.t. } (A\to\alpha)\in D}} \tilde{p}(A\to\alpha) \prod_{B\in V} (1 + tu_B)^{n(B;\alpha)} - 1 - tu_A. \qquad (5.35) \end{aligned}$$

It is seen that $h'(0) = \mathbf{M}u - u$, where $\mathbf{M}$ is the mean matrix corresponding to $\tilde{p}$. Every $h_A(t)$ is a convex function. Since $h_A(0) = h_A(1) = 0$, $h'_A(0) \le 0$, hence

$$\mathbf{M}u \le u.$$

We show that for at least one A, $(\mathbf{M}u)_A < u_A$. Note that $h'_A(0) = 0$ only if $h_A(t)$ is linear. Assume all $(\mathbf{M}u)_A = u_A$, then all $h_A(t)$ are linear. $h'_A(0) = 0$ and $h_A(0) = 0$ then imply that every $h_A(t)$ equals 0. Choose $t < 0$ such that $1 + tu_A > 0$ for all A. Then $f(\mathbf{1} + tu) - \mathbf{1} - tu = \mathbf{0}$. So we get a non-negative solution of $f(s) = s$ which is strictly less than $\mathbf{1}$. This is contradicting to the fact that $\mathbf{1}$ is the smallest non-negative solution of $f(s) = s$. So now we have

$$\mathbf{M}u \le u,\ (\mathbf{M}u)_A < u_A \quad \text{for at least one } A$$

By Theorem Theorem 2, there is a strictly positive left eigenvector ν such that $\nu\mathbf{M} = \rho\nu$. Because u is also strictly positive, we then have $\nu\mathbf{M}u < \nu u$, or $\rho\nu u < \nu u$. Hence $\rho < 1$. This completes the proof. □

5.7 Identifiability of Parameters and a Relation between Production Probabilities and Frequencies of Productions

Identifiability of parameters has a close relation with the consistency of estimators. The issue of the consistency of the ML estimator (5.4) of PCFGs is quite easy. If p imposes a consistent distribution on the set of parse trees, then as the size of i.i.d. samples goes to infinity, with probability one, the ML estimator $\hat{p}$ converges to p. To see this, think of the n parse tree samples as taken from the following experiment. Starting from S, we begin a branching process governed by the production probabilities. Once the process terminates, we start from S a new branching process. We repeat n times and collect the n resulting samples. It is seen that during the experiment, all the instances of variables choose their productions independently from each other. Then by the law of large numbers, the ratio between the number of instances of $A \to \alpha$ and the number of instances of A, which is exactly $\hat{p}(A \to \alpha)$, converges to $p(A \to \alpha)$, with probability one.

Because the ML estimator (5.4) for PCFGs is consistent, production probabilities are identifiable parameters of PCFGs. In other words, different systems of production probabilities impose different distributions on languages. This fact can be presented as

Proposition 14. If p_1, p_2 impose distributions P_1, P_2, respectively, on Ω and $p_1 \neq p_2$, then $P_1 \neq P_2$.

Proof. Assume $P_1 = P_2$. Then draw n i.i.d. samples from P_1. Because the ML estimator $\hat{p}$ is consistent, as $n \to \infty$, with probability one,

$$\hat{p}(A \to \alpha) \to p_1(A \to \alpha).$$

With the same reason

$$\hat{p}(A \to \alpha) \to p_2(A \to \alpha).$$

Hence $p_1 = p_2$, a contradiction. □

We mentioned in §5.5 that the ML estimators (5.23) and (5.24) may give us infinitely many solutions when the language generated by a CFG is equipped with distributions given by (5.21). This phenomenon of multi-solutions comes from the *non-identifiability* of parameters of Gibbs distributions, i.e., different parameters may yield the same Gibbs distribution.

To see why parameters of Gibbs distribution (5.21) are non-identifiable, we note that the f's have the following relations:

$$\begin{aligned}\sum_{(A\to\alpha)\in D} f(A \to \alpha; \omega) &= \sum_{(B\to\beta)\in D} n(A;\beta) f(B \to \beta; \omega), \qquad \text{if } A \neq S,\\ \sum_{(S\to\alpha)\in D} f(S \to \alpha; \omega) &= \sum_{(B\to\beta)\in D} n(S;\beta) f(B \to \beta; \omega) + 1.\end{aligned}$$

From these relations, it can be shown that f's are linearly dependent. In other words, there exists a $\lambda_0 \neq 0$, such that for any ω, $\lambda_0 \cdot f(\omega) = 0$. If $\hat{\lambda}$ is a solution for (5.23), the for any number t,

$$e^{(\hat{\lambda}+t\lambda_0)\cdot f(\omega)} = e^{\hat{\lambda}\cdot f(\omega)},$$

and

$$Z_{\hat{\lambda}+t\lambda_0} = Z_{\hat{\lambda}}.$$

Hence $P_{\hat{\lambda}+t\lambda_0}(\omega) = P_{\hat{\lambda}}(\omega)$. Therefore for any t, $\hat{\lambda} + t\lambda_0$ is also a solution for (5.23). This tells us that the parameters of Gibbs distribution (5.21) are non-identifiable.

Finally, we consider a relation between production probabilities and the expected values of frequencies of productions in parse trees. If under the distribution imposed by $p(A \to \alpha)$, the entropy of the language is finite, then for i.i.d. $\omega_1, \dots, \omega_n$, by consistency of the ML estimator given by (5.4),

$$\hat{p}(A \to \alpha) = \frac{\sum_{i=1}^{n} f(A \to \alpha; \omega_i)/n}{\sum_{\substack{\beta \text{ s.t.} \\ (A \to \beta) \in D}} \sum_{i=1}^{n} f(A \to \alpha; \omega_i)/n} \to p(A \to \alpha). \quad \text{w.p. } 1$$

Because the entropy is finite, for every production rule $(A \to \beta) \in D$,

$$\frac{1}{n} \sum_{i=1}^{n} f(A \to \beta; \omega) \to E_p(f(A \to \beta; \omega)), \quad \text{w.p. } 1,$$

where E_p is expectation under p. Therefore,

$$p(A \to \alpha) = \frac{E_p(f(A \to \alpha; \omega))}{\sum_{(A \to \beta) \in D} E_p(f(A \to \beta; \omega))}.$$

If the entropy is infinite, the above argument does not work. However, we have the following

Proposition 15. Suppose p's impose a consistent distribution P on Ω. Then for any increasing sequence of finite subsets Ω_n of Ω with $\Omega_n \uparrow \Omega$, i.e., $\Omega_1 \subset \Omega_2 \ldots \subset \Omega$, Ω_n finite and $\cup \Omega_n = \Omega$,

$$p(A \to \alpha) = \lim_{n \to \infty} \frac{E_p(f(A \to \alpha; \omega) | \omega \in \Omega_n)}{\sum_{(A \to \beta) \in D} E_p(f(A \to \beta; \omega) | \omega \in \Omega_n)},$$

where $E_p(f(A \to \alpha; \omega) | \omega \in \Omega_n)$ is the conditional expectation of $f(A \to \alpha; \omega$ on Ω given by

$$E_p(f(A \to \alpha; \omega) | \omega \in \Omega_n) = \frac{\sum_{\omega \in \Omega_n} f(A \to \alpha; \omega) p(\omega)}{\sum_{\omega \in \Omega_n} p(\omega)}$$

Proof. Considered as a vector, $p = \{p(A \to \alpha)\}$ belongs to the following set of vectors indexed by D,

$$\Big\{ v = \{v(A \to \alpha)\}_{(A \to \alpha) \in D},\ v(A \to \alpha) > 0 : \ \text{ for all } \ (A \to \alpha) \in D \Big\}.$$

Note we do not require that for each A, the components $v(A \to \alpha)$ add up to 1. On this set, define a function

$$L_n(v) = \sum_{\omega \in \Omega_n} P(\omega | \Omega_n) \log \frac{P(\omega | \Omega_n)}{\prod_{(A \to \alpha) \in D} v(A \to \alpha)^{f(A \to \alpha; \omega)}} \tag{5.36}$$

Let $\hat{p}_n$ be the (unique) minimizer of (5.36) subject to

$$\sum_{(A\to\alpha)\in D} v(A\to\alpha) = 1.$$

It is easy to see that

$$\hat{p}_n(A\to\alpha) = \frac{\sum_{\omega\in\Omega_n} f(A\to\alpha;\omega)P(\omega)}{\sum_{\omega\in\Omega_n} f(A;\omega)P(\omega)}.$$

Here $f(A;\omega)$ is defined as the sum of $f(A\to\alpha;\omega)$ over all $\alpha\in(V\cup T)^*$ with $(A\to\alpha)\in D$. In order to show that $\hat{p}_n\to p$, consider the auxiliary function

$$F_n(v) = L_n(v) + \sum_{A\in V^+}\sum_{\omega\in\Omega_n} P(\omega|\Omega_n)f(A;\omega) \sum_{\substack{\alpha\in(V\cup T)^* \\ \text{s.t. } (A\to\alpha)\in D}} v(A\to\alpha). \tag{5.37}$$

Then

$$\frac{\partial F_n}{\partial v(A\to\alpha)} = -\frac{\sum_{\omega\in\Omega_n} P(\omega|\Omega_n)f(A\to\alpha;\omega)}{v(A\to\alpha)} + \sum_{\omega\in\Omega_n} P(\omega|\Omega_n)f(A;\omega) \tag{5.38}$$

$$\frac{\partial^2 F_n}{\partial v(A\to\alpha)\partial v(A\to\alpha)} = \frac{\sum_{\omega\in\Omega_n} P(\omega|\Omega_n)f(A;\omega)}{v^2(A\to\alpha)}, \tag{5.39}$$

$$\frac{\partial^2 F_n}{\partial v(A\to\alpha)\partial v(B\to\beta)} = 0, \text{ if } (A\to\alpha)\neq(B\to\beta). \tag{5.40}$$

Hence $\partial F_n/\partial\hat{p}_n = 0$ and F_n is strictly convex. This implies that $\hat{p}_n$ is the unique minimizer of F_n. Since for all n, $\hat{p}_n(A\to\alpha) < 1$, the Hessian of F_n is uniformly lower bounded from 0. Then there is an $a > 0$, such that for *all* n, $F_n(p) - F_n(\hat{p}_n) \geq a\|p-\hat{p}\|^2$.

Now for each A, both $p(A\to\alpha)$ and $\hat{p}_n(A\to\alpha)$ add up to 1. Then from (5.37) and the fact that $\hat{p}_n$ is the minimizer of L_n,

$$F_n(p) - F_n(\hat{p}_n) = L_n(p) - L_n(\hat{p}_n) \geq 0.$$

By Jensen's inequality,

$$\begin{aligned} -L_n(\hat{p}_n) &= \sum_{\omega\in\Omega_n} P(\omega|\Omega_n)\log\frac{\prod_{(A\to\alpha)\in D}\hat{p}_n(A\to\alpha)^{f(A\to\alpha;\omega)}}{P(\omega|\Omega_n)} \\ &\leq \log\sum_{\omega\in\Omega_n}\prod_{(A\to\alpha)\in D}\hat{p}_n(A\to\alpha)^{f(A\to\alpha;\omega)} \\ &\leq \log 1 = 0, \end{aligned}$$

therefore $L_n(\hat{p}_n) \geq 0$. Then it follows that

$$L_n(p) \geq F_n(p) - F_n(\hat{p}_n) \geq 0.$$

Since $L_n(p) = -\log p(\Omega_n) \to 0$ as $n\to\infty$, then $F_n(p) - F_n(\hat{p}_n) \to 0$. Hence $\hat{p}_n \to p$. □

Bibliography

[1] S. P. Abney. Stochastic Attribute-Value Grammars. To appear in *Computational Linguistics.*

[2] J. K. Baker. Trainable Grammars for Speech Recognition. In *Speech Communications Papers of 97'th Meeting of Acoustical Society of America*, pages 547-550, Cambridge, Massachusetts. 1979.

[3] L. E. Baum. An Inequality and Associated ML in Statistical Estimation of Probabilistic Functions of Markov Processes. *Inequalities*, 3:1-8. 1972.

[4] J. Besag. Spatial Interaction and Statistical Analysis of Lattice Systems (with discussion). *Journal of Royal Statist. Soc. , Ser. B*, 36:192-236. 1974.

[5] J. Besag. Efficiency of Pseudolikelihood for Simple Gaussian Field. *Biometrika* 64:616-619. 1977.

[6] T. L. Booth and R. A. Thompson. Applying Probability Measures to Abstract Languages. *IEEE Trans. on Computers,* C-22:442-450. 1973.

[7] Z. Chi and S. Geman. Estimation of Probabilistic Context-Free Grammars. To appear in *Computational Linguistics.* 1998.

[8] A. Dempster, N. Laird, and D. Rubin. Maximum-Likelihood from Incomplete Data via the EM Algorithm. *Journal of Royal Statist. Soc., Ser. B*, 39:1-38. 1977.

[9] S. Geman and C. Graffigne. Markov Random Field Models and Their Applications to Computer Vision. In M. Gleasor (editor), *Proceedings of the International Congress of Mathematicians* (1986), pages 1496-1517, Amer. Math. Soc. Providence. 1987.

[10] U. Grenander. *Lectures in Pattern Theory Volume 1, Pattern Synthesis.* Springer-Verlag, New York. 1976.

[11] T. E. Harris. *The Theory of Branching Processes.* Springer-Verlag, Berlin. 1963.

[12] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison Wesley, Reading, Massachusetts. 1979.

[13] K. E. Mark, M. I. Miller and U. Grenander. Constrained Stochastic Language Models. In S. E. Levinson and L. Shepp (editors). *Image Models (and Their Speech Model Cousins)*, pages 131-140. Springer. 1996.

[14] K. E. Mark, M. I. Miller, U. Grenander, and S. P. Abney. Paramter Estimation for Constrained Context-Free Language Models. In *Proceedings of the DARPA Speech and Natural Language Workshop,* pages 146-149, Harriman, New York. February, 1992. Morgan Kaufmann.

[15] K. E. Mark. Markov Random Field Models for Natual Language. PhD Thesis. Department of Electrical Engineering, Washington University. May, 1997.

[16] M. I. Miller and J. A. O'Sullivan. Entropies and Combinatorics of Random Branching Processes and Context-Free Languages. *IEEE Trans. on Information Theory,* Vol. 38, No. 4. July, 1992.

[17] J. A. Sánchez and J. M. Benedi. Consistency of Stochastic Context-Free Grammars from Probabilistic Estimation Based on Growth Transformation. *IEEE Trans. on PAMI,* Vol. 19, No. 9. September, 1997.

[18] C. S. Wetherell. Probabilistic Languages: A Review and Some Open Questions. *Computing Surveys*, 12:361-379. 1980.

[19] P. Walters. *An Introduction to Ergodic Theory.* Springer-Verlag, New York. 1982.

[20] L. Younes. Maximum-Likelihood Estimation for Gibbsian Fields. 1991.

Chapter 6

Probabilistic Feature Based Grammars

6.1 Introduction

Statistical language models are becoming increasingly important in linguistics. The development of such models aims to solve problems that traditional categorical grammars face. Sometimes called non-probabilistic grammars, categorical grammars provide extremely detailed syntactic and semantic analyses of a range of sentences. They also have the merit of being sensitive to a wide variety of linguistic interactions. However, categorical grammars have several drawbacks which hinder their utility. First of all, because the grammars fail to address the ranking of grammatical analyses, they suffer serious inefficiency problem when dealing with sentences which have tremendous amount of different analyses. For the same reason, they also lack robustness when coming across unexpected or ill-formed input. Furthermore, with no practical automatic learning mechanism to categorical grammars, such grammars have to be hand-crafted and usually become so complex that they are difficult or impossible to understand and maintain.

Statistical language models are probabilistic versions of categorical grammars, with all analyses allowed by the grammars being equipped with probabilities. The assignment of probability measures automatically enables statistical language models to systematically treat grammatical analyses differently. When good statistical models are established for languages, analyses empirically more likely to be chosen are allocated higher probabilities, hence more likely to be selected by parsing algorithms. Good statistical models also make it possible that analyses of ill-formed input have very low probabilities, making them easily detected by the parsing algorithms. Because of the discriminating power of probabilities, the rules by statistical models need not be as detailed and complex as categorical grammars when modeling the same languages. In addition, statistical models can be adjusted by tuning their parameters and can be learned from the training corpus because the parameters can be estimated.

The simplest statistical language models are probabilistic regular grammars (PRGs) and probabilistic context-free grammars (PCFGs). They are actually the same things as Markov chains and stochastic branching processes, respectively. Both models have had remarkable applications to simple tasks in speech recognition and computer vision (Chou [4]). However, these grammars' non-probabilistic prototypes, i.e., regular grammars (RGs) and context-free grammars (CFGs), are widely deemed linguistically inadequate, because they lack the context sensitivity that is ubiquitous in natural languages. In order to apply statistical methods more effectively to linguistics, it is necessary to develop probabilistic versions of more expressive grammars.

Standard grammars in computational linguistics are attribute-value grammars of some variety. In this article, we will call attribute-value grammars feature based grammars. RGs and CFGs are two types of feature based grammars, but among the least expressive ones. The more expressive feature based grammars cope with context sensitivity by addressing features that contain non-local information of languages. Efforts have been made to develop general probabilistic feature based grammars (Mark et al. [7], Abney [1]). Invariably, all the probabilities proposed for feature based grammars take the form of Gibbs distribution. The argument for the Gibbs form is based on the "maximum entropy" principle (Jaynes [6]). In Mark et al. [7], a Gibbs distribution was derived for a simple case, where the probabilistic models are combinations of a PCFG and n-gram language models, by invoking maximum entropy estimation. Similar argument can be applied to more general cases to get the Gibbs distributions as discussed in Abney [1]. However, this was not pursued in either of the two articles. In §6.2, we will derive the Gibbs form of distributions on features based grammars and some of its variants.

The emphasis of this article is on the technical issues of parameter estimation. In §6.3 and §6.4, we will propose two schemes for estimation. Both schemes are easy to prove to be consistent. We will argue that the second scheme, which is a pseudo-likelihood type scheme for estimation, is efficient, if the goal of parameter estimation is to analyze sentences rather than sample sentences.

6.2 Gibbs Distributions for Feature Based Grammars

Given a grammar G, let Ω be the set of all parse trees allowed by G. Elements in Ω are denoted as ω. Because a natural language has only countably many sentences, and each sentence has only finitely many parse trees allowed by G, Ω is countable. Let $f_1(\omega), \dots, f_N(\omega)$ be N real functions, or "features", on Ω. Suppose under certain unknown distribution on Ω, the expectation of $f_1(\omega), \dots, f_N(\omega)$ are $\bar{f}_1, \dots, \bar{f}_N$, respectively. With only $\bar{f}_1, \dots, \bar{f}_N$ being known, we want to make a reasonable guess about the unknown distribution.

For this end, the maximum entropy principle suggests using the solution of the following constrained maximization problem,

$$p = \underset{\tilde{p} \text{ prop on } \Omega}{\arg\max} \left\{ -\sum_{\omega \in \Omega} \tilde{p}(\omega) \log \tilde{p}(\omega) \right\},$$

subject to

$$E_p(f_i(\omega)) = \sum_{\omega \in \Omega} f_i(\omega) p(\omega) = \bar{f}_i, \quad i = 1, \dots, N \tag{6.1}$$

and

$$\sum_{\omega\in\Omega} p(\omega) = 1. \tag{6.2}$$

The philosophy for the above approximation is that while $p(\omega)$ satisfies the given constraints on f_i, it should be made as random (or un-informative) as possible in other unconstrained dimensions, i.e., $p(\omega)$ should represent information no more than what is available and in this sense, the maximum entropy principle is often called the minimum prejudice principle (Zhu et al. [5]).

By introducing the Lagrange multipliers λ_i, $i = 1, \ldots, N$, and β, the constrained maximization problem is changed to

$$\frac{\partial}{\partial p(\omega)} \left\{ -\sum_{\omega\in\Omega} p(\omega) \log p(\omega) + \sum_{i=1}^{N} \lambda_i \sum_{\omega\in\Omega} f_i(\omega) p(\omega) + \beta \sum_{\omega\in\Omega} p(\omega) \right\} = 0.$$

Solving this equation gives

$$p(\omega) = \frac{1}{Z(\lambda)} e^{\lambda\cdot f(\omega)}, \tag{6.3}$$

where $\lambda = (\lambda_1, \ldots, \lambda_N)$ and $f(\omega) = (f_1, \ldots, f_N)$, and $Z(\lambda) = \sum e^{\lambda\cdot f(\omega)}$.

The maximum entropy principle can be generalized to the "minimum discriminant principle" (Mark et al. [7]). Suppose we have a distribution $\pi(\omega)$ on Ω, then the minimum discriminant principle requires the guess of the unknown distribution minimize the following quantity,

$$\sum_{\omega\in\Omega} p(\omega) \log \frac{p(\omega)}{\pi(\omega)},$$

subject to (6.1) and (6.2). Then we get the solution with the form

$$p(\omega) = \frac{1}{Z(\lambda)} \pi(\omega) e^{\lambda\cdot f(\omega)} = \frac{1}{Z(\lambda)} e^{\lambda\cdot f(\omega) + \log \pi(\omega)}, \tag{6.4}$$

which is still a Gibbs form.

If π is finite, an explanation for the constrained minimization is as follows. Write

$$\sum_{\omega\in\Omega} p(\omega) \log \frac{p(\omega)}{\pi(\omega)} = -\log \pi(\Omega) + \sum_{\omega\in\Omega} p(\omega) \log \frac{p(\omega)}{\kappa(\omega)},$$

where $\kappa(\omega) = \pi(\omega)/\pi(\Omega)$ is a probability distribution on Ω. The minimization then finds the distribution which satisfies the constraints and is closest to the distribution κ in terms of Kullback-Leiber distance. However, this explanation does not apply to the case where $\pi(\Omega) = \infty$.

Because the set of all parses is infinite, both (6.3) and (6.4) have the possible problem that the partition number $Z(\lambda)$ might be infinity, which makes the distribution not well-defined. An alternative to the Gibbs forms (6.3) and (6.4) is the following distribution,

$$p(\omega) = \pi(Y(\omega)) \frac{e^{\lambda\cdot f(\omega)}}{\sum\limits_{Y(\omega')=Y(\omega)} e^{\lambda\cdot f(\omega')}}, \tag{6.5}$$

where $Y(\omega)$ is the "yield" of ω, which is the terminal string associated with the parse tree ω, and π is a probability distribution on the language.

From the information point of view, we can think of π as a description of the mechanism to generate sentences. It can be different from the Gibbs distribution with the potential function $\lambda \cdot f(\omega)$. On the other hand, the rules to analyze individual sentences, which are given by λ and f, with λ being the parameter, are uniform across all sentences.

The potential function $\lambda \cdot f(\omega)$ can be looked on as the first order expansion of a function $\varphi(f(\omega))$. Even when f gives all the information about Ω, i.e., the σ-algebra $\mathcal{F}(f)$ contains all the singleton sets $\{\omega\}$, the Gibbs distribution (6.3) can still be very far from the true distribution. As an example, suppose $\Omega = \mathbf{N}$ and $f(\omega)$ for $\omega \in \Omega$ is the numerical value of the element. If p is a distribution on Ω with $p(2) \gg p(\Omega \backslash \{2\})$, then the Gibbs distribution (6.3) can never get close to p.

A solution to this problem is to learn the function φ, on the set of all possible values of $f(\omega)$. To do this, one can approximate $\varphi(f(\omega))$ by a higher order expansion and estimate λ's in the expansion,

$$\sum_{i \leq k} \lambda_i (f(\omega))^i,$$

where $i = (i_1 \dots i_n)$ is a multiple index composed of non-negative integers. $i \leq k$ means $i_1 + \cdots + i_n \leq k$, and f^i means $f_1^{i_1} \cdots f_N^{i_N}$. For the example given just now, a second order expansion can do well enough in the sense that

$$\sum_{\omega} \left| \sqrt{p(\omega)} - \sqrt{p_\lambda(\omega)} \right|$$

is small, where $p_\lambda(\omega)$ is a Gibbs distribution with the potential function $\lambda_1 f(\omega) + \lambda_2 (f(\omega))^2$. It turns out that λ_1 and λ_2 should satisfy $\lambda_1 \approx -4\lambda_2$ and $\lambda_1 \gg 0$.

One can also learn $\varphi(f(\omega))$ by dividing the range of f into several bins $B_1, \dots, B_k$, and approximating $\varphi(f)$ by a function which is constant in each bin (Zhu et al. [5]). The potential function is then changed to

$$\sum_{i=1}^{k} \lambda_i \mathbf{1}_i(f(\omega)),$$

where $\mathbf{1}_i$ is the indicator function of the bin B_i.

Next we will consider how to estimate parameter λ of the Gibbs forms. From now on, we will always use N as the notation for the dimension of f.

6.3 Maximum-Likelihood (ML) Type Estimation of Parameters

Suppose we are given n i.i.d. samples. Let us consider two cases about the data.

Case One:

In this case, the n samples are fully observed parse trees $\omega_1, \ldots, \omega_n$. Under the assumption that the distribution of ω is given by (6.3) with parameter λ_0, if $|f|$ has finite mean and if n is large, then by the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} f(\omega_i) \approx E_{\lambda_0}(f),$$

where $E_{\lambda_0}(f)$ is the expectation of $f(\omega)$ under the distribution $e^{\lambda_0 \cdot f(\omega)}/Z(\lambda_0)$.

Therefore, we take any solution to the following equation in λ as an estimate of λ_0,

$$E_\lambda(f) = \frac{1}{n}\sum_{i=1}^{n} f(\omega_i). \tag{6.6}$$

The estimation formulated by (6.6) is a maximum-likelihood type estimation. Indeed, *if* there is a solution to the following maximization problem,

$$\hat{\lambda} = \arg\max_{\lambda} \prod_{i=1}^{n} \frac{e^{\lambda \cdot f(\omega_i)}}{Z(\lambda)},$$

then the solution, $\hat{\lambda}$, is a solution to (6.6). However, because the set of all parse trees is infinite, we can not compute $Z(\lambda)$, therefore $E_\lambda(f)$ in (6.6) is unknown.

In order to get around this problem, we modify the estimation as follows. Let $Y_n = \{Y(\omega_1), \ldots Y(\omega_n)\}$. Then we replace (6.6) by

$$\frac{1}{n}\sum_{i=1}^{n} f(\omega_i) = \frac{\sum\limits_{Y(\omega)\in Y_n} f(\omega) e^{\lambda \cdot f(\omega)}}{\sum\limits_{Y(\omega)\in Y_n} e^{\lambda \cdot f(\omega)}} = E_\lambda[f(\omega)|Y(\omega) \in Y_n],$$

or

$$\bar{f} - E_\lambda[f(\omega)|Y(\omega) \in Y_n] = 0, \tag{6.7}$$

where $\bar{f}$ is the average of $f(\omega_1), \ldots f(\omega_n)$. It can be shown the left hand side of (6.7) is the gradient of the function

$$L_n(\lambda, \omega_1, \ldots, \omega_n) = \lambda \cdot \bar{f} - \log\left(\sum_{Y(\omega)\in Y_n} e^{\lambda \cdot f(\omega)}\right),$$

which is convex in λ. Note that since there can be multiple parse trees with the same yield, the set $\{\omega : Y(\omega) \in Y_n\}$ might be strictly larger than $\{\omega_1, \ldots, \omega_n\}$. Since $L_n(\lambda, \omega_1, \ldots, \omega_n)$ is convex in λ, any solution to the following maximization problem is a solution to (6.7), and vice versa,

$$\hat{\lambda}_n = \arg\max_{\lambda} \{L_n(\lambda, \omega_1, \ldots, \omega_n)\}. \tag{6.8}$$

In the remaining part of this section, we will use $L(\lambda)$ as short for $L_n(\lambda, \omega_1, \ldots, \omega_n)$. That the maximization problem (6.8) has a solution is not guaranteed. For example, suppose we have 3 ω's, ω_1, ω_2 and ω_3, and $f(\omega_1) = (0,0)$, $f(\omega_2) = (0,1)$, and $f(\omega_3) = (1,0)$. Suppose only ω_2 and ω_3 are observed, with each being observed once, and $Y(\omega_i)$, $i = 1,2,3$ are the same. Then $\bar{f}$ is the average of $f(\omega_2)$ and $f(\omega_3)$, i.e., $(1/2, 1/2)$, and

$$L(\lambda) = \lambda \cdot \bar{f} - \log\left(\sum_{i=1}^{3} e^{\lambda \cdot f(\omega_i)}\right) = \frac{\lambda_1 + \lambda_2}{2} - \log\left(1 + e^{\lambda_1} + e^{\lambda_2}\right).$$

The above function can not achieve its maximum. Indeed,

$$\nabla L(\lambda) = \left(\frac{1}{2} - \frac{e^{\lambda_1}}{1 + e^{\lambda_1} + e^{\lambda_2}}, \frac{1}{2} - \frac{e^{\lambda_2}}{1 + e^{\lambda_1} + e^{\lambda_2}}\right).$$

Since ∇L can never be 0, there are no extreme points for L.

In order to get the condition for the existence of solution to (6.8), let $\Omega_n = \{\omega : Y(\omega) \in Y_n\}$ and C be the convex closure of the set $\{f(\omega) : \omega \in \Omega_n\}$. The boundary of C is the union of all the facets of C and denoted as ∂C. The inner part of C is defined as $C \backslash \partial C$. Because $\bar{f}$ is the average of some of the $f(\omega)$'s with $\omega \in \Omega_n$, $\bar{f} \in C$.

Proposition 16. Suppose $f(\omega)$ are not all the same for $\omega \in \Omega_n$. Then the maximization problem (6.8) has a solution if and only if $\bar{f} \in C \backslash \partial C$.

Remark. If $f(\omega)$ are the same for all $\omega \in \Omega_n$, then $L(\lambda)$ is a constant.

Proof. What we need to show is that the function

$$L(\lambda) = \lambda \cdot \bar{f} - \log\left(\sum_{\omega \in \Omega_n} e^{\lambda \cdot f(\omega)}\right)$$

can achieve its maximum if and only if $\bar{f} \in C \backslash \partial C$.

Let k be the dimension of convex set C. Recall that N is the dimension of λ. Clearly, $k \leq N$. If $k < N$, then there is an N-dimensional vector $\beta \neq 0$ and a constant c, such that $\beta \cdot f(\omega) = c$ for all $\omega \in \Omega_n$. Without loss of generality, suppose the last component of β, $\beta_N \neq 0$. Then for all $\omega \in \Omega_n$,

$$f_N(\omega) = \frac{c}{\beta_N} - \frac{\beta_1}{\beta_N} f_1(\omega) - \cdots \frac{\beta_{N-1}}{\beta_N} f_{N-1}(\omega).$$

Then

$$L(\lambda) = \lambda' \cdot \bar{g} - \log\left(\sum_{\omega \in \Omega_n} e^{\lambda' \cdot g(\omega)}\right) \overset{\triangle}{=} L'(\lambda'),$$

where

$$\lambda' = \left(\lambda_1 - \frac{\beta_1 \lambda_N}{\beta_N}, \ldots, \lambda_{N-1} - \frac{\beta_{N-1}\lambda_N}{\beta_N}\right),$$

is an $N-1$ dimensional vector, and

$$g(\omega) = (f_1(\omega), \ldots, f_{N-1}(\omega))\,.$$

Let C' be the convex closure of $\{g(\omega) : \omega \in \Omega_n\}$. Then C' is still a k dimensional convex polygon but embedded in an $N-1$ dimensional space and $\bar{g} \in C' \backslash \partial C'$ if and only if $\bar{f} \in C \backslash \partial C$. Obviously, $L(\lambda)$ can get to its maximum if and only if $L'(\lambda')$ can. From the above procedure we see that we can reduce the dimension of λ until it equals k, without affecting the final conclusion.

In the remaining part of the proof we only consider the case where $k = N$. Let S be the $N-1$ dimensional unit sphere, which consists of all N dimensional vectors with $|v| = 1$. If $\bar{f} \in C \backslash \partial C$, then for any $v \in S$,

$$M(v) > v \cdot \bar{f},$$

where

$$M(v) = \max_{\omega \in \Omega_n} \{v \cdot f(\omega)\}\,.$$

The function $M(v) - v \cdot \bar{f}$ is continuous, therefore, by compactness of S, there is a constant $A > 0$ such that $M(v) - v \cdot \bar{f} > A$ for all $v \in S$.

For each $v \in S$, taking $L(tv)$ as a function in t, we have

$$L'(tv) = v \cdot \bar{f} - \frac{\sum\limits_{\omega \in \Omega_n} v \cdot f(\omega) e^{tv \cdot f(\omega)}}{\sum\limits_{\omega \in \Omega_n} e^{tv \cdot f(\omega)}} \rightarrow v \cdot \bar{f} - M(v) < -A, \quad t \rightarrow \infty,$$

and $L''(tv) < 0$. For each $t > 0$, let $B_t = \{v \in S : L'(tv) < 0\}$. From the above result we see $S \subset \cup_{t>0} B_t$. Because of the continuity of $L'(tv)$ in v, B_t is open. Because S is compact, $S \subset \cup B_{t_i}$ for some $t_1 < t_2 < \ldots < t_m$. For each $v \in S$, because $L(tv)$, when taken as a function in t, is concave, therefore, if $L'(t_0 v) < 0$, then for any $t > t_0$, $L'(tv) < 0$, which means $B_{t_i} \subset B_{t_m}$. This implies that for all $v \in S$ and $t > t_m$, $L'(tv) < 0$. Thus if $|\lambda| > t_m$, then $L(\lambda) < L(t_m v)$, where $v = \lambda/|\lambda|$. Therefore $L(\lambda)$ must get its maximum in the region $\{\lambda : |\lambda| \leq t_m\}$.

Conversely, assume $L(\lambda)$ achieves its maximum at some λ_0, then $\nabla L(\lambda_0) = 0$, and therefore

$$\bar{f} = \frac{\sum\limits_{\omega \in \Omega_n} f(\omega) e^{\lambda_0 \cdot f(\omega)}}{\sum\limits_{\omega \in \Omega_n} e^{\lambda_0 \cdot f(\omega)}}.$$

If the vertices of C are $v_1, \ldots, v_p$, then every $f(\omega)$ can be written as $a_1 v_1 + \ldots a_p v_p$, where $a_i \geq 0$ and $a_1 + \ldots a_p = 1$. Because each $e^{\lambda_0 \cdot f(\omega)} > 0$, from the above equality, $\bar{f} = b_1 v_1 + \ldots b_p v_p$ with each b_i being positive. Hence $\bar{f} \in C \backslash \partial C$. □

As mentioned earlier, (6.8) may not have a solution. One way to handle this problem is to modify the maximum-likelihood estimation (6.8) to the following form,

$$\hat{\lambda}_n = \arg \max_{|\lambda| \leq n} \{L_n(\lambda, \omega_1, \ldots, \omega_n)\}\,. \tag{6.9}$$

Case Two:
In this case, only the yields of the parse trees are observed. Let $y_1, \ldots, y_n$ be the n sentences and let $Y_n = \{y_1, \ldots, y_n\}$. Under the assumption that the distribution of ω is given by (6.3) with parameter λ_0, if $|f|$ has finite mean and if n is large, then by the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} E_{\lambda_0}[f(\omega)|Y(\omega) = y_i] \approx \sum_{y\in Y} E_{\lambda_0}[f(\omega)|Y(\omega) = y]P_{\lambda_0}(y) = \sum_{\omega\in\Omega} E_{\lambda_0}(f(\omega)),$$

where $P_{\lambda_0}(y)$ is the sum of all $P_{\lambda_0}(\omega)$ with $Y(\omega) = y$, and $Y = \{Y(\omega) : \omega \in \Omega\}$. On the other hand, as n is large enough,

$$\sum_{y\in Y_n} E_{\lambda_0}[f(\omega)|Y(\omega) = y]P_{\lambda_0}(y|Y_n) \approx \sum_{y\in Y} E_{\lambda_0}[f(\omega)|Y(\omega) = y]P_{\lambda_0}(y),$$

hence

$$\frac{1}{n}\sum_{i=1}^{n} E_{\lambda_0}[f(\omega)|Y(\omega) = y_i] \approx \sum_{y\in Y_n} E_{\lambda_0}[f(\omega)|Y(\omega) = y]P_{\lambda_0}(y|Y_n).$$

With a similar argument as in the first case, we take any solution to the following equation as an estimate of λ_0,

$$\frac{1}{n}\sum_{i=1}^{n} E_{\lambda}[f(\omega)|Y(\omega) = y_i] - \sum_{y\in Y} E_{\lambda}[f(\omega)|Y(\omega) = y]P_{\lambda}(y|Y_n) = 0. \tag{6.10}$$

To transform (6.10) into an optimization problem, define the log-likelihood function in λ,

$$L_n(\lambda, y_1, \ldots, y_n) = \frac{1}{n}\sum_{i=1}^{n} \log P_{\lambda}(y_i|Y_n).$$

Then

$$\nabla_{\lambda} L_n(\lambda, y_1, \ldots, y_n) = \frac{1}{n}\sum_{i=1}^{n} E_{\lambda}[f(\omega)|Y(\omega) = y_i] - \sum_{y\in Y} E_{\lambda}[f(\omega)|Y(\omega) = y]P_{\lambda}(y|Y_n).$$

Therefore, any maximizer of $L_n(\lambda, y_1, \ldots, y_n)$ is a solution to (6.10).

A condition that $L_n(\lambda, y_1, \ldots, y_n)$ can reach its maximum is as follows. As in the first case, suppose the dimension of λ is N.

Proposition 17. Given a convex set C in R^N, a plane P is called a support of C if $\emptyset \neq P \cap C \subset \partial C$. For each sentence y, let $C(y)$ be the convex closure of the set $\{f(\omega) : y(\omega) = y\}$.

If there is no such a plane P that it is a common support of $C(y_1), \ldots, C(y_n)$ and all $C(y)$'s are on the same side of P, then L_n achieves its maximum.

Proof. As in the proof of Proposition Proposition 16, let S be the unit sphere in $\mathbf{R}^N$. By the assumption, for any $v \in S$,

$$\max_{y \in Y_n} \max_{y(\omega)=y} \{v \cdot f(\omega)\} > \min_{y \in Y_n} \max_{y(\omega)=y} \{v \cdot f(\omega)\}.$$

The functions on both sides of the above inequality are continuous in v. Because S is compact, there is a constant $\delta > 0$, such that for any $v \in S$,

$$\max_{y \in Y_n} \max_{y(\omega)=y} \{v \cdot f(\omega)\} - \min_{y \in Y_n} \max_{y(\omega)=y} \{v \cdot f(\omega)\} > \delta$$

Based on this, using the same argument as Proposition Proposition 16, we can show that L_n achieves its maximum. □

If $L_n(\lambda, y_1, \ldots, y_n)$ can achieve its maximum, then the maximizers of the function are solutions to (6.10). However, unlike $L_n(\lambda, \omega_1, \ldots, \omega_n)$ in case one, $L_n(\lambda, y_1, \ldots, y_n)$ is not necessarily convex. Unless $y_1, \ldots y_n$ satisfy the condition of Proposition Proposition 17, $L_n(\lambda, y_1, \ldots, y_n)$ might not be able to achieve its maximum. As an alternative, we take

$$\hat{\lambda}_n = \arg \max_{|\lambda|<n} L_n(\lambda, y_1, \ldots, y_n), \tag{6.11}$$

as the estimate of λ_0.

For the estimation given by (6.11), we have the following consistency result.

Proposition 18. Assume the distribution on Ω is given by

$$P_{\lambda_0}(\omega) = \frac{e^{\lambda_0 \cdot f(\omega)}}{Z_{\lambda_0}}.$$

Suppose $\omega_1, \ldots, \omega_n$ are i.i.d. samples from P_{λ_0}. Let $y_i = Y(\omega_i)$, $i = 1, \ldots, n$, and $Y_n = \{y_1, \ldots, y_n\}$. Define $\Omega_n = \{\omega \in \Omega : \; Y(\omega) \in Y_n\}$. Let $\hat{\lambda}_n$ be the estimates given by (6.9) or (6.11). Define the distribution P_n on Ω such that

$$P_n(\omega) = \begin{cases} \dfrac{e^{\hat{\lambda}_n \cdot f(\omega)}}{\sum\limits_{\omega' \in \Omega_n} e^{\hat{\lambda}_n \cdot f(\omega')}} & \text{if } \omega \in \Omega_n \\ 0 & \text{otherwise} \end{cases}$$

(1) If $\hat{\lambda}_n$ are given by (6.9) and if $H = -\sum\limits_{\omega \in \Omega} P_{\lambda_0}(\omega) \log P_{\lambda_0}(\omega) < \infty$, then with probability 1, as $n \to \infty$, P_n weakly converges to P_{λ_0} on Ω, i.e.,

$$P_n(\omega) \to P_{\lambda_0}(\omega), \quad \text{for any } \omega \in \Omega.$$

(2) If $\hat{\lambda}_n$ are given by (6.11) and if $H = -\sum\limits_{y \in Y} P_{\lambda_0}(y) \log P_{\lambda_0}(y) < \infty$, then with probability 1, as $n \to \infty$, P_n weakly converge to P_{λ_0} on Y, i.e.,

$$P_n(y) \to P_{\lambda_0}(y), \quad \text{for any } y \in Y.$$

Proof. We only prove (2). The proof of (1) is very similar to the proof of (2).

Write $L_n(\lambda)$ for $L_n(\lambda, y_1, \ldots, y_n)$. For any integer $n > |\lambda_0|$, by (6.11), $L_n(\lambda_n) \geq L_n(\lambda_0)$. But

$$L_n(\lambda_0) = \frac{1}{n}\sum_{i=1}^{n} \log P_{\lambda_0}(y_i) + \log Z(\lambda_0) - \log\left(\sum_{\omega\in\Omega_n} e^{\lambda_0\cdot f(\omega)}\right).$$

With probability 1, $L_n(\lambda_0) \to H$, hence

$$\liminf \frac{1}{n}\sum_{i=1}^{n} \log P_n(y_i) \geq H.$$

Let $I_n(y)$ denote the empirical probability of y, i.e.,

$$I_n(y) = \frac{|\{i:\ y_i = y\}|}{n}.$$

Then

$$\begin{aligned} L_n(\lambda_n) &= \frac{1}{n}\sum_{i=1}^{n} \log P_n(y_i) \\ &= \sum_{y\in Y_n} I_n(y)\log P_n(y) \\ &\leq \sum_{y\in Y_n} I_n(y)\log I_n(y). \end{aligned} \tag{6.12}$$

Fix $\epsilon > 0$, there is a finite $Y' \subset Y$, such that

$$\sum_{y\in Y'} P_{\lambda_0}(y)\log P_{\lambda_0}(y) \leq \sum_{y\in Y} P_{\lambda_0}(y)\log P_{\lambda_0}(y) + \epsilon. \tag{6.13}$$

With probability 1, when n is large enough, $Y_n \supset Y'$, then

$$\sum_{y\in Y_n} I_n(y)\log I_n(y) \leq \sum_{y\in Y'} I_n(y)\log I_n(y). \tag{6.14}$$

Letting $n \to \infty$, with probability 1,

$$\sum_{y\in Y'} I_n(y)\log I_n(y) \to \sum_{y\in Y'} P_{\lambda_0}(y)\log P_{\lambda_0}(y). \tag{6.15}$$

By (6.12)-(6.15),

$$\limsup \frac{1}{n}\sum_{i=1}^{n} \log P_n(y_i) \leq H.$$

Therefore,

$$\lim \frac{1}{n}\sum_{i=1}^{n} \log P_n(y_i) = H.$$

The above arguments also show that

$$\lim \frac{1}{n} \sum_{i=1}^{n} \log I_n(y_i) = H.$$

Then for large n,

$$\sum_{y \in Y_n} I_n(y) \log P_n(y) \geq \sum_{y \in Y_n} I_n(y) \log I_n(y) - \epsilon.$$

Since $\{P_n\}$ is a sequence of probability measures on the countable set Y, it contains convergent subsequences. Let $\tilde{P}$ be the limit of a convergent subsequence $\{P_{n_i}\}$. Then $\tilde{P}$ is a measure on Y with $\sum \tilde{P}(y) \leq 1$. From

$$\sum_{y \in Y_{n_i}} I_{n_i}(y) \log P_{n_i}(y) \geq \sum_{y \in Y_{n_i}} I_{n_i}(y) \log I_{n_i}(y) - \epsilon,$$

we get

$$\sum_{y \in Y} P_{\lambda_0}(y) \log \tilde{P}(y) \geq \sum_{y \in Y} P_{\lambda_0}(y) \log P_{\lambda_0}(y),$$

which can happen only if $\tilde{P} = P_{\lambda_0}$. Therefore any convergent subsequence of $\{P_n\}$ converges to P_{λ_0}. Therefore $P_n \to P_{\lambda_0}$. □

Corollary 5. If λ_0 is identifiable, i.e., for any $\lambda \neq \lambda_0$, $P_\lambda \neq P_{\lambda_0}$, then the estimation (6.9) is consistent, which means with probability 1, $\hat{\lambda}_n \to \lambda_0$ as $n \to \infty$. □

6.4 Pseudo-Likelihood (PL) Type Estimation of Parameters

The estimation procedures given in §6.3 are basically of maximum-likelihood type. They estimate the "global" distribution, i.e., the distribution on the set of all parse trees or the distribution on the set of all sentences. In the context of parsing, however, global distributions are irrelevant. What is really relevant for efficient parsing is that, given a sentence, all the possible parses of the sentence are properly assigned *conditional probabilities* so that the correct parses to the sentence are preferred in the sense that they have higher conditional probabilities. This observation suggests using the pseudo-likelihood (PL) type procedure for parameter estimation (Besag [2], [3]).

The idea for the PL estimation is as follows. Let (Ω, P) be a space. Suppose Ω is partitioned into disjoint subsets Ω_α. Then for each $\omega \in \Omega$, there is a unique Ω_α, denoted as $\Omega(\omega)$ such that $\omega \in \Omega(\omega)$. If we are given a parametric family of probability distributions $\{P_\theta\}_{\theta \in \Theta}$ and $P = P_{\theta_0}$, then for i.i.d. samples $\omega_1, \ldots, \omega_N$ from P, the PL estimate for θ_0 is

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \prod_{i=1}^{N} \left\{ P_\theta(\omega_i | \Omega(\omega_i)) \right\}.$$

Now let Ω be the set of all parses. In the context of parsing, we are interested in the comparison of all the parses for each single sentence, but not the comparison of parses for different sentences. Therefore, the partition we choose is such that, for $\omega \in \Omega$,

$$\Omega(\omega) = \{\omega' \in \Omega : Y(\omega') = Y(\omega)\}.$$

If $Y(\omega) = y$, then clearly, for any distribution P on Ω,

$$P(\omega|\Omega(\omega)) = P(\omega|Y(\omega) = y) = \frac{P(\omega)}{\sum\limits_{Y(\omega')=y} P(\omega')}$$

The global distribution of sentences is irrelevant for parsing, and we assume it to be $\pi(y)$, which might be unknown. The conditional probability distribution of parses, given a sentence y, is assumed to be a Gibbs distribution. In certain sense, the Gibbs distribution of parses, given y, should depend on y, i.e.,

$$P(\omega|Y(\omega) = y) = \frac{e^{\lambda_y \cdot f_y(\omega)}}{\sum\limits_{Y(\omega')=y} e^{\lambda_y \cdot f_y(\omega')}},$$

where λ_y are parameters depending on y and f_y are features depending on y. However, it is reasonable to assume that across all the sentences, the parsing rules are the same. Therefore, we suppose the conditional distributions have the same λ and the same f, for all y.

The distribution of all the parses then takes the form given by (6.5). Given i.i.d. samples $\omega_1, \ldots, \omega_n$, let $y_i = Y(\omega_i)$. The PL estimate is

$$\hat{\lambda}_n = \arg\max_{\lambda} \left\{ \prod_{i=1}^{n} P_\lambda(\omega_i|\Omega(\omega_i)) \right\} = \arg\max_{\lambda} \left\{ \prod_{i=1}^{n} \frac{e^{\lambda \cdot f(\omega)}}{\sum\limits_{Y(\omega)=y_i} e^{\lambda \cdot f(\omega)}} \right\},$$

or, using the notion of log-likelihood,

$$\hat{\lambda}_n = \arg\max_{\lambda} \left\{ \lambda \cdot \bar{f} - \frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{Y(\omega)=y_i} e^{\lambda \cdot f(\omega)} \right) \right\}. \tag{6.16}$$

Let $PL(\lambda, \omega_1, \ldots, \omega_n)$ be the function being maximized on the right hand side of (6.16). If the maximization has a solution $\hat{\lambda}_n$, then $\nabla PL(\lambda_0, \omega_1, \ldots, \omega_n) = 0$, i.e.,

$$\bar{f} = \frac{1}{n} \sum_{i=1}^{n} E_\lambda \left[f(\omega) | Y(\omega) = y_i \right]. \tag{6.17}$$

The formula (6.17) has an explanation which has nothing to do with the Gibbs form. If $|f(\omega)|$ has finite mean, then by the law of large numbers, as $n \to \infty$, with probability one, $\bar{f} \to E(f)$. On the other hand, for any sentence y,

$$\frac{|\{i : \; y_i = y\}|}{n} \to \pi(y),$$

and therefore,

$$\frac{1}{n}\sum_{i=1}^{n} E\left[f(\omega)|Y(\omega)=y_i\right] = \sum_{y\in Y}\frac{|\{i:\ y_i=y\}|}{n}E\left[f(\omega)|Y(\omega)=y_i\right] \to E(f).$$

In the above formula we omit the subscript of E_λ to make it clear that the distribution considered here is not necessarily given by (6.5).

If the true distribution belongs to a parametric family $\{P_\theta\}$ and its parameter is θ_0, then as n is large,

$$\bar{f} \approx \frac{1}{n}\sum_{i=1}^{n} E_{\theta_0}\left[f(\omega)|Y(\omega)=y_i\right],$$

and it is reasonable to let (any) solution of

$$\bar{f} = \frac{1}{n}\sum_{i=1}^{n} E_{\theta}\left[f(\omega)|Y(\omega)=y_i\right]$$

be an estimate of θ_0.

The estimation given by (6.16) is consistent in the following sense.

Proposition 19. Let $\{P_\lambda\}$ be a parametric family of probability distributions on Ω, such that for each λ,

$$P_\lambda(\omega) = \pi(Y(\omega))\frac{e^{\lambda\cdot f(\omega)}}{\sum\limits_{\omega'\in\Omega(\omega)} e^{\lambda\cdot f(\omega')}}.$$

Assume $\omega_1,\ldots,\omega_n$ are i.i.d. samples from P_{λ_0}. Let $\hat{\lambda}_n$ be the estimates given by (6.16). For each n, let $P_n = P_{\hat{\lambda}_n}$. If

$$-\sum_{y\in Y}\pi(y)\sum_{Y(\omega)=y} P_{\lambda_0}(\omega|Y(\omega)=y)\log P_{\lambda_0}(\omega|Y(\omega)=y) < \infty,$$

then with probability 1, for each sentence y, and for each ω with $Y(\omega)=y$,

$$P_n(\omega|Y(\omega)=y) \to P_{\lambda_0}(\omega|Y(\omega)=y),$$

Proof. With the similar arguments as in Proposition Proposition 18, it can be shown that with probability 1,

$$\frac{1}{n}\sum_{i=1}^{n}\log P_n(\omega_i|\Omega(\omega_i)) \to \sum_{y\in Y}\pi(y)\sum_{Y(\omega)=y} P_{\lambda_0}(\omega|\Omega(\omega))\log P_{\lambda_0}(\omega|\Omega(\omega))$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\log I_n(\omega_i|\Omega(\omega_i)) \to \sum_{y\in Y}\pi(y)\sum_{Y(\omega)=y} P_{\lambda_0}(\omega|\Omega(\omega))\log P_{\lambda_0}(\omega|\Omega(\omega)).$$

But

$$\frac{1}{n}\sum_{i=1}^{n}\log P_n(\omega_i|\Omega(\omega_i)) = \sum_{y\in Y} I_n(y) \sum_{Y(\omega)=y} I_n(\omega|\Omega(\omega))\log P_n(\omega|\Omega(\omega)),$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\log I_n(\omega_i|\Omega(\omega_i)) = \sum_{y\in Y} I_n(y) \sum_{Y(\omega)=y} I_n(\omega|\Omega(\omega))\log I_n(\omega|\Omega(\omega)).$$

For each y, since $I_n(y) \to \pi(y)$ and

$$\sum_{Y(\omega)=y} I_n(\omega|\Omega(\omega))\log P_n(\omega|\Omega(\omega)) \leq \sum_{Y(\omega)=y} I_n(\omega|\Omega(\omega))\log I_n(\omega|\Omega(\omega)),$$

we conclude that

$$\sum_{Y(\omega)=y} I_n(\omega|\Omega(\omega))\log P_n(\omega|\Omega(\omega)) - \sum_{Y(\omega)=y} I_n(\omega|\Omega(\omega))\log I_n(\omega|\Omega(\omega)) \to 0.$$

Now since the set $\{\omega :\ y(\omega) = y\}$ is finite, we get

$$P_n(\omega|\Omega(\omega)) \to P_{\lambda_0}(\omega|\Omega(\omega)).$$

The proof is complete. □

Corollary 6. If for each $\lambda \neq \lambda_0$, there is a y and an ω with $Y(\omega) = y$, such that $P_\lambda(\omega|Y(\omega) = y) \neq P_{\lambda_0}(\omega|Y(\omega) = y)$, then with probability one, $\hat{\lambda}_n \to \lambda_0$. □

Bibliography

[1] S. P. Abney. Stochastic Attribute-Value Grammars. *Computational Linguistics.* Accepted for publication.

[2] J. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems (with discussion). *J. Roy. Statist. Soc. Ser. B* 36, 192-236. 1974.

[3] J. Besag. Statistical Analysis of Non-Lattice Data. *The Statistician* 24, 179-195. 1975.

[4] P. A. Chou. Recognition of Equations Using a Two-Dimensional Stochastic Context-Free Grammar. *Visual Communications and Image Processing IV.* SPIE – The International Society for Optical Engineering. November, 1989.

[5] M. Johnson. NSF Grant Proposal. Department of Cognitive and Linguistic Sciences, Brown University. 1998.

[6] E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review* 106, 620-630. 1957.

[7] K. E. Mark. Markov Random Field Models for Natural Languages. PhD thesis. Department of Electrical Engineering, Washington University. May, 1997.

[8] S. C. Zhu, Y. N. Wu, and D. B. Mumford. Minimax Entropy Principle and Its Application to Texture Modeling. *Neural Computation* 9, 1627-1660. 1997.

Chapter 7

Scale Invariance of Natural Images

7.1 Introduction

Scale invariance refers to the phenomenon that the marginal distributions of many statistics of natural images are unchanged after the images get scaled. From the information point of view, scale invariance implies that even though individual natural images do change after being scaled, the information from the population of all the scaled natural images is no different than from the population of the original ones. As we shall see, scale invariance is a very robust property of natural images, and despite its simple form, we will argue that it is a non-trivial characteristic of natural images, and therefore an interesting natural phenomenon in its own right.

Scale invariance of natural images is of great interest in vision. It is widely believed that the statistical properties of natural images determine the basic aspects of the visual system. Scale invariance is among the most prominent statistical characteristics of natural images people have ever found. In Knill *et al.* [1], it was demonstrated that human visual system, when discriminating fractal images, is most sensitive to those which are approximately scale invariant. Because of the fractal nature of many texture images, this result suggests the role of scale invariance in texture discrimination by the visual system. Scale invariance is also an important ingredient in various theories on sensory coding. In Field [3], it was proposed that the visual system adopts a sparse coding scheme. One of the reasons that sparse representation is effective, the author argued, is that natural scenes are scale invariant.

Scale invariance has a lot of applications in computer science, especially in computer vision and image compression, and the reasons for this are very much the same as in vision.

This article is concerned with scale invariance of natural images itself rather than its importance to other areas of science. We will consider the following fundamental problem about scale invariance: Why are natural images scale invariant? To pursue an answer to this problem not only helps better understanding scale invariance, but also gives insight into statistical characterization of natural images. In the next sections, efforts are devoted to establishing a model on the origin of scale invariance of natural images.

A remark follows. Natural images are always presumed to be translation invariant, or stationary. Stationarity means that over the ensemble of natural images, the statistics at one location are the same as any other location. This is a reasonable assumption because, intuitively speaking, we can not observe "special" locations in images where the statistics tend to be peculiar. It implies that over the ensemble of natural images, all features have the same probability of occuring in one location versus another. From now on, when we say scale invariance of natural images, we always mean scale and translation invariance of these images.

There have been only a few models on the origin of scale invariance of natural images. One recent example is given in Rudderman [3]. According to this model, images are generated randomly by superimposing "objects" at random locations on a plane. Objects are planar patches with independent random shapes and sizes. Each object is also independently painted by a single random color. It was argued that if sizes of objects are distributed by a power-law, then images of the plane, when the plane is fully covered by the objects, have some scale invariant statistics.

Another model is presented in Mumford [4]. As the model in [3], images are made up of independent objects. Unlike that model, however, objects are patches of patterns, shadows, textons, etc., which means that within each object, the color is not a constant, but a function of location inside the object. The formation of images is by superimposing independent randomly scaled objects on a plane at random locations. The biggest difference between these two models lies in their explanations of the cause of scale invariance. In [3], it is the occlusion that is the main reason for scale invariance. On the other hand, in [4], only when occlusion is ignored, can images obtained in the above way be considered as scale invariant. The first model can get scale invariance only for some statistics, while the second one, when ignoring occlusion, guarantees scale invariance of all statistics.

Different as they are, both models consider objects as patches distributed on a plane. Such objects can only be considered as intermediate because they do not have clear physical meaning. After all, natural images are perspective projections of the real world, which is three dimensional, onto a planar surface. With high order approximation, it can be assumed that the projection is through an ideal camera in which the effects of diffraction, aberrations, and discrete sampling are absent. Because the world can be broken up into physical objects, it is therefore reasonable to presume that images consist of perspective projections, or 2D views, of the objects. A Poisson law is proposed as the law of distribution of objects in the three dimensional world. It is argued that the Poisson law of distribution of objects and the perspective projection of objects onto the camera image plane lead to approximate scale invariance of natural scenes. As in [4], only when the effects of occlusion are neglectable, can this argument be correct.

A by-product of our model is the representation of natural scenes as sums of wavelets. This representation was also proposed in [4]. As said earlier, the model proposed in this article gives wavelets a natural explanation. Another representation of natural scenes by sums of wavelets was given in [3]. However, it lacks the randomness which characterizes the wavelet representation derived from the Poisson model.

The article proceeds as follows. Section 7.2 discusses some evidence of scale invariance of natural images. We will argue that scale invariance is a very special property which separates natural images from other visual signals. Then we will formulate scale (and translation) invariance mathematically. In order to establish a model on the origin of scale invariance, we will first study some simple properties of scale invariant images. Section 7.3 gives one of such properties, which is the law of size of object in scale and translation invariant images. We will motivate the law by two arguments. Section 7.4 gives details of our model and section 7.5 concludes by showing the numerical results.

7.2 Evidence of Scale Invariance of Natural Images

This section presents evidence of scale invariance of natural images. But we will start by making clear what scaling for images is.

7.2.1 Scaling of Images

A (digitized) image I on an $M \times N$ lattice is simply a matrix with M rows and N columns. We adopt the convention of C language, so that the elements of I are represented by $I(i,j)$, where i is the row number running from 0 to $M-1$, and j is the column number running from 0 to $N-1$.

Scaling is achieved in the following way. To scale down an $M \times N$ image I by factor k, we take the disjoint $k \times k$ blocks $B_{ij} = [ik, (i+1)k-1] \times [jk, (j+1)k-1]$ in I and compute the average intensity value of each block. The average intensity of the block B_{ij} is taken as the intensity value at (i,j) in the down-scaled image. In mathematical terms, if $I^{(k)}$ denotes the down-scaled image, then it is a $\lfloor M/k \rfloor \times \lfloor N/k \rfloor$ matrix such that for $0 \le i \le \lfloor M/k \rfloor - 1$ and $0 \le j \le \lfloor N/k \rfloor - 1$,

$$I^{(k)}(i,j) = \frac{1}{k^2} \sum_{n=0}^{k-1} \sum_{m=0}^{k-1} I(ik+n, jk+m). \tag{7.1}$$

Why is scaling defined in this way? Naturally, we can imagine that every finite image is part of an infinite image, still denoted I, which is defined on the whole integer grid. Assume for each infinite image I, there is an underlying function $\phi(x,y)$ defined on $\mathbf{R}^2$, such that the value of $I(i,j)$ is the average of $\phi(x,y)$ over the square $S_{ij} = [id, (i+1)d] \times [jd, (j+1)d]$, where $d > 0$ is a constant, i.e.

$$I(i,j) = \frac{1}{d^2} \int_{S_{ij}} \phi(x,y) dx dy. \tag{7.2}$$

In other words, I is a digitized version of ϕ at "sampling rate" $1/d$. In order that the average of ϕ over S_{ij} makes sense, we assume ϕ is "regular", e.g., measurable. This is an ideal model, because in real images, pixel intensity values are responses of complicated filters to the visual signal. The filters may not be distributed on a square lattice, and their supports can overlap with each other.

Under the ideal model given by (7.2), for each $k \ge 1$, define an infinite image $I^{(k)}$ by (7.1), with i, j running through all integers. Then

$$I^{(k)}(i,j) \quad = \quad \frac{1}{k^2} \sum_{n=0}^{k-1} \sum_{m=0}^{k-1} I(ik+n, jk+m)$$

$$
\begin{aligned}
&= \frac{1}{k^2d^2}\sum_{n=0}^{k-1}\sum_{m=0}^{k-1}\int_{S_{ik+n,jk+m}}\phi(x,y)dxdy \\
&= \frac{1}{k^2d^2}\int_{[ikd,(i+1)kd]\times[jkd,(j+1)kd]}\phi(x,y)dxdy \\
&= \frac{1}{d^2}\int_{S_{ij}}\phi^{(k)}(x,y)dxdy,
\end{aligned}
$$

where

$$\phi^{(k)}(x,y) = \phi(kx,ky),$$

which demonstrates that $I^{(k)}(i,j)$ is the average of $\phi^{(k)}$ on S_{ij}. By the definition of scaling for functions defined on continuum, $\phi^{(k)}$ is the down-scaled by factor k version of f. It is therefore natural to define $I^{(k)}$ as the down-scaled by factor k version of the image I.

Scaling simulates two situations. Firstly, suppose a natural scene produces a (continuous) image $\phi(x,y)$ on a camera's image plane. Usually the distance between a natural scene and the camera is much larger than the focal distance of the camera, and therefore the camera image plane is almost located right at the focus. As the focal distance of the camera changes while the camera itself stands still, in order to get focused images of the same scene, the camera image plane needs to move closer or farther away from the camera lens, depending on whether the focal distance decreases or increases. In this case, in first order approximation, images produced on the image plane are scaled versions of each other. If the focal distance is k times smaller or k times larger, then the images are down-scaled or up-scaled by factor k, respectively. The error of the approximation lies in the fact that a natural scene is composed of objects with different distances from the camera. Only objects at a specific distance can produce truly focused images on the camera image plane. All other objects only produce blurred images. However, since both the diameter of the camera lens and the focal distance are much smaller than the distances of the objects from the camera, the blurring can be ignored. The second situation is called aperture imaging and is less familiar. The apparatus for aperture imaging is almost identical to a camera except that there is a tiny hole instead of a convex lens in the front of the apparatus to let light in. As the apparatus stands still while its image plane moves forward and backward, the images generated on the image plane, instead of being approximately scaled, as in the first situation, are truly scaled versions of each other.

One may think that if a natural scene is viewed from different distances, the images that it produces on the observer's retina or the camera's image plane will be scaled versions of each other. This is however incorrect. Because of perspective effects, as the observer or the camera gets closer to the scene, the nearer objects get larger faster than the farther objects. On the other hand, when the observer or the camera moves away from the scene, the nearer objects get smaller faster than the farther objects. Mathematically, if an object is originally at distance d, then as the observer or the camera moves farther away by distance x, the image of the object is down-scaled by factor $d/(d+x)$ which is a variable in d instead of a constant. This implies that images of objects are not scaled by a common factor, and hence the whole images are not scaled versions of each other.

7.2.2 Experiments

The images that we use are collected from the Internet. All the images are 256 × 256 matrices with integer intensity values between 1 and 256. Figure 6 shows six pictures in the collection.

It was reported in Zhu *et al.* [5] that the marginal distributions of x and y-derivatives of natural images are scale invariant. We conduct an experiment on our images and confirm the result. For digitized images, derivatives at a pixel are approximated by differences between the intensity values of the pixel and its neighboring pixels. For instance, at a pixel with location (i, j) in an image I, ∇_x and ∇_y are computed by

$$\begin{aligned}\nabla_x I(i,j) &= I(i,j+1) - I(i,j)\\ \nabla_y I(i,j) &= I(i+1,j) - I(i,j),\end{aligned}$$

Notice that i corresponds to the y coordinate while j corresponds to the x coordinate.

In order to get the empirical marginal distribution of derivatives, we first compute the histogram of derivatives for each image. Each histogram has 101 bins evenly dividing the interval $[-255, 255]$ and is normalized so that the sum of the histogram is 1. The average normalized histogram over all the images is then the empirical marginal distribution.

The results are presented in Figure 7.2. To demonstrate that the marginal distributions are really close to each other after images are scaled, we plot the logarithms of the marginal distributions. Figures 7.2a and b plot those of $\nabla_x I^{(k)}$, for $k = 2$ to 5, against $\nabla_x I$. Figures 7.2c and d plot those of $\nabla_y I^{(k)}$, for $k = 2$ to 5, against $\nabla_y I$.

From Figure 7.2, we can clearly see that the marginal distributions are almost unchanged to scaling. Notice the symmetry of the marginal distribution of $\nabla_x I$. The symmetry can be explained as the nature lacks obvious preference of the left over the right or vice versa. There is, however, no such apparent reason for the symmetry of the marginal distribution of $\nabla_y I$.

It is also noticeable that even many individual images have scale invariant marginal distribution. Figure 7.3 shows logarithms of normalized histograms of $\nabla_x I$ for the images in Figure 7.1. As can be seen, some of the histograms have strong scale invariance.

To see if the scale invariance we have observed is approximately independent of calibration, we generate, for each image I, a new image J by the following formula

$$J(i,j) = \log I(i,j).$$

Then we compute the marginal distributions of $\nabla_x J$. The results are given in Figure 7.4. Still, we observe strong scale invariance.

Not only differentiations, but also many other linear filterings produce responses that have scale invariant marginal distributions. We have tested two other kinds of linear filters. The first one is the isotropic center-surround filters, i.e., the Laplacian of Gaussian filters,

$$LG(x, y, s) = C \cdot (x^2 + y^2 - s^2) \exp\left(-\frac{x^2 + y^2}{s^2}\right),$$

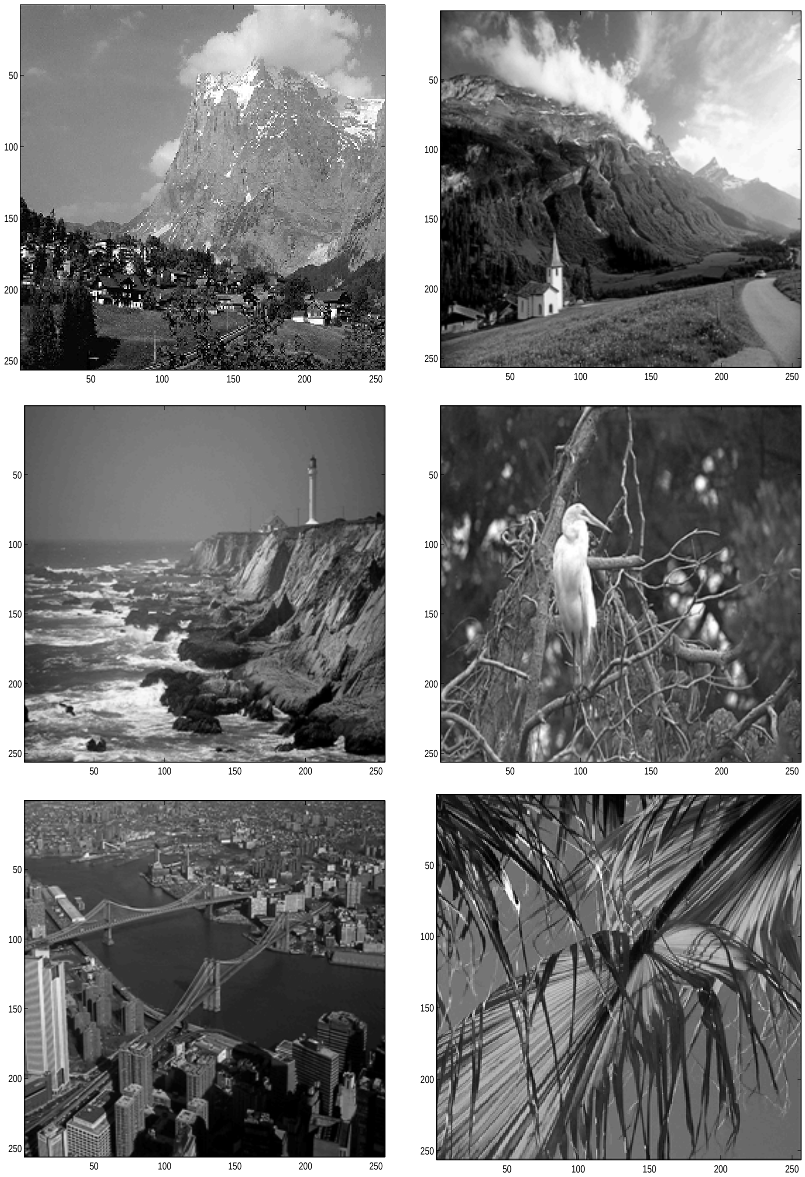
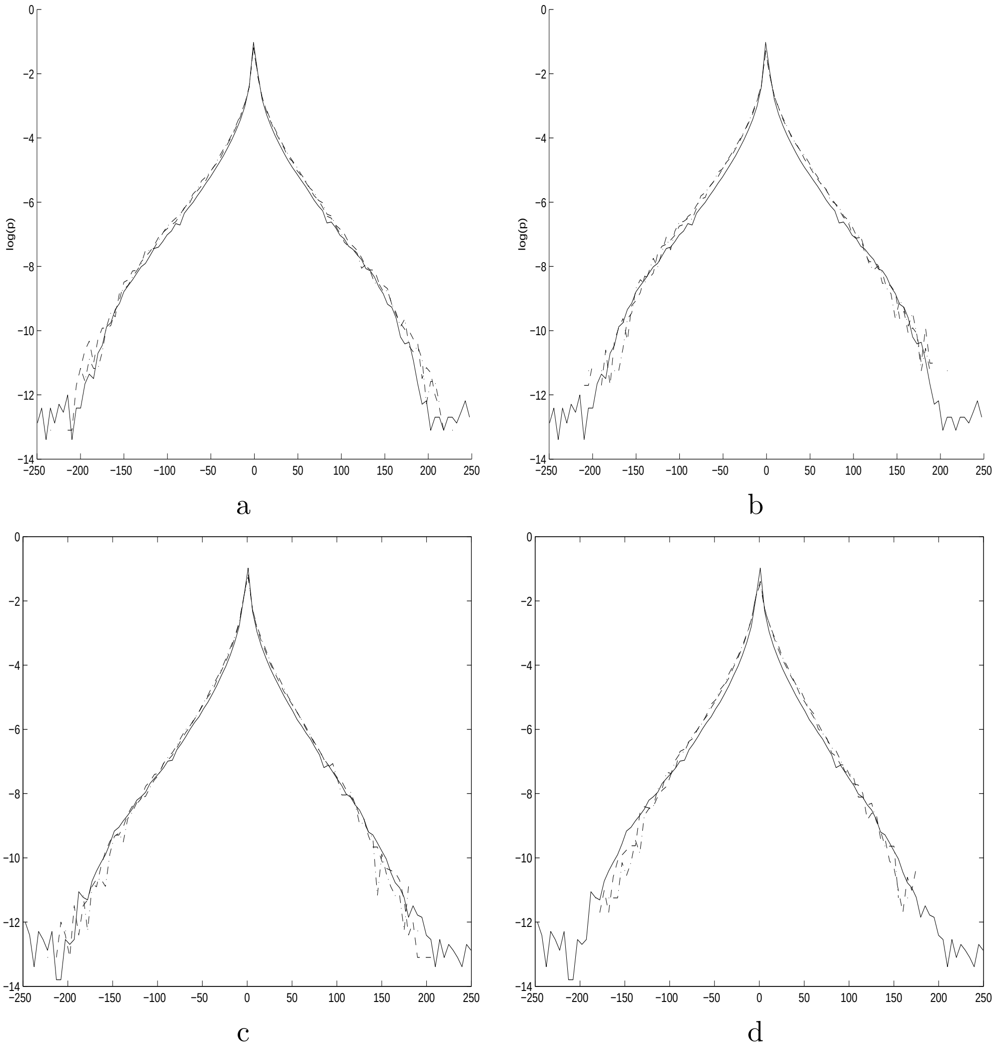


Figure 7.1: 6 out of the 30 collected images



a

b

c

d

Figure 7.2: Logarithms of marginal distributions of derivatives, solid curves are for unscaled images a. $\nabla_x I^{(k)}$, $k = 2$ (dashed), $k = 3$ (dash-dotted); b. $\nabla_x I^{(k)}$, $k = 4$ (dashed), $k = 5$ (dash-dotted); c. $\nabla_y I^{(k)}$, $k = 2$ (dashed), $k = 3$ (dash-dotted); d. $\nabla_y I^{(k)}$, $k = 4$ (dashed), $k = 5$ (dash-dotted).

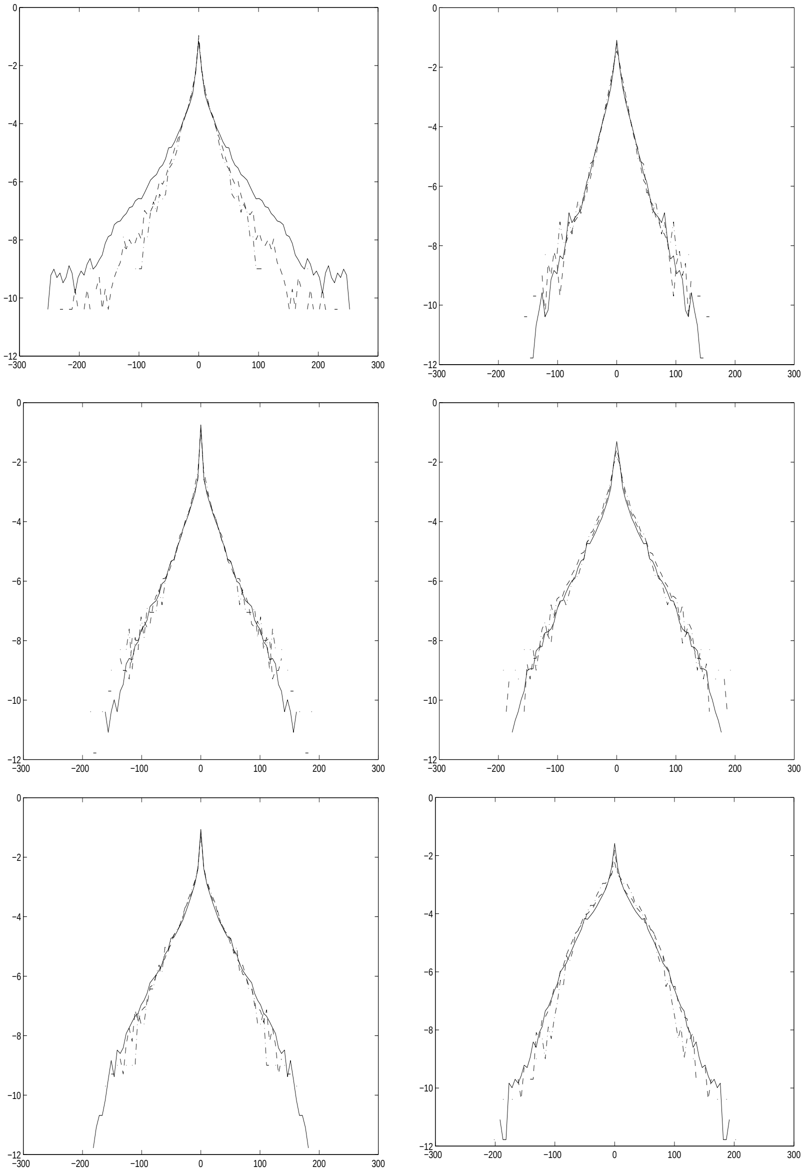


Figure 7.3: Logarithms of normalized histograms of $\nabla_x I^{(k)}$ for images in Figure 7.1, $k = 1$ (solid), $k = 2$ (dashed), and $k = 4$ (dash-dotted)

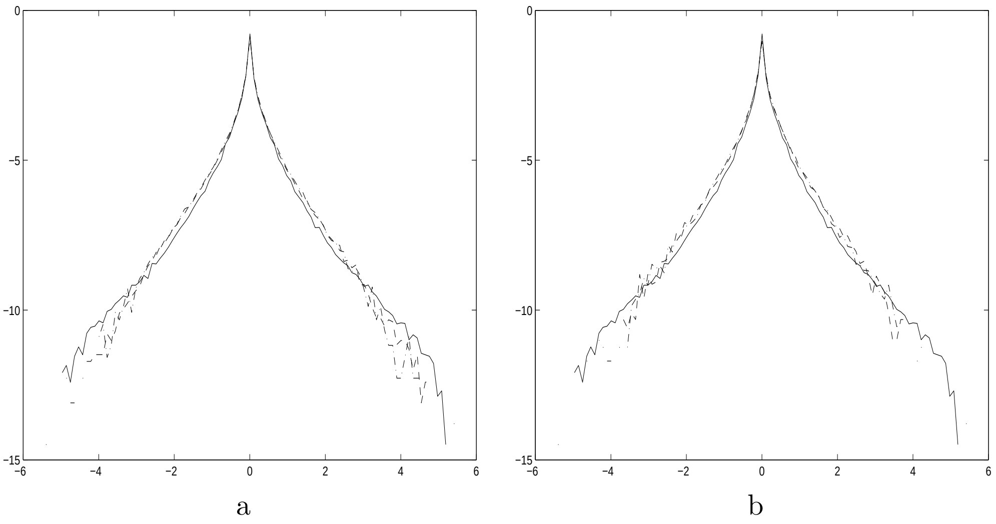


Figure 7.4: Logarithms of marginal distributions of $\nabla_x J^{(k)}$, $J = \log I$. a. $\nabla_x I^{(k)}$, $k = 2$ (dashed), $k = 3$ (dash-dotted); b. $\nabla_x I^{(k)}$, $k = 4$ (dashed), $k = 5$ (dash-dotted);

where C is a constant and s stands for the scale of the filter. We denote these filters by $LG(s)$. The second one is Gabor filters, which is defined as

$$G(x, y, s, \theta) = C \cdot \exp\left(-\frac{4z^2 + w^2}{2s^2}\right) \exp\left(-i\frac{2\pi z}{s}\right),$$

where

$$\begin{pmatrix} z \\ w \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

The real and image parts of the filters are denoted by $Gcos(s, \theta)$ and $Gsin(s, \theta)$, respectively.

In Figure 7.5, we plot logarithms of marginal distributions of responses to these two kinds of filters with different parameters, and again we observe scale invariance of the marginal distributions.

7.2.3 Discussion

That natural images have rich structures and scale invariant distributions makes them distinguished from noise signals. Firstly, Cauchy noise images do scale. Indeed, for a Cauchy noise image I, $I(i, j)$ are i.i.d. random variables with density function,

$$f(x) = \frac{1}{\pi}\frac{1}{1 + x^2}, \quad -\infty < x < \infty,$$

and characteristic function $\psi(u) = e^{-|u|}$. For each $k \geq 1$, $I^{(k)}(i, j)$ is the average of k^2 independent random variables from f. It is then seen that the characteristic function of $I^{(k)}(i, j)$ is still $e^{-|u|}$, implying $I^{(k)}$ and I have the same distribution, and therefore Cauchy

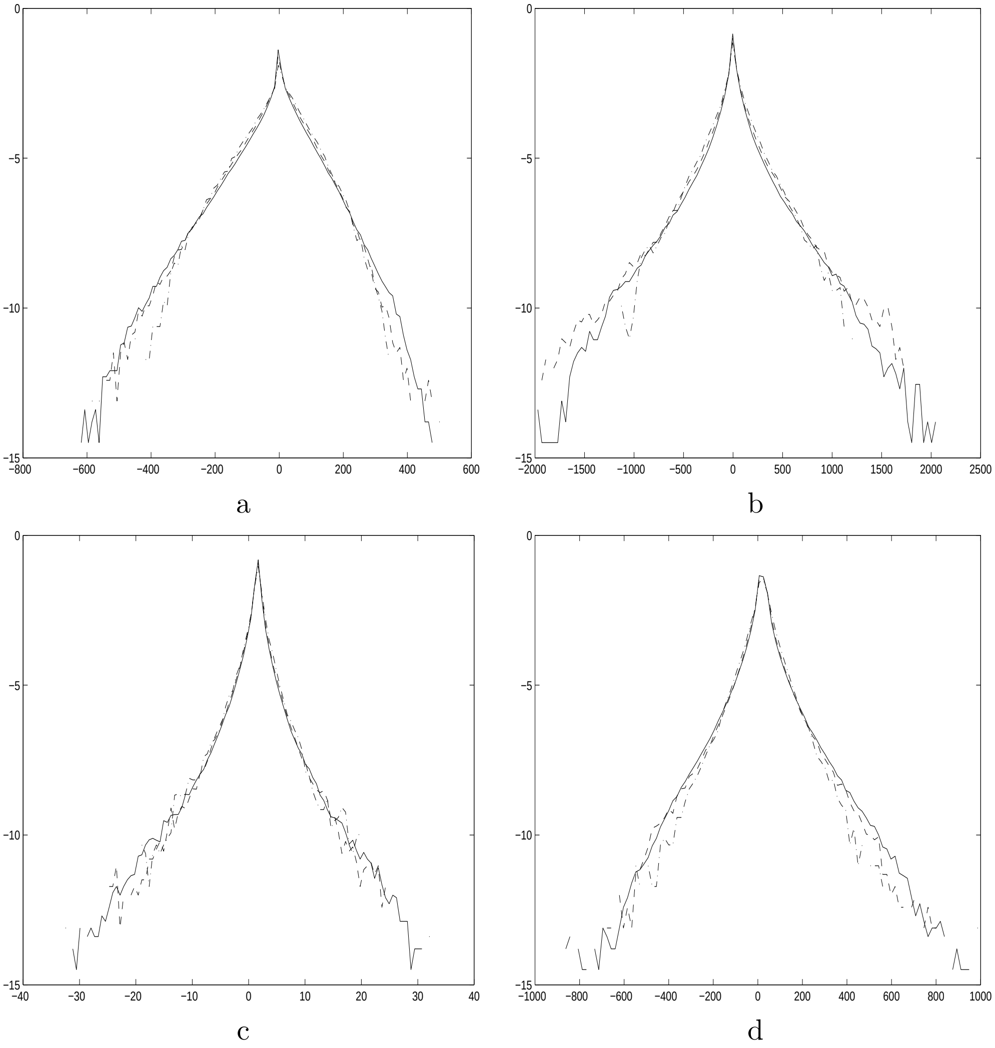


Figure 7.5: Logarithms of marginal distributions of $F * I^{(k)}$, $k = 1$ (solid), $k = 2$ (dashed), $k = 4$ (dash-dotted). a. $F = LG(2.5)$, b. $F = Gsin(4, 0)$, c. $F = Gcos(4, \pi/2)$, d. $F = Gcos(3, \pi/4)$

noise images are scale invariant. However, it is very easy to distinguish natural images from Cauchy noise images because the former ones always contain much richer structures.

Secondly, the marginal distribution of derivatives in white noise images is not scale invariant. Indeed, if $I(i,j)$ are i.i.d. $\sim \mathcal{N}(0,1)$, then for each $k \geq 1$, the marginal density function of $\nabla_x I^{(k)}$ is $\mathcal{N}(0, \sigma_k^2)$ with $\sigma_k = \sqrt{2}/k$. Having decreasing variance, the normalized histogram of $I^{(k)}$ becomes "narrower" as k increases. To see this, first note that the values of $\nabla_x I$ are not independent, because for each (i,j), $\nabla_x I(i,j) = I(i,j+1) - I(i,j)$ and $\nabla_x I(i,j+1) = I(i,j+2) - I(i,j+1)$ have dependency. However, $\nabla_x I(i,1)$, $\nabla_x I(i,3), \ldots \nabla_x I(i, 1+2s), \ldots$ are independent to each other. By the law of large numbers, the normalized histogram of $\{\nabla_x I(i, 1+2s)\}$ converges to the marginal distribution of $\nabla_x I(i,j)$ as the size of I goes to infinity. Similarly, the normalized histogram of $\{\nabla_x I(i, 2s)\}$ converges to the marginal distribution of $\nabla_x I(i,j)$ as the size of I goes to infinity. The normalized histogram of $\nabla_x I$ is the average of the two histograms and therefore tends to the marginal distribution of $\nabla_x(i,j)$. Since the marginal distribution is $\mathcal{N}(0, \sigma_k^2)$ with $\sigma_k = \sqrt{2}/k$, then the normalized histogram is increasingly concentrated around 0 as k increases.

One may point out that the numerical results we have shown do not directly involve intensity values of images and there might be some distribution μ, such that if $I(i,j)$ are i.i.d. $\sim \mu$, then I's are perceptually similar to natural images, and, even though I's themselves are not scale invariant, we still can get the same numerical results. However, we argue that this is unlikely to be true. We observed that the normalized histogram of $\nabla_x I$ is scale invariant, which, by an argument similar to last paragraph, implies the marginal distribution of $\nabla_x I(i,j)$ is scale invariant. Since

$$\nabla_x I(i,j) \overset{\mathcal{D}}{=} \nabla_x I^{(k)}(i,j),$$

rewriting both sides in terms of differences between pixel values, there is

$$I(i,j+1) - I(i,j) \overset{\mathcal{D}}{=} \frac{1}{k^2} \sum_{n=0}^{k-1} \sum_{m=0}^{k-1} [I(ik+n, (j+1)k+m) - I(ik+n, jk+m)].$$

Because $I(ik+n, (j+1)k+m) - I(ik+n, jk+m)$, $0 \leq n, m \leq k-1$ are independent to each other and have the same distribution as $\nabla_x I(i,j) = I(i,j+1) - I(i,j)$, the distribution of $\nabla_x I(i,j)$ is not only infinitely divisible but also a Cauchy distribution. We then get that the sub-image $\{\nabla_x I(i, 1+2j)\}$ is a Cauchy noise image. However, this is not the case for natural images. Because even in the subsample $\{\nabla_x I(i, 1+2j)\}$ of a natural image, we can observe a lot of structures.

All the experiments we have conducted are on images defined on finite lattice. However, we have seen it is convenient and natural to consider images as defined on $\mathbf{R}^2$. From now on, we use $\phi(x,y)$ to represent an image defined on $\mathbf{R}^2$ and $I(i,j)$ a digitized image defined on a finite or infinite integer lattice.

For digitized images $I(i,j)$, only down-scaling by an integer factor is appropriate. Up-scaling and scaling by a non-integer factor are not well defined. However, because across the ensemble of digitized natural images, we observe scale invariance of many filter responses, no matter how high the image sampling rate is, it makes sense to think that the underlying continuous images have marginal distributions invariant to scaling.

We formulate scale invariance of natural images as follows. Recall we always implicitly require stationarity of images.

Definition. Let $\mathcal{E}$ be a space of functions defined on $\mathbf{R}^2$ (think of $\mathcal{E}$ as the ensemble of natural images), such that for any $\phi \in \mathcal{E}$, any $\lambda > 0$ and any $(a, b) \in \mathbf{R}^2$, $\phi(\lambda x+a, \lambda y+b) \in \mathcal{E}$. A probability distribution on $\mathcal{E}$ is scale and translation invariant if for any $\lambda > 0$ and $(a, b) \in \mathbf{R}^2$,

$$\phi(\lambda x + a, \lambda y + b) \overset{\mathcal{D}}{=} \phi(x, y).$$

We need to fill a gap between the observation and our claim. We have observed that many filterings produce responses that have scale invariant marginal distributions. It is natural then to speculate that *all* filterings produce responses which have scale invariant marginal distributions [1]. But why does this imply that the distribution of natural images itself is scale invariant? Indeed, if ϕ is an image and F is a linear filter, than the filter response of ϕ to F is the convolution $F * \phi$ on $\mathbf{R}^2$,

$$F * \phi(x, y) = \int \phi(u, v)F(x - u, y - v)du\ dv.$$

Assuming ergodicity of the distribution of natural images, with probability one, the histogram of $F * \phi$ is the distribution of $F * \phi(0, 0)$, in the sense that, for any $a < b$, as $M \to \infty$,

$$\frac{1}{4M^2} m(\{(x, y) \in [-M, M]^2 : F * \phi(x, y) \in [a, b)\}) \to \text{Prob}(\langle \phi, \bar{F} \rangle \in [a, b)),$$

where $m(\cdot)$ is the Lebesgue measure. But $F * \phi(0, 0) = \langle \phi, \bar{F} \rangle$, where $\bar{F}(x, y) = F(-x, -y)$. Since the histogram of $F * \phi$ is scale invariant, the distribution of $\langle \phi, \bar{F} \rangle$ is scale invariant. Together with the always implicitly assumed stationarity, this leads to

$$E(e^{i\langle \phi(x,y), \bar{F}(x,y) \rangle}) = E(e^{i\langle \phi(\lambda x+a, \lambda y+b), \bar{F}(x,y) \rangle}).$$

If this is true for all filters, then the *characteristic functional* of the probability distribution on images is scale and translation invariant. Since a probability distribution on images is uniquely determined by its characteristic functional, the distribution is scale and translation invariant.

In section 7.2.2 we mentioned that scale invariance of marginal distribution of derivatives be approximately independent of calibration. Indeed, if the sampling rate of a digitized image I is high, for each (i, j), the underlying continuous image ϕ is about constant over the square $S_{ij} = [id, (i + 1)d] \times [jd, (j + 1)d]$, where d is the inverse of the sampling rate. Thus, for any smooth calibration κ,

$$\kappa(I(i, j)) = \kappa \left(\frac{1}{d^2} \int_{S_{ij}} \phi(x, y)dx\ dy \right) \approx \frac{1}{d^2} \int_{S_{ij}} \kappa \circ \phi(x, y)dx\ dy.$$

If $\phi(x, y) \overset{\mathcal{D}}{=} \phi(\lambda x, \lambda y)$, then $\kappa \circ \phi(x, y) \overset{\mathcal{D}}{=} \kappa \circ \phi(\lambda x, \lambda y)$. Letting $J = \kappa(I)$, from the above approximation, we get

$$\left\{ J^{(k)}(i, j) \right\} \approx \left\{ \frac{1}{d^2} \int_{S_{ij}} \kappa \circ \phi(kx, ky)dx\ dy \right\} \overset{\mathcal{D}}{=} \left\{ \frac{1}{d^2} \int_{S_{ij}} \kappa \circ \phi(x, y)dx\ dy \right\} \approx \{J(i, j)\}.$$

[1] Strictly speaking, the "raw intensity" of images is not expected to be scale invariant.

Therefore, the distribution of J is approximately scale invariant, verifying our suggestion that the scale invariance of marginal distributions is approximately independent of calibration.

Finally, we establish a connection between the version of scale invariance given in the above definition and a result on scale invariance in the literature. It is well known that natural images have power spectrum of the form [3]

$$S(k) = \frac{A}{k^{2-\eta}},$$

where k is the spatial frequency, A is a constant, and η is close to 0. $S(k)$ is defined as

$$S(k) = \frac{1}{2\pi}\int_0^{2\pi} d\theta \int_{\mathbf{R}^2} \langle \phi(\mathbf{x})\phi(\mathbf{x}+\mathbf{y})\rangle e^{-ik\mathbf{v}(\theta)\cdot\mathbf{y}} d^2y,$$

where for fixed $\mathbf{y}$, $\langle \phi(\mathbf{x})\phi(\mathbf{x}+\mathbf{y})\rangle$ is the average of $\phi(\mathbf{x})\phi(\mathbf{x}+\mathbf{y})$ over all $\mathbf{x}$ and all ϕ, and $\mathbf{v}(\theta) = (\cos\theta, \sin\theta)$. The ideal case is that $\eta = 0$. To see the reason for this, note that under the assumption of ergodicity of the distribution of images,

$$\langle \phi(\mathbf{x})\phi(\mathbf{x}+\mathbf{y})\rangle = E(\phi(0)\phi(\mathbf{y}))$$

where E is over all ϕ. Therefore, by scale invariance,

$$\begin{aligned}
S(k) &= \frac{1}{2\pi}\int_0^{2\pi} d\theta \int_{\mathbf{R}^2} E(\phi(0)\phi(\mathbf{y}))e^{-ik\mathbf{v}(\theta)\cdot\mathbf{y}} d^2y \quad (\mathbf{z} = k\mathbf{y}) \\
&= \frac{1}{k^2}\frac{1}{2\pi}\int_0^{2\pi} d\theta \int_{\mathbf{R}^2} E(\phi(0)\phi(k^{-1}\mathbf{z}))e^{-i\mathbf{v}(\theta)\cdot\mathbf{z}} d^2z \quad (\phi(k^{-1}\mathbf{z}) \sim \phi(\mathbf{z})) \\
&= \frac{1}{k^2}\frac{1}{2\pi}\int_0^{2\pi} d\theta \int_{\mathbf{R}^2} E(\phi(0)\phi(\mathbf{z}))e^{-i\mathbf{v}(\theta)\cdot\mathbf{z}} d^2z \\
&= \frac{S(1)}{k^2}.
\end{aligned}$$

7.3 The $\frac{1}{r^3}$ Law of Size of Object

Our goal is to build a model on the origin of scale (and translation) invariance of natural images. As a first step to this goal, we consider the distribution of sizes of objects in images. Let r be the one dimension size of object, such as diameter and periphery. As a first order approximation, the density function of r is Cr^{-3}, where C is a constant. There are several arguments to get this the result. We will demonstrate two of them. A third argument, which is based on compositional rules, can be found in Geman [6]. We start from Mumford's line segment argument.

7.3.1 Poisson Line Segment Argument

We consider images composed only of straight line segments with finite lengths. A random image is generated in the following way. First, we produce a sample $\{(x_i, y_i)\}$ from a homogeneous Poisson point process on $\mathbf{R}^2$. With probability 1, the sample is countable. For each (x_i, y_i), we independently sample an r from a distribution with density function $f(r)$ and an angle θ uniformly from $[0, \pi]$. We then put a line segment with length r and orientation $(\cos\theta, \sin\theta)$ at (x_i, y_i), with (x_i, y_i) being its middle point. All the random line segments then compose an image I.

Now suppose I produced by the above random procedure are distributed by a scale and translation invariant law. We want to know the form of $f(r)$, which is the law of size of object in this case.

In order to get $f(r)$, define a function $N(a, b; R)$ such that for any $0 < a < b$, $N(a, b; R)$ is the expected number of line segments of I with midpoints falling into the square $S_R = [0, R] \times [0, R]$ and lengths between a and b. Define $N_2(a, b; R)$ similarly for $I^{(2)}$. Since $I \sim I^{(2)}$, we get

$$N(a, b; R) = N_2(a, b; R).$$

Because $I^{(2)}$ is a down-scaled by factor 2 version of I, any line segment of $I^{(2)}$ contained in S_R is a down-scaled by factor 2 version of a line segment of I in the square S_{2R}, and the latter line segment has length twice larger than the first one. We get

$$N_2(a, b; R) = N(2a, 2b; 2R).$$

The square S_{2R} consists of four disjoint squares, each being identical to S_R. Because the random processes involved to generate the images are homogeneous, we get

$$N(2a, 2b; 2R) = 4N(2a, 2b; R) \Rightarrow N(a, b; R) = 4N(2a, 2b; R).$$

On the other hand, we have

$$N(a, b; R) \propto \int_a^b f(r)dr, \quad a < b.$$

Therefore,

$$\int_a^b f(r)dr = 4\int_{2a}^{2b} f(r)dr.$$

More generally, we can replace 2 by any positive number s to get

$$\int_a^b f(r)dr = s^2 \int_{sa}^{sb} f(r)dr.$$

Differentiating with respect to b, we finally get

$$f(r) = s^3 f(sr) \Rightarrow f(r) = \frac{C}{r^3}.$$

It is obvious that $1/r^3$ can not be a density function because it is singular at 0. The problem comes from the assumption that the images can be up-scaled by any factor. This assumption implies that the density of the Poisson point process must be infinity which is impossible. On way to fix the problem is to require that the line segments have lengths larger than a threshold, say, ϵ and only down-scaling of images be allowed. When an image is down-scaled, all line segments with lengths less than ϵ are thrown away. Then with the same argument, we still can get the $1/r^3$ law, except that r should be larger then ϵ.

7.3.2 Coding Theory Argument

Still, we consider images composed of line segments and assume that the images are scale and translation invariant. Our first step is to discretize the images so that end points of digitized line segments are on the lattice $\{(nd_N, md_N)\}_{n,m\in\mathbf{Z}}$, where $d_N = 1/N$ is the resolution.

We want to compare the probability of a digitized line segment l and the probability of its down-scaled by factor k version $l^{(k)}$. To this end we code digitized line segments by a k-ary code. A k-ary code is of the form $a_{n-1}\ldots a_1a_0$, where $a_i \in \{0,\ldots,k-1\}$ and $a_{n-1} > 0$. For each line segment l, let $c(l;k)$ be its k-ary code. If the coding is optimal, then by Shannon's theorem,

$$\mathrm{Prob}(l) = \frac{1}{k^{|c(l;k)|}},$$

where $|c(l;k)|$ is the length of the code $c(l;k)$.

Suppose the end points of $l^{(k)}$ are (m_1d_N, n_1d_N) and (m_2d_N, n_2d_N). Then each of the k^4 line segments with end points $((km_1+i_1)d_N, (kn_1+j_1)d_N)$ and $((km_2+i_2)d_N, (kn_2+j_2)d_N)$, $i_1, i_2, j_1, j_2 = 0,\ldots,k-1$, has $l^{(k)}$ as its down-scaled by factor k version. As N is large enough, $d_N = 1/N$ is small and all the k^4 line segments are spatially close to each other. By continuity, the information about these k^4 line segments is evenly distributed. Therefore, given the k-ary code of $l^{(k)}$, in order to get the whole information about l, we need $\log_k k^4 = 4$ extra bits. This implies

$$|c(l;k)| \approx 4 + |c(l^{(k)};k)| \Rightarrow \mathrm{Prob}(l) \approx \frac{1}{k^4}\mathrm{Prop}(l^{(k)}).$$

For any line segment ℓ on $\mathbf{R}^2$, let l_N be its digitized version at resolution d_N. Then $\ell^{(k)}$ is digitized as $l_N^{(k)}$ at resolution $1/d_N$. We then have

$$\lim_{N\to\infty} \frac{\mathrm{Prob}(l_N)}{\mathrm{Prob}(l_N^{(k)})} = \frac{p(\ell)}{p(\ell^{(k)})} \Rightarrow \frac{p(\ell)}{p(\ell^{(k)})} = \frac{1}{k^4},$$

where $p(\ell)$ is the density of ℓ. The above relation holds for any line segment and any positive integer scaling factor. By continuity, it holds for an arbitrary positive scaling factor. Because orientations of line segments are uniformly distributed and the distribution of images is translation invariant, for any line segments l_r and l_{sr} with lengths r and sr, respectively,

$$p(l_r) = s^4 p(l_{sr}) \Rightarrow p(l_r) = \frac{C}{r^4}.$$

In order to get the marginal distribution of r, we integrate the density function over all line segments with length r and with one end point at the origin. The other end points of these line segments are on a circle with radius r. Therefore the marginal distribution of r is

$$f(r) = 2\pi r \cdot \frac{C}{r^4} = \frac{C'}{r^3}.$$

7.4 The Poisson Model

In this section we build a Poisson model on the origin of scale and translation invariance of natural images. The model has two components (1) distribution of objects in the 3D world, and (2) surface processes that describe intensity distributions inside 2D views of objects. We start by modeling objects of 3D world and after establishing the Poisson model, we will discuss implications and limitations of the model.

7.4.1 Modeling Objects

As discussed in section 7.1, natural images are perspective projections of the 3 dimensional world on an image plane. World breaks up into physical objects with different shapes, surface colors and sizes and so natural images also break up into the viewed surfaces of objects. Thus the first problem that comes up is how to model physical objects.

As a coarse approximation, objects in our model are independent rigid planar templates parallel to the image plane. Each template has a reference point. The position of an object is the spatial location of its reference point.

Several correlated important aspects of real objects are ignored in our model.

(1) Occlusion and orientation of object. The surface of a 3D object always has several different aspects. Because of occlusion by the other aspects of the same surface, an aspect which is visible when the object is at one place can become invisible when the object moves to another place. Even when objects are modeled as planar templates, if an object is not parallel to the camera image plane, then because of perspective effect, when the object moves on a plane parallel to the camera image plane, the farther away it moves from the camera laterally, the larger its 2D view becomes. This effect is not accounted for by our argument. In real situation, since the angle of view of a camera is usually small, the effects incurred by occlusion and orientation are small. In our model, however, we allow an arbitrary large, but fixed angle of view. In order to avoid complications, we require templates be parallel to the image plane.

(2) Dependence between objects. It often happens that in a large region of the world, objects have long range dependence. For example, each window on a building can be considered as an object. The position of a window is obviously not independent to the positions of the other windows on the same building. A solution to this is to define two objects as two parts of a larger object whenever they have dependence. But this can also cause problem. For instance, houses built by a street stand approximately along a straight line. If all the houses are put together as a single object, then the perspective effect on the 2D view of the object, as mentioned in (1), can not be ignored. Another example is a long river flowing on a plain. It is even hard to model it as a template parallel to the image plane.

(3) Volume of object. Because a real object has a certain volume, when it occurs at a location in the space, other objects can not occupy space arbitrarily close to it. On the other hand, since a template is a planar shape with zero volume, other templates can get arbitrarily close to it.

Let us describe the 3D world by a Euclidean coordinate system. Every point in the world is represented by (x, y, z). Assume that the camera lens is at the origin of the coordinate system. The direction of view of the camera is along the positive x-axis. Suppose the distance between the camera image plane and the lens is 1. Then the image plane is the plane $\{(-1, y, z) : \ y, z \in \mathbf{R}\}$. We also need a coordinate system for the image plane. Let every point on the image plane be represented by (u, v), and let the origin of the image plane coordinate system be the intersection point of the x-axis and the image plane, which is the space point $(-1, 0, 0)$. Because the perspective projection of an object is upside down and left and right reversed, we define the direction of the u-axis as the opposite direction of the y-axis, and the direction of the v-axis as the opposite direction of the z-axis. Then the projection of a spatial point (x, y, z), $x > 0$ is $(y/x, z/x)$ on the image plane.

7.4.2 Distribution of Objects

Under the set-ups of section 7.4.1, we make the following assumption on the distribution of objects in the world.

Assumption 1. Objects are distributed by a homogeneous Poisson law.

We must decide the support of the Poisson law, i.e., the region in which an object can be any where with positive probability. Let us show that the support can not be the whole 3D space. For simplicity of discussion, for now we assume all the objects are identical, i.e., they share the same template. Refer to Figure 7.6. By our set-ups, D in the figure is 1. If the distance between an object with size R and the camera lens is d, then the size of the projection of the object is $RD/d \propto 1/d$. Letting r be the size of the projection, we have

$$r \propto \frac{1}{d}.$$

Now we derive the probability density function $f(r)$. As in section 7.3.1, fix a finite square on the image plane. Images in the square are projections of an infinite cone in the space, illustrated as the shaded area in Figure 7.6. All objects in the cone with distance d from the camera are on a planar region with area proportional to d^2. Assume the distribution of objects is homogeneous 3D Poisson, then the density $g(d)$ of objects in the cone with distance d is also proportional to d^2. From $r \propto 1/d$ and $g(d) \propto d^2$, we get the law of size of object $f(r) \propto 1/r^4$. This is inconsistent with the result in section 7.3, where $f(r) \propto r^{-3}$. The heuristic argument suggests that the support of the Poisson law be assumed other than the whole 3D space. In other words, objects should be modeled as being distributed in a sub-region in the 3D space by a homogeneous Poisson law.

We have to look at the nature more closely. Natural images are taken on the earth. The surface of the earth, within our visible distance, is flat. Although objects can be any where above the ground, they are overwhelmingly distributed below a certain altitude. Therefore, we modify our previous assumption to the following.

Assumption 2. There is a constant $H > 0$, such that objects are distributed by a homogeneous Poisson law in the region between the earth and the height H.

Let the plane $z = 0$ represent the earth. Let $\{(x_i, y_i, z_i)\}$ be the positions of objects. Then

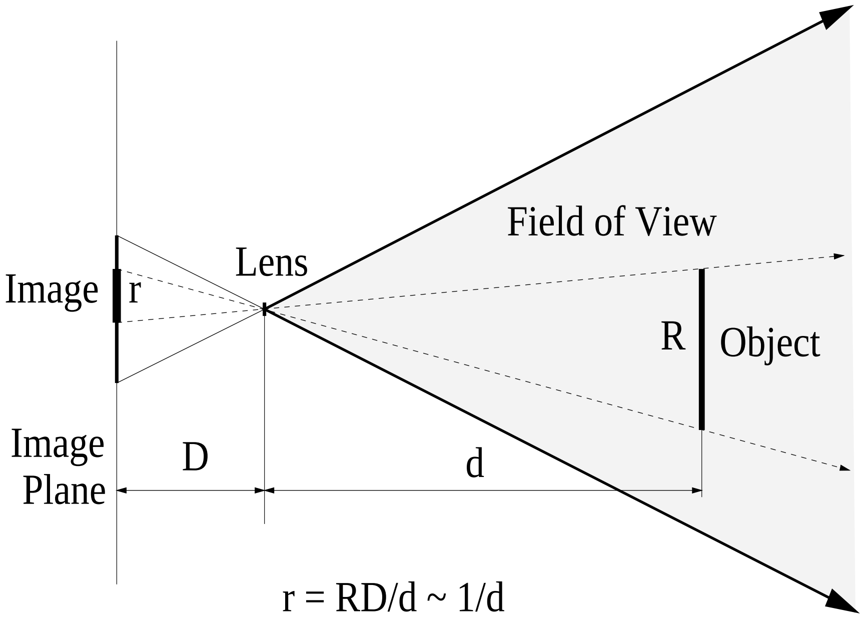


Figure 7.6: Perspective view of an object

the assumption means that $\{(x_i, y_i, z_i)\}$ is a sample from a homogeneous Poisson process in the region $\mathbf{R}^2 \times [0, H] = \{(x, y, z) : \ x, y \in \mathbf{R}, 0 \le z \le H\}$.

To see this assumption is consistent with the r^{-3} law of size of object, refer to Figure 7.6 again. When d is large enough, any object in the shaded area with distance d is on a rectangular planar region with width proportional to d and with fixed altitude H. Thus $g(d) \propto d$ and this together with $r \propto 1/d$ leads to $f(r) \propto 1/r^3$.

7.4.3 Surface Processes of Objects

At different distances, the 2D view of an object not only has different sizes, but also shows different surface colors, textures, etc. The distribution of intensity values inside the 2D view of an object, which we want to call "surface process", can be very complicated. It can be not only a smooth function, but also an "irregular" function, e.g., a sample from a random process like white noise. Such irregular functions are called generalized functions in mathematical terms. In any case, the surface process inside the projection of an object is a function of $\mathbf{u} = (u, v)$ on the image plane, with support inside the projection. We use $\psi(\mathbf{u}; \mathbf{x}, T)$ to represent the surface process inside the projection of a template T which is located at $\mathbf{x}$.

Given a template T, suppose P is a point on T with location relative to the reference point of T being $\mathbf{v} = (a, b)$. If T is located at $\mathbf{x} = (x, y, z)$, then the spatial location of P is $\mathbf{y} = (x, y + a, z + b)$ and its projection on the image plane is

$$\mathbf{u} = \left(\frac{y + a}{x}, \frac{z + b}{x} \right).$$

Under ideal conditions, where the effects of decay, scattering, interference, diffraction, etc., as light travels in the space, are absent, and where the camera is ideal, if there are no other objects between P and $\mathbf{u}$, the intensity at $\mathbf{u}$ equals the intensity of the light setting off from P to $\mathbf{u}$. The intensity of the light depends not only on P's own physical condition, which we assume to be unchanged no matter where T is, but also on the lighting condition around the spatial location of P as well as the direction the light goes. We assume that the lighting condition is uniform all over the space. We also assume that the intensity of light from P is constant on all directions. Then the light that goes from P to $\mathbf{u}$ has intensity depending only on P but not on its location in the space. This implies that the light intensity is a function only on the relative location of P on T, and therefore the intensity at $\mathbf{u}$ is determined by $\mathbf{v}$. By our notations, this can be written as

$$\psi(\mathbf{u}; \mathbf{x}, T) = I(\mathbf{v}; T).$$

Write $\mathbf{p} = (y, z)$. Then $\mathbf{x} = (x, \mathbf{p})$ and $\mathbf{u} = x^{-1}(\mathbf{p} + \mathbf{v})$. Therefore,

$$\psi(\mathbf{u}; (x, \mathbf{p}), T) = I(x\mathbf{u} - \mathbf{p}; T).$$

From the equation it is seen that the surface process of a template object T located at $(x, \mathbf{p})$ in space is a scaled and translated version of I. We call this change of surface process by location "color rendering".

We now consider the whole picture. Because the distribution of objects is homogeneous Poisson in $\mathbf{R}^2 \times [0, H]$, with probability one, there are countably many objects in the region $\mathbf{R}^+ \times \mathbf{R} \times [0, H]$. Let the positions of these objects be $\{(x_i, \mathbf{p}_i)\}$, $\mathbf{p}_i = (y_i, z_i)$. At each location $(x_i, \mathbf{p}_i)$, a template T_i is independently selected from a certain distribution. Writing $I_i(\mathbf{u}) = I(\mathbf{u}; T_i)$, then I_i are i.i.d. *If we ignore occlusion, then the whole image I is the arithmetic sum of the projections of all the objects and can be written as*

$$I(\mathbf{u}) = \sum_i I_i(x_i\mathbf{u} - \mathbf{p}_i), \quad I_i \text{ i.i.d.} \tag{7.3}$$

7.4.4 Discussion

The first consequence of the above results is as follows. Fix a large enough rectangular image I. For $0 < a < b$, define

A_a^b = Expected total area of regions in I which are covered by projections of objects with distance from the camera between a and b.

If we ignore occlusion, then for any $\lambda > 0$, $A_{\lambda a}^{\lambda b} = A_a^b$.

Call an object a T-object if the object is a template T. If a T-object is x away from the lens, then its projection has area proportional to x^{-2}. By the Poisson distribution, the density of T-objects showing up in I with distance from the camera being x is proportional to x. Without occlusion, the total area of the projections of T-objects with distance between a and b is proportional to

$$\int_a^b \frac{1}{x^2} x dx = \log \frac{b}{a}.$$

Integrating over all possible templates, we get, without occlusion,

$$A_a^b \propto \log \frac{b}{a} \Rightarrow \text{for all } \lambda > 0, A_{\lambda a}^{\lambda b} = A_a^b.$$

The same argument also applies to explain the scale invariance of marginal distribution of derivatives. Without loss of generality, consider ∇_u. Given a template T, let

$$\begin{aligned} D_a^b(x;T) &= \text{Total area of regions in the projection of } T \text{ where values of } \nabla_u \text{ are} \\ & \quad\ \text{between } a \text{ and } b \text{ when the distance between } T \text{ and the camera is } x. \end{aligned}$$

Since the surface process inside the projection of a T-object with location $\mathbf{x} = (x, \mathbf{p})$ is $\psi(\mathbf{u}; \mathbf{x}, T) = I(x\mathbf{u} - \mathbf{p}; T)$, the derivative at $\mathbf{u}$ is $x\nabla_u I(x\mathbf{u} - \mathbf{p}; T)$, therefore

$$\nabla_u \psi(\mathbf{u}; (x, \mathbf{p}), T) \in [a, b] \Leftrightarrow \nabla_u I(x\mathbf{u} - \mathbf{p}; T) \in [ax^{-1}, bx^{-1}].$$

By the fact that the area of the projection of the T-object is proportional to $1/x^2$,

$$D_a^b(x;T) = \frac{1}{x^2} D_{a/x}^{b/x}(1;T).$$

Let D_a^b be the expected total area of regions in I where $\nabla_u I$ is between a and b. Neglecting occlusion and integrating the above equation over all $x \in (0, \infty)$ and all T, D_a^b is proportional to

$$\begin{aligned} \int d\mu(T) \int_0^\infty D_a^b(x;T) g_T(x) dx &= \int d\mu(T) \int_0^\infty \frac{1}{x^2} D_{a/x}^{b/x}(1;T) c_T x dx \\ &= \int_0^\infty \frac{1}{x} K\left(\frac{a}{x}, \frac{b}{x}\right) dx, \end{aligned}$$

where $d\mu$ is the distribution of T and $g_T(x)$ is the density of T-objects with distance x. By section 7.4.2, $g_T(x) = c_T x$, where c_T is a constant depending only on T. For any $\lambda > 0$,

$$\int_0^\infty \frac{1}{x} K\left(\frac{a}{\lambda x}, \frac{b}{\lambda x}\right) dx = \int_0^\infty \frac{1}{x} K\left(\frac{a}{x}, \frac{b}{x}\right) dx \Rightarrow D_{a/\lambda}^{b/\lambda} = D_a^b.$$

If I is scaled by factor λ, then in the scaled image $I^{(\lambda)}$, the derivative at $\mathbf{u}$ equals λ times the derivative at $\lambda^{-1}\mathbf{u}$ in I. Thus the area of regions in $I^{(\lambda)}$ where derivatives are between a and b, denoted $\bar{D}_a^b$, is proportional to $D_{a/\lambda}^{b/\lambda} = D_a^b$. Therefore, for any $a < b$,

$$\frac{\bar{D}_b^a}{Area(I^{(\lambda)})} = \frac{D_b^a}{Area(I)}.$$

Under the assumption of ergodicity of images,

$$\frac{D_b^a}{Area(I)} = \text{marginal propability that } \nabla_u I \in [a, b],$$

and

$$\frac{\bar{D}_b^a}{Area(I^{(\lambda)})} = \text{marginal propability that } \nabla_u I^{(\lambda)} \in [a, b].$$

Then it is seen the marginal distribution of derivatives of $I^{(\lambda)}$ is the same as I.

The expression (7.3) strongly suggests using randomly scaled and translated "template functions" to represent images. These template functions, as called in [4], are random

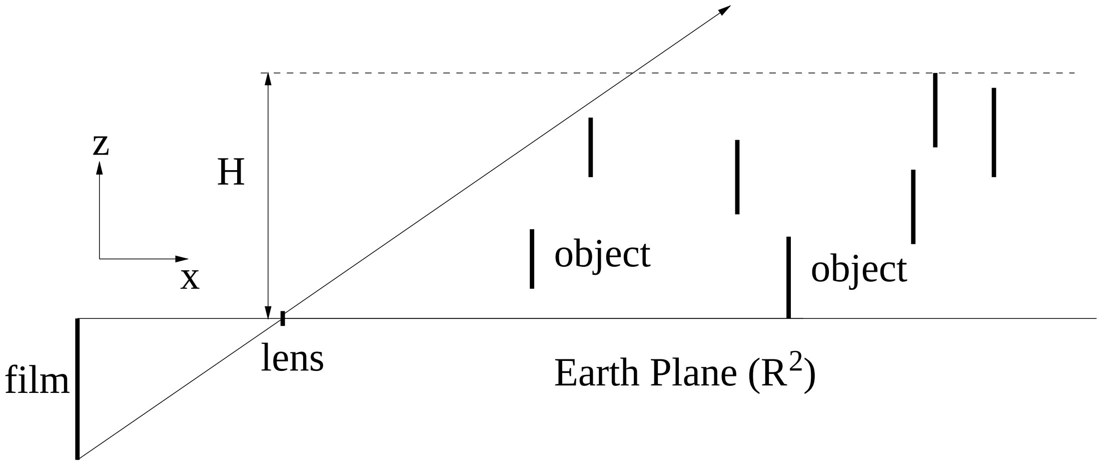


Figure 7.7: Side View of the "World" and the Camera

wavelets. There random wavelets are explained as random patches superimposed on a planar region. Here we find a natural explanation for random wavelets: they are the projections of objects randomly distributed in the 3D world. It is interesting that by modeling images in different ways, the same form of random wavelet representation is obtained. Further study of using random wavelet expansion to construct scale and translation invariant distributions on images is given in next chapter.

7.5 Numerical Experiment

For real images, occlusion can not be ignored. Unfortunately, there are few methods to analyze the effects of occlusion on our model. We resort to numerical experiments to check how well the Poisson model approximates scale invariance.

We simulate putting objects in the spatial region $\mathbf{R} \times \mathbf{R} \times [0, H]$ and projecting them on a finite rectangle camera film. To prevent images from being covered by the projections of only a few objects which are very close to the camera, the simulation only allows objects with distance from the camera larger than a lower bound. An upper bound is also selected for the distance, so that if an object has distance larger than the maximum value, its 2D view is smaller than a pixel. Only objects with distance between the lower and upper bounds are generated.

Figure 7.7 illustrates the side view of the camera as well as the "world" in the simulation. The lens is located on the earth, i.e., with z-coordinate equal to 0. Note that to project objects, which are distributed above the earch, onto the film, the film has to be put "under" the earth, as show in the picture.

The actual implementation does not involve sampling objects in space. When plotting 2D views of objects, we need first know their positions on the image as well as their scaling factors. The positions and scaling factors can be sampled based on the following observation. Let the film be the rectangle $[-1, 1] \times [0, 1]$. Given the distance x of an object, the scaling factor of its 2D view is x and the positions $\{(u_i, v_i)\}$ of all the 2D views with scaling factor x that occur on the image film compose a sample from a Poisson point process with density

λx on the region $[-1, 1] \times [0, H/x]$.

A pseudo-code for sampling positions and scaling factors of 2D views is as follows. Note that the scaling factors are discretized.

```
Position-Scaling
  fix the film as the plane region $[-1, 1] \times [0, 1]$
  fix $D_{\min}$, $D_{\max}$ and $H$
  fix density $\lambda$ and step size $\epsilon$
  $P \leftarrow \emptyset$
  $D \leftarrow D_{\min}$
  while $D \leq D_{\max}$ do
    sample $N$ from the Poisson distribution $\text{Prob}(N = n) = \frac{e^{-D\lambda}}{n!}(D\lambda)^n$
    sample $N$ $(u, v)$'s independently from $[-1, 1] \times [0, H/D]$
    $P \leftarrow P \cup \{(D, u_i, v_i)\}_{i=1}^{N}$
    $D \leftarrow d + \epsilon$
  return $P$
```

Given their positions and scaling factors, the second step is to plot the 2D views of the objects. To simulate occlusion, we start from those with the largest scaling factor, which corresponds to the largest distance from the camera. When two 2D views overlap, the one with larger scaling factor is overwritten by the other one. The input of the following subroutine is $P = \{(D_i, y_i, z_i)\}_{i=1}^{n}$ with $D_1 \geq D_2 \geq \dots D_n > 0$.

```
Draw($P$)
  for $i \leftarrow 1$ to $n$ do
    Supper-Impose($D_i, y_i, z_i$)
```

To display the image, we digitize the film $[-1, 1] \times [0, 1]$ by dividing it into $2N \times N$ squares indexed by (i, j), $i = -N, -N+1, \dots, N-1$, $j = 0, \dots, N-1$. The value at pixel (i, j) is the average intensity value of the image over the square $S_{ij} = [id, (i+1)d] \times [jd, (j+1)d]$, where $d = 1/N$.

The templates we use are rectangles and circles with random sizes. The surface processes are smooth functions plus white noise. Suppose we want to plot the 2D view of a T-object located in the space at $(x, \mathbf{p})$ with surface process $I(x\mathbf{u} - \mathbf{p}; T)$. Then $I(\mathbf{u}; T) = I_s(\mathbf{u}; T) + W(\mathbf{u}; T)$, where I_s is a smooth function and W is a white noise with variance σ^2. Therefore

$$\frac{1}{d^2} \int_{[0,d] \times [0,d]} W(\mathbf{u}) d^2 u$$

is Gaussian random variable with distribution $N(0, \sigma^2/d^2)$. Then for a pixel (i, j) with the square S_{ij} being inside the 2D view of the T-object, its intensity value is

$$\begin{aligned} & \frac{1}{d^2} \int_{S_{ij}} I_s(x\mathbf{u} - \mathbf{p}; T) d^2 u + \frac{1}{d^2} \int_{S_{ij}} W(x\mathbf{u} - \mathbf{p}; T) d^2 u \\ = \quad & \frac{1}{d^2} \int_{S_{ij}} I_s(x\mathbf{u} - \mathbf{p}; T) d^2 u + \frac{1}{x} \xi(i, j), \end{aligned}$$

where $\xi(i, j)$ are i.i.d. $\sim N(0, \sigma^2/d^2)$.

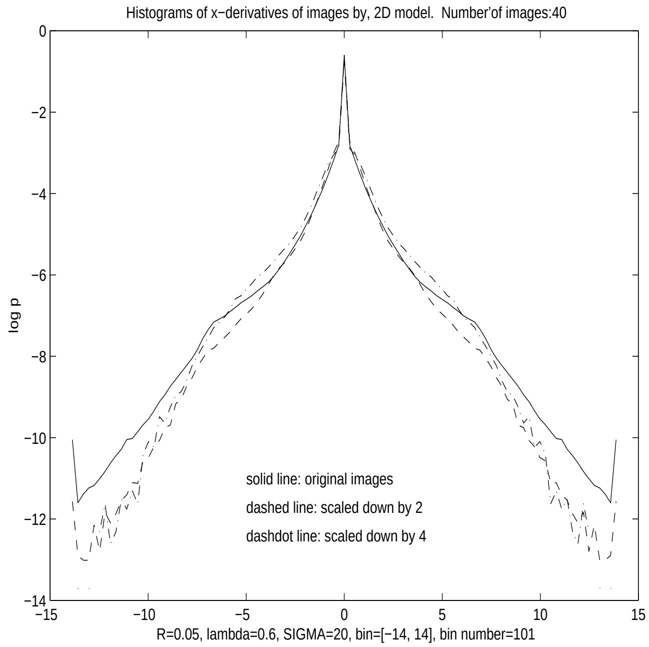


Figure 7.8: Logarithms of the marginal distributions of $\nabla_x I$, $i = 1$ (solid), $i = 2$ (dashed), and $i = 3$ (dash-dotted)

A pseudo-code for the above procedure is as follows. For simplicity, we only show how to plot a scaled disc with a random surface process. In addition, the constant d is assumed to be 1 in the code and therefore $\sigma^2/d^2 = \sigma^2$.

```
Supper-Impose($x, y, z$)
  pick $s$ randomly from {"disc", "rectangle", ...}
  if $s =$"disc"  then
    sample a random radius $r$, a smooth function $I_s$ and a variance $\sigma$ from certain
    distributions
    for $i \leftarrow -N$ to $N - 1$
      for $j \leftarrow 0$ to $N - 1$
        if $|(id, jd) - (y, z)| \leq \frac{r}{x}$  then
          sample a $\xi$ from $N(0, \sigma^2)$
          $I(i, j) \leftarrow \frac{1}{d^2} \int_{S_{ij}} I_s(x\mathbf{u} - \mathbf{p}) d^2 u + \frac{1}{x}\xi$
  else if $s =$"rectangle"  then
    ...
```

Figure 7.8 plots logarithms of marginal distributions of $\nabla_x I$ for images generated by the simulation. It shows good scale invariance. In Figure 7.9, we present a sampled scene. From the picture we see that the "color rendering" makes 2D views of closer objects look like having more details and 2D views of farther objects look smoother.

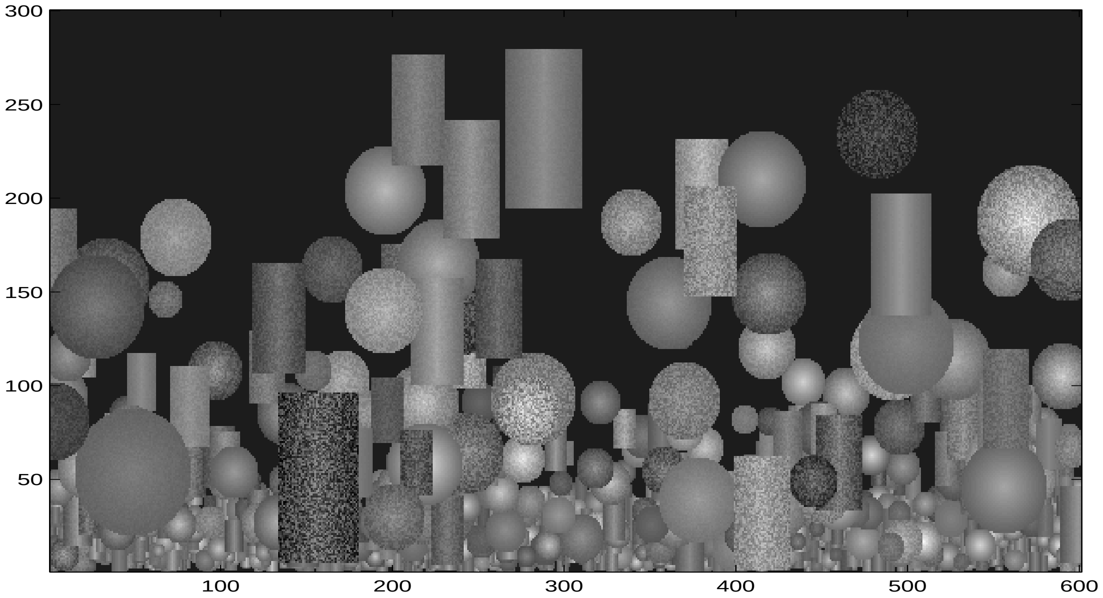


Figure 7.9: A sample scene.

Bibliography

[1] D. C. Knill, D. J. Field, and D. Kersten. Human Discrimination of Fractal Images. *Journal of the Optical Society of America A, Optics and Image Science*, Vol. 7, No. 6. June, 1990.

[2] D. J. Field. What Is the Goal of Sensory Coding? *Neural Computation* **6**, 559-601. 1994.

[3] D. L. Rudderman. Origins of Scaling in Natural Images. *Vision Research.* December, 1996.

[4] D. B. Mumford. Stochastic Models for Generic Images. Division of Applied Mathematics, Brown University. 1998. In preparation.

[5] S. C. Zhu and D. B. Mumford. Prior Learning and Gibbs Reaction-Diffusion. *IEEE Transactions on PAMI.* 1997.

[6] S. Geman. Invariant Binding in Composition Systems. Technical report, Division of Applied Mathematics, Brown University. 1998. In preparation.

Chapter 8

Construction of Scale and Translation Invariant Distributions on Generalized Functions and Functions Defined on Integer Lattice

8.1 Introduction

Scale and translation invariant distribution, also called self-similar stationary distribution, is of great interest in various areas such as random processes (Samorodnitsky & Taqqu [1]), physics (Shlesinger *et al.* [2]) and psychology (Knill *et al.* [3]). Recently it also gains interest in vision (Ruderman [4], Zhu & Mumford [5], Chi & Geman [6], Mumford *et al.* [7]). In order to establish probabilistic models that can help analyzing and understanding different classes of images, it is necessary to study various kinds of scale and translation invariant distributions. Therefore a general approach to construct scale and translation invariant distributions will be very useful.

The scale and translation invariant distributions that we want in vision are different from the commonly defined self-similar stationary distributions. Usually one considers a self-similar stationary distribution as a distribution on a set of functions of time or space. In vision, distributions are defined on the space of images. Mumford [7] noted that a fundamental difference between visual signals and other types of sensory signals is that visual signals do not have characteristic scale, while others have. In other words, images are scale invariant. Based on this distinguished property of images, Mumford for the first time argued that images are better modeled as *generalized functions*, more often called *distributions* in mathematics, instead of functions. Generalized functions are continuous linear functionals on a certain function space. Functions in this function space are called

[1]Supported by the Army Research Office (DAAL03-92-G-0115), the National Science Foundation (DMS-9217655), and the Office of Naval Research (N00014-96-1-0647).

test functions. Mumford went on to formulate in [8] the problem of constructing scale and translation invariant distributions on images in terms of generalized functions on $\mathbf{R}^2$. He showed that, in order to construct scale and translation invariant distributions with finite variance, there should be some constraints on test functions. He then introduced the method of random wavelet expansion and constructed generalized functions which has the following form,

$$I(x,y) = \sum_i J_i(\lambda_i x + a_i, \lambda_i y + b_i), \tag{8.1}$$

where $\{(\lambda_i, a_i, b_i)\}$ is a sample from a Poisson distribution on $\mathbf{R}\times\mathbf{R}^2$ with density $\lambda^{-1}d\lambda da db$ and J_i are independent random generalized functions from some distribution ν_0. It is easy to see that generalized functions given by (8.1) are formally distributed by a scale and translation invariant law.

A similar construction is also established by Chi and Geman in [6], but from a different point of view. They built a model to explain scale and translation invariance in natural images. In this model, a vision signal consists of images of different objects distributed in the nature via a Poisson law. The distribution of color intensities inside the image of an object is called a surface process. Chi and Geman noted that, when viewed from different distances, same object has different surface processes. They then argued that the Poisson law of the distribution of objects and the change of surface process by distance lead to *approximate* scale invariance in natural *scenes*. Here we emphasize the word “scene” because a scene is a function on the plane instead of a generalized function. If occlusion is ignored, then this model gives the same form of expression as (8.1).

This paper generalizes the above results on scale and translation invariant distributions. We will consider the problem of constructing scale and translation invariant distributions on generalized functions defined on an arbitrary Euclidean space. We will establish a general framework under which other kinds of scale and translation invariant distributions as well as the distribution given by (8.1) can be built. In certain sense, this framework generalizes the random wavelet expansion method. However, instead of staying in $\mathbf{R}^2$ and using random wavelets directly as the building blocks of the construction, as seen in (8.1), this method first expands images by wavelets. It then goes to the space of wavelet expansions of images, and constructs a distribution on that space. We will see that, if the distribution on the wavelet expansions of images has certain invariance, then it induces a scale and translation invariant distribution on images. We will give examples to show that this approach enables us to construct scale and translation invariant distributions in a much easier way and build several important scale and translation invariant distributions.

The approach can also be used to construct scale and translation invariant distributions on functions defined on $\mathbf{Z}^d$. The construction is pretty much the same as on generalized functions with some technical modifications involved.

The outline of this paper is as follows. In §8.2, we survey the mathematical formulation of scale and translation invariant distributions of images given by [8] and generalize it to generalized functions on an arbitrary Euclidean space. In §8.3, after showing the motivation from continuous wavelet transforms, we establish a method to construct scale and translation invariant distributions. Then in the next several sections, we construct various scale and translation invariant distributions. In §8.8, we apply the method established in §8.3 to construct scale and translation invariant distributions on functions defined on integer lattice. Finally, in §8.9, we make a discussion on generalizations of the construction.

8.2 Mathematical Set-ups

In this section we give the mathematical notations and definitions that we will use in this paper.

A multiple index α is an n-tuple of integers, i.e., $\alpha = (\alpha_1, \ldots, \alpha_n)$, $\alpha_i \in \mathbf{Z}$. For two multiple indices α and β, $\alpha \geq \beta$ means that $\alpha_i \geq \beta_i$ for each i and $\alpha \geq 0$ means that $\alpha_i \geq 0$ for each i. The absolute value of α, denoted $|\alpha|$, is defined as $\sum_i |\alpha_i|$.

For any $x \in \mathbf{R}^n$, $|x|$ is the length of x, i.e.,

$$|x| = \left(x_1^2 + x_2^2 + \ldots + x_n^2\right)^{\frac{1}{2}}.$$

x^α is defined as $x_1^{\alpha_1} \cdots x_n^{\alpha_n}$. For any function ϕ, $\partial_\alpha \phi$ is defined by

$$\partial_\alpha \phi(x) = \frac{\partial^{\alpha_1 + \ldots + \alpha_n}}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}} \phi(x).$$

The set of all infinitely differentiable functions on $\mathbf{R}^n$ is denoted $C^\infty(\mathbf{R}^n)$. Define the set

$$C_0^\infty(\mathbf{R}^n) = \{\phi \in C^\infty(\mathbf{R}^n) : \text{ supp}(\phi) \text{ is compact }\}.$$

Define the set

$$C^{\infty,0}(\mathbf{R}^n) = \left\{\phi \in C^\infty(\mathbf{R}^n) : \int_{\mathbf{R}^n} \phi(x) dx = 0\right\}.$$

The convergence in $C^\infty(\mathbf{R}^n)$ is defined as follows. $\phi_j \to \phi$, $\phi_j, \phi \in C^\infty(\mathbf{R}^n)$ if for any multiple index α,

$$\max_x |\partial_\alpha \phi_j(x) - \partial_\alpha \phi(x)| \to 0.$$

A rapidly decreasing function ϕ is a function in $C^\infty(\mathbf{R}^n)$, such that for any $k \geq 0$ and $\alpha \geq 0$,

$$\lim_{|x| \to \infty} (|x|^k + 1)|\partial_\alpha \phi(x)| = 0.$$

The set of all real rapidly decreasing functions on $\mathbf{R}^n$ is usually denoted as $\mathcal{D}(\mathbf{R}^n)$ or $\mathcal{D}$. Define convergence in $\mathcal{D}(\mathbf{R}^n)$ as follows. $\phi_j \to \phi$, $\phi_j, \phi \in \mathcal{D}(\mathbf{R}^n)$, if and only if for any $k \geq 0$ and $\alpha \geq 0$,

$$\sup_x (|x|^k + 1)|\partial_\alpha (\phi_j - \phi)(x)| \to 0.$$

An important subspace of $\mathcal{D}$ is $\mathcal{D}_0$, which consists of all functions in $\mathcal{D}$ whose integrals are 0. As an extension, we define $\mathcal{D}_k$ as the set of all functions in $\mathcal{D}$ whose first k moments are vanishing, i.e.,

$$\mathcal{D}_k = \{\phi \in \mathcal{D} : \int x^\alpha \phi(x) dx = 0, |\alpha| \leq k\},$$

It is easy to see that $\mathcal{D}_k$ is closed under the convergence in $\mathcal{D}$.

Given a test function space $\mathcal{E}$, a linear functional F on $\mathcal{E}$ is said to be continuous if for any $\phi_n \to \phi$, ϕ_n, $\phi \in \mathcal{E}$, $\langle F, \phi_n \rangle \to \langle F, \phi \rangle$. The set of all continuous functionals defined on $\mathcal{E}$, which is a linear space, is denoted as $\mathcal{E}'$. Thus, the set of all continuous linear functionals on $\mathcal{D}$ and $\mathcal{D}_k$ are denoted as $\mathcal{D}'$ and $\mathcal{D}'_k$, respectively.

Mathematically, an image is a continuous linear functional whose test functions are defined on $\mathbf{R}^2$. Because of this, we also say an image is defined on $\mathbf{R}^2$. For the sake of generality, from now on we will consider linear functionals defined on an arbitrary Euclidean space $\mathbf{R}^n$.

Given a test function space $\mathcal{E}$, the scaling operator S_t, $t > 0$, on $\mathcal{E}$, is defined as

$$S_t : \quad \phi(x) \mapsto \frac{\phi(t^{-1}x)}{t^n}, \quad \phi \in \mathcal{E}.$$

The translation operator $T_{\vec{v}}$, $\vec{v} \in \mathbf{R}^n$, on $\mathcal{E}$, is defined as

$$T_{\vec{v}} : \quad \phi(x) \mapsto \phi(x - \vec{v}), \quad \phi \in \mathcal{E}.$$

In order that the construction of scale and translation invariant distributions on $\mathcal{E}'$ makes sense, $\mathcal{E}$ has to be closed under scaling and translation, that is, for any $\phi \in \mathcal{E}$, $t > 0$ and $\vec{v} \in \mathbf{R}^n$, $S_t\phi \in \mathcal{E}$ and $T_{\vec{v}}\phi \in \mathcal{E}$. Clearly, $\mathcal{D}_k$ are closed under S_t and $T_{\vec{v}}$.

The scaling operator S_t^* and translation operator $T_{\vec{v}}^*$ on $\mathcal{E}'$ are defined as the adjoint operators of S_t and $T_{\vec{v}}$, respectively. That is, for any $F \in \mathcal{E}'$, and any $\phi \in \mathcal{E}$,

$$\begin{aligned} \langle S_t^* F, \phi \rangle &= \langle F, S_t \phi \rangle, \\ \langle T_{\vec{v}}^* F, \phi \rangle &= \langle F, T_{\vec{v}} \phi \rangle. \end{aligned}$$

Explicitly, if $F \in \mathcal{E}'$ is a function, then $S_t^*(F)(x) = F(tx)$ and $T_{\vec{v}}^*(F)(x) = F(x + \vec{v})$.

We now define scale and translation invariance of a probability distribution on linear functionals.

Definition 8. Suppose $H \in \mathbf{R}$. A probability distribution μ on $\mathcal{E}'$ is called scale and translation invariant with index H if for any $t > 0$ and $\vec{v} \in \mathbf{R}^n$, for any measurable subset $A \subset \mathcal{E}'$,

$$\begin{aligned} \mu(A) &= \mu(\{t^H S_t^* F, \ F \in A\}), \\ \mu(A) &= \mu(\{T_{\vec{v}}^* F, \ F \in A\}). \end{aligned}$$

In the discussion below, we will not use the notation μ explicitly. Instead, we use the notations

$$\begin{cases} F \sim t^H S_t^* F \\ F \sim T_{\vec{v}}^* F \end{cases}$$

to represent the fact that the probability distribution on $\mathcal{E}'$ is scale and translation invariant with index H.

8.3 A General Method to Construct Scale and Translation Invariant Distributions Using Wavelet Expansions

Let us first give the mathematical motivation of our approach. In the theory of continuous wavelet transforms (Heil & Walnut [9]), a representation V of the group $\mathbf{R} \times \mathbf{R}^n$ on $L^2(\mathbf{R}^n)$ is defined by:

$$V(u, \vec{v})f(x) = e^{-nu/2}f(e^{-u}x - v) = D_{e^u}T_{\vec{v}}f(x), \quad f \in L^2(\mathbf{R}^n).$$

In the above expression, D_u, $u > 0$ is called the dilation operator and is defined by

$$D_u f(x) = \frac{f(u^{-1}x)}{u^{n/2}}.$$

D_u is an isometry on $L^2(\mathbf{R}^n)$. If $f, g \in L^2(\mathbf{R}^n)$, then

$$\int_{\mathbf{R}\times\mathbf{R}^n} |(f, V(u, \vec{v})g)|^2 du d\vec{v} = \int_{\mathbf{R}^n} |\hat{f}(\xi)|^2 c_{g,\omega} d\xi, \tag{8.2}$$

where

$$\omega = \frac{\xi}{|\xi|}$$

and

$$c_{g,\omega} = \int_0^\infty \frac{|\hat{g}(t\omega)|^2}{t} dt,$$

and (f, g) is the inner product of the two functions:

$$(f, g) = \int_{\mathbf{R}^n} f(x)\overline{g(x)}dx.$$

For a proof of (8.2), see Appendix.

The function g is called admissible if the integral in (8.2) is convergent for $f = g$. Obviously, if g is admissible, then for almost all ω, $c_{g,\omega} < \infty$. Therefore, $\hat{g}(0) = 0$, which implies that

$$\int_{\mathbf{R}^n} g = 0.$$

For an admissible g, the Φ-transform of g is the operator Φ_g given by

$$\Phi_g f(u, \vec{v}) = (f, V(u, \vec{v})g).$$

In the theory of continuous wavelet transforms, interest is about the case in which for any $\omega \in \mathbf{R}^n$ with $|\omega| = 1$, $c_{g,\omega} = 1$. In this case, Φ_g is an isometry from $L^2(\mathbf{R}^n)$ to $L^2(\mathbf{R} \times \mathbf{R}^n)$ and, in certain sense, for every $f \in L^2(\mathbf{R}^n)$,

$$\begin{aligned} f(x) &= \int_{\mathbf{R}} \int_{\mathbf{R}^n} \Phi_g f(u, \vec{v}) D_{e^u} T_{\vec{v}} g(x) du d\vec{v} \\ &= \int_{\mathbf{R}} \int_{\mathbf{R}^n} (f, V(u, \vec{v})g) V(u, \vec{v}) g(x) du d\vec{v} \end{aligned}$$

In our study, g may not satisfy $c_{g,\omega} = 1$ for all ω with $|\omega| = 1$. Instead, we assume that g satisfies $0 < \inf_\omega c_{g,\omega} \leq \sup_\omega c_{g,\omega} < \infty$. From (8.2), we see that if we define a continuous linear operator A from $L^2(\mathbf{R}^n)$ to $L^2(\mathbf{R}^n)$ such that

$$(Af)^\wedge(\xi) = \frac{\hat{f}(\xi)}{\sqrt{c_{g,\omega}}},$$

then

$$\|\Phi_g Af\|_{L^2} = \int_{\mathbf{R}\times\mathbf{R}^n} |(Af, V(u,\vec{v})g)|^2 du d\vec{v} = \int_{\mathbf{R}^n} |\hat{f}(\xi)|^2 d\xi = \|f\|_{L^2},$$

Therefore $\Phi_g A$ is an isometry from $L^2(\mathbf{R}^n)$ to $L^2(\mathbf{R}\times\mathbf{R}^n)$. Note that since $(Af, g) = (f, Ag)$, A is a self-adjoint operator.

In certain sense, for every $f \in L^2(\mathbf{R}^n)$,

$$f(x) = \int_{\mathbf{R}} \int_{\mathbf{R}^n} \Phi_g Af(u,\vec{v}) A D_{e^u} T_{\vec{v}} g(x) du d\vec{v} \tag{8.3}$$

For an explanation of (8.3), see Appendix.

Equation (8.3) tells us that a function $f \in L^2(\mathbf{R}^n)$ can always be expanded in both scale u and space coordinates $\vec{v}$. It is natural to generalize and guess that for "most" of the linear functionals F in $\mathcal{E}'$, the representation (8.3) is also valid, i.e.,

$$F(x) = \int_{\mathbf{R}} \int_{\mathbf{R}^n} \Phi_g AF(u,\vec{v}) A D_{e^u} T_{\vec{v}} g(x) du d\vec{v}$$

Then for any $\phi \in \mathcal{E}$, the action of F on ϕ can be written, formally, as

$$\begin{aligned} \langle F, \phi \rangle &= \int_{\mathbf{R}} \int_{\mathbf{R}^n} \Phi_g AF(u,\vec{v}) (A D_{e^u} T_{\vec{v}} g, \phi) du d\vec{v} \\ &= \int_{\mathbf{R}} \int_{\mathbf{R}^n} \Phi_g AF(u,\vec{v}) (D_{e^u} T_{\vec{v}} g, A\phi) du d\vec{v}. \end{aligned}$$

By variable substitutions $u \to -u$, $\vec{v} \to -\vec{v}$, and grouping terms, we can rewrite $\langle F, \phi \rangle$ in the following form

$$\langle F, \phi \rangle = \int_{\mathbf{R}} \int_{\mathbf{R}^n} K(u,\vec{v}) (e^{Hu} S^*_{e^u} T^*_{\vec{v}} g, A\phi) du d\vec{v}.$$

We explain this equation as follows. Corresponding to g, there is a mapping Ψ_g from $\mathcal{E}$ to a space of functions, say $\mathcal{F}$, on $\mathbf{R} \times \mathbf{R}^n$ such that, for any $\phi \in \mathcal{E}$, $\Psi_g \phi \in \mathcal{F}$, and for any $u \in \mathbf{R}$ and $\vec{v} \in \mathbf{R}^n$,

$$\Psi_g \phi(u,\vec{v}) = (e^{Hu} S^*_{e^u} T^*_{\vec{v}} g, \phi). \tag{8.4}$$

Note that, then

$$\Psi_g \phi(u,\vec{v}) = (e^{Hu} g(e^u x + \vec{v}), \phi(x)) = e^{(H-n/2)u} (V(-u,-v)g, \phi).$$

Taking $K(u,\vec{v})$ as a linear functional on $range(\Psi_g)$, we get

$$\langle F, \phi \rangle = \langle K, \Psi_g A\phi \rangle$$

Under certain conditions, there is a continuous operator χ, such that

$$\Psi_g A\phi = \chi \Psi_g \phi.$$

Then, letting $W = \chi^* K$, we get the form

$$\langle F, \phi\rangle = \langle W, \Psi_g \phi\rangle$$

i.e.

$$F = \Psi_g^* W, \tag{8.5}$$

where Ψ_g^* is the adjoint operator of Ψ_g.

Let us compute how $\langle F, \phi\rangle$ changes when ϕ undergoes a scaling and translation transformation. For any u_0 and $\vec{v}_0$,

$$\begin{aligned}\Psi_g T_{\vec{v}_0} S_{e^{u_0}} \phi(u, \vec{v}) \quad &= (e^{Hu} S^*_{e^u} T^*_{\vec{v}} g, T_{\vec{v}_0} S_{e^{u_0}} \phi) \\ &= (e^{Hu} S^*_{e^{u_0}} T^*_{\vec{v}_0} S^*_{e^u} T^*_{\vec{v}} g, \phi) \\ &= (e^{Hu} S^*_{e^{u+u_0}} T^*_{\vec{v}+e^u \vec{v}_0} g, \phi) \\ &= e^{-Hu_0} \Psi_g \phi(u + u_0, \vec{v} + e^u \vec{v}_0).\end{aligned}$$

Define an operator $U_{u_0,\vec{v}_0}$ on functions on $\mathbf{R} \times \mathbf{R}^n$ via

$$U_{u_0,\vec{v}_0} f(u, \vec{v}) = f(u + u_0, \vec{v} + e^u \vec{v}_0). \tag{8.6}$$

Then

$$\Psi_g T_{\vec{v}_0} S_{e^{u_0}} \phi(u, \vec{v}) = e^{-Hu_0} U_{u_0,\vec{v}_0} \Psi_g \phi(u, \vec{v}).$$

Therefore

$$\langle S^*_{e^{u_0}} T^*_{\vec{v}_0} F, \phi\rangle = \langle F, T_{\vec{v}_0} S_{e^{u_0}} \phi\rangle = e^{-Hu_0} \langle W, U_{u_0,\vec{v}_0} \Psi_g \phi\rangle = e^{-Hu_0} \langle U^*_{u_0,\vec{v}_0} W, \Psi_g \phi\rangle, \tag{8.7}$$

where $U^*_{u_0,\vec{v}_0}$ is the adjoint of $U_{u_0,\vec{v}_0}$. Explicitly, if W is a function, then $U^*_{u_0,\vec{v}_0}$ transforms W to

$$U^*_{u_0,\vec{v}_0} W(u, \vec{v}) = W(u - u_0, \vec{v} - e^{u-u_0} \vec{v}_0). \tag{8.8}$$

Now suppose that F in (8.5) is scale and translation invariant with index H, then for any $u_0 \in \mathbf{R}$ and $\vec{v}_0 \in \mathbf{R}^n$,

$$e^{Hu_0} \langle S^*_{e^{u_0}} T^*_{\vec{v}_0} F, \phi\rangle \sim \langle F, \phi\rangle.$$

Comparing with (8.7), we get

$$\langle U^*_{u_0,\vec{v}_0} W, \Psi_g \phi\rangle \sim \langle W, \Psi_g \phi\rangle.$$

This implies that in order that F in (8.5) is distributed by a scale and translation invariant law, we only need to make sure that W are distributed by a law which is invariant under $U^*_{u,\vec{v}}$, for any $(u, \vec{v}) \in \mathbf{R} \times \mathbf{R}^n$.

We put the above result in a slightly different form.

Proposition 20. Suppose $g \in \mathcal{D}$ satisfies

$$\int_{\mathbf{R}^n} g = 0.$$

Define the operator Ψ_g by (8.4). Suppose $\mathcal{E}$ is a space of functions on $\mathbf{R}^n$ which is closed under scaling and translation, and $\mathcal{F}$ is a space of functions on $\mathbf{R} \times \mathbf{R}^n$. If Ψ_g maps $\mathcal{E}$ into $\mathcal{F}$ and μ is a probability distribution on $\mathcal{F}'$, then Ψ_g^* induces a distribution ν on $\mathcal{E}'$ by

$$\nu(A) = \mu((\Psi_g^*)^{-1}A), \quad \text{for any } \ A \subset \mathcal{E}' \ \text{ measurable.}$$

If μ is invariant under $U_{u,\vec{v}}^*$, for any $(u, \vec{v}) \in \mathbf{R} \times \mathbf{R}^n$, then the distribution ν is scale and translation invariant with index H.

Remark 7. It is easy to see that the $U_{u,\vec{v}}^*$ in (8.6) is a diffeomorphism with Jacobi 1. This observation is the key to our constructions. For example, since homogeneous Poisson distribution is invariant under $U_{u,\vec{v}}^*$, we can construct Poisson type scale and translation invariant distributions. Since inner product of $L^2(\mathbf{R} \times \mathbf{R}^n)$ is invariant under $U_{u,\vec{v}}^*$, we can construct Gaussian type scale and translation invariant distributions. We will make the arguments clearer in following sections.

In the next several sections, we will use Proposition Proposition 20 to construct several types of scale and translation invariant distributions.

8.4 Poisson Type Scale and Translation Invariant Distributions

Without giving the exact definition, we mention that Poisson type distributions are defined on $\mathcal{F}'$, where $\mathcal{F} = C^\infty(\mathbf{R} \times \mathbf{R}^n) \cap L^1(\mathbf{R} \times \mathbf{R}^n)$. In order that the construction given by Proposition Proposition 20 makes sense, $\mathcal{E}$ should satisfy the condition that $\Psi_g(\mathcal{E}) \subset \mathcal{F}$. The following result shows that there are $\mathcal{E}$ which satisfy this condition.

Proposition 21. Suppose $g \in C_0^\infty(\mathbf{R}^n)$.

(1) If g has vanishing first $k-1$ moments and $0 < H < k$ in the definition of Ψ_g, then Ψ_g is a continuous mapping from $\mathcal{D}(\mathbf{R}^n)$ into $C^{\infty,0}(\mathbf{R} \times \mathbf{R}^n)$ and into $L^p(\mathbf{R} \times \mathbf{R}^n)$, for any $1 \le p \le \infty$;

(2) If g has vanishing integral and $-k < H \le 0$ in the definition of Ψ_g, then Ψ_g is a continuous mapping from $\mathcal{D}_{k-1}(\mathbf{R}^n)$ into $C^{\infty,0}(\mathbf{R} \times \mathbf{R}^n)$ and into $L^p(\mathbf{R} \times \mathbf{R}^n)$, for any $1 \le p \le \infty$.

For a proof of Proposition Proposition 21, see the Appendix.

For the sake of simplicity, from now on we only consider the case $k = 1$. We also assume that any Poisson point process involved in the discussion in the remaining part of the paper has density 1.

The following Proposition Proposition 22 and Proposition Proposition 23 construct scale and translation invariant distributions directly without using characteristic functionals.

Proposition 22. (Poisson Type 1) Let $\mathcal{W}(u, \vec{v})$ be a function such that

$$\int_{\mathbf{R}\times\mathbf{R}^n} \frac{|\mathcal{W}(u, \vec{v})|}{e^{nu}} dud\vec{v} < \infty.$$

and

$$\int_{\mathbf{R}\times\mathbf{R}^n} |\mathcal{W}(u, \vec{v})| dud\vec{v} < \infty.$$

For each $(x, \vec{y}) \in \mathbf{R} \times \mathbf{R}^n$, define a function

$$w_{x,\vec{y}}(u, \vec{v}) = \mathcal{W}(u - x, \vec{v} - e^{u-x}\vec{y}).$$

Let $\{(u_i, \vec{v}_i)\}$ be a homogeneous Poisson process in $\mathbf{R} \times \mathbf{R}^n$. For each sample $\{(u_i, \vec{v}_i)\}$, define W on $C^\infty(\mathbf{R} \times \mathbf{R}^n) \cap L^1(\mathbf{R} \times \mathbf{R}^n)$, such that for $f \in C^\infty(\mathbf{R} \times \mathbf{R}^n) \cap L^1(\mathbf{R} \times \mathbf{R}^n)$,

$$\langle W, f \rangle = \sum_i (w_{u_i,\vec{v}_i}, f). \tag{8.9}$$

Then for almost all samples $\{(u_i, \vec{v}_i)\}$ of the Poisson process, the linear functional F given by (8.5) is well defined and continuous on (1) $\mathcal{D}(\mathbf{R}^n)$, if $0 < H < 1$ and (2) $\mathcal{D}_0(\mathbf{R}^n)$, if $-1 < H \leq 0$. Moreover, F is distributed by a scale and translation invariant law and has finite covariance.

Proof. By (8.8), for each $(u_i, \vec{v}_i)$,

$$\begin{aligned} & U^*_{u_0,\vec{v}_0} w_{u_i,\vec{v}_i}(u, \vec{v}) \\ = \; & w_{u_i,\vec{v}_i}(u - u_0, \vec{v} - e^{u-u_0}\vec{v}_0) \\ = \; & \mathcal{W}(u - u_0 - u_i, \vec{v} - e^{u-u_0}\vec{v}_0 - e^{u-u_0-u_i}\vec{v}_i) \\ = \; & w_{u_0+u_i,\vec{v}_i+e^{u_i}\vec{v}_0}(u, \vec{v}). \end{aligned}$$

Hence

$$U^*_{u_0,\vec{v}_0} W = \sum_i w_{u_i+u_0,\vec{v}_i+e^{u_i}\vec{v}_0}(u, v).$$

Since $\{(u_i, \vec{v}_i)\}$ is a homogeneous Poisson process on $\mathbf{R} \times \mathbf{R}^n$, and the determinant of the transform

$$\begin{aligned} T: \quad & u \to u + u_0 \\ & \vec{v} \to \vec{v} + e^u\vec{v}_0 \end{aligned}$$

is one, then $\{(u_i + u_0, \vec{v}_i + e^{u_i}\vec{v}_0 + \vec{v}_i)\} \sim \{(u_i, \vec{v}_i)\}$. Since W is determined by $\{(u_i, \vec{v}_i)\}$, this proves that the distribution of W is invariant under $U^*_{u_0,\vec{v}_0}$. Therefore by Proposition Proposition 20, $F = \Psi^*_g W$ is distributed by a scale and translation invariant law.

The proof for the other claims in Proposition Proposition 22 is lengthy, so we put it in the Appendix. □

Proposition 23. (Poisson Type 2) Let $\{(u_i, \vec{v}_i)\}$ be a homogeneous Poisson process in $\mathbf{R} \times \mathbf{R}^n$. Define W on $C^\infty(\mathbf{R} \times \mathbf{R}^n) \cap L^1(\mathbf{R} \times \mathbf{R}^n)$ as

$$W = \sum_i \delta(u - u_i)\delta(\vec{v} - \vec{v}_i).$$

i.e., for $f \in C^\infty(\mathbf{R} \times \mathbf{R}^n) \cap L^1(\mathbf{R} \times \mathbf{R}^n)$,

$$\langle W, f \rangle = \sum_i f(u_i, \vec{v}_i), \tag{8.10}$$

Then for almost all samples $\{(u_i, \vec{v}_i)\}$ of the Poisson process, the linear functional $F = \Psi_g^* W$ is well defined and continuous on (1) $\mathcal{D}(\mathbf{R}^n)$, if $0 < H < 1$ and (2) $\mathcal{D}_0(\mathbf{R}^n)$, if $-1 < H \leq 0$. Moreover, F is distributed by a scale and translation invariant law and has finite covariance.

Proof. The proof that F is distributed by a scale and translation invariant law is similar to the proof of Proposition Proposition 22. The proof for the other parts is given in the Appendix. □

We can generalize the construction of (8.10) as follows. Fix a probability distribution ν on $\mathbf{R}$. As in Proposition Proposition 23, let $\{(u_i, \vec{v}_i)\}$ be a homogeneous Poisson distribution on $\mathbf{R} \times \mathbf{R}^n$ with density 1. Define W on $C^\infty(\mathbf{R} \times \mathbf{R}^n) \cap L^p(\mathbf{R} \times \mathbf{R}^n)$, such that for $f \in C^\infty(\mathbf{R} \times \mathbf{R}^n) \cap L^p(\mathbf{R} \times \mathbf{R}^n)$,

$$\langle W, f \rangle = \sum_i a_i f(u_i, \vec{v}_i), \tag{8.11}$$

where a_i are i.i.d.$\sim \nu$.

Lemma 10. (Poisson Type 3) If the characteristic function of ν is

$$\kappa(s) = \int_{\mathbf{R}} e^{isx} \nu(dx) = e^{-|s|^\alpha},$$

where $1 \leq \alpha \leq 2$, i.e., ν is the Lévy distribution with index α, then with probability one, the sum in (8.11) converges. Moreover, W is continuous *in probability*, i.e., as $f_n \to f$ in $C^\infty(\mathbf{R} \times \mathbf{R}^n) \cap L^\alpha(\mathbf{R} \times \mathbf{R}^n)$, $\langle W, f_n \rangle \to \langle W, f \rangle$ in probability.

Proof. We given an intuitive argument for this result. Rewrite (8.11) as

$$\langle W, f \rangle = \sum_{(u,\vec{v}) \in \mathbf{R} \times \mathbf{R}^n} a(u, v) f(u, v) 1_{\{(u_i, \vec{v}_i)\}}(u, \vec{v}),$$

where $a(u, v)$ are i.i.d. $\sim \nu$. Each term in the summation is independent from the others. Therefore, in order to show that the sum converges with probability one, it is enough to show that the product of the characteristic functions of all these terms converges. In a small box with size $dud\vec{v}$ around $(u, \vec{v})$, the probability that an $(u_i, \vec{v}_i)$ appears is $dud\vec{v}$. Conditioning on the appearance of an $(u_i, \vec{v}_i)$ in the small box, the characteristic function of $a(u_i, \vec{v}_i) f(u_i, \vec{v}_i)$ is $\kappa(sf(u_i, \vec{v}_i))$. The probability that no $(u_i, \vec{v}_i)$ appears in the box is $1 - dud\vec{v}$. Hence the characteristic function of the total sum is

$$E\left(e^{is\langle W, f \rangle}\right) = \prod_{(u,\vec{v})} \left(1 + (\kappa(sf(u, \vec{v})) - 1) dud\vec{v}\right)$$

$$
\begin{aligned}
&= \exp\left(\sum_{(u,\vec{v})} \log(1+(\kappa(sf(u,\vec{v}))-1))dud\vec{v}\right) \\
&= \exp\left(\int_{\mathbf{R}\times\mathbf{R}^n} (\kappa(sf(u,\vec{v}))-1)dud\vec{v}\right) \\
&= \exp\left(\int_{\mathbf{R}\times\mathbf{R}^n} (e^{-|sf(u,\vec{v})|^\alpha}-1)dud\vec{v}\right) \qquad (8.12)
\end{aligned}
$$

Since

$$0 \le 1 - e^{-|sf(u,\vec{v})|^\alpha} \le |sf(u,\vec{v})|^\alpha,$$

$|f(u,\vec{v})|^\alpha$ is integrable. Therefore, the above product converges. Hence the first part of Proposition Lemma 10 is proved.

When $f_n \to f$ in $C^\infty(\mathbf{R}\times\mathbf{R}^n)\cap L^1(\mathbf{R}\times\mathbf{R}^n)$, then the characteristic function of $\langle W, f_n - f\rangle$ is:

$$\exp\left(\int_{\mathbf{R}\times\mathbf{R}^n} (e^{-|s\left(f_n(u,\vec{v})-f(u,\vec{v})\right)|^\alpha} - 1)dud\vec{v}\right) \xrightarrow{\text{u.c.}} 1,$$

where u.c. means uniformly convergence on any finite closed interval. Hence $\langle W, f_n - f\rangle \to 0$ in probability. □

By Proposition Proposition 21, Ψ_g is a continuous mapping into $C^\infty(\mathbf{R}\times\mathbf{R}^n)\cap L^\alpha(\mathbf{R}\times\mathbf{R}^n)$. Therefore, W being continuous in probability implies that $F = \Psi_g^* W$ is also continuous in probability. Unfortunately, it is in general not true that with probability one, F defined in this way is well defined for all functions in $\mathcal{D}(\mathbf{R}^n)$ or $\mathcal{D}_0(\mathbf{R}^n)$.

To get around this problem, we observe that $\mathcal{D}(\mathbf{R}^n)$ and $\mathcal{D}_k(\mathbf{R}^n)$ are nuclear spaces. Therefore we can use the following Minlos' theorem:

Theorem Let C be a function on a nuclear space $\mathcal{E}$ and have the following properties:

1. C is continuous on $\mathcal{E}$;
2. C is positive definite; i.e., for any $\phi_1,\ldots,\phi_n \in \mathcal{E}$ and any $\xi_1,\ldots,\xi_n \in \mathbf{C}$,
$$\sum_{i,j} \xi_i\bar{\xi}_j C(\phi_i - \phi_j) \ge 0;$$
3. $C(0) = 1$;

Then there exists a unique probability distribution μ_C on $\mathcal{E}'$, so that for all $\phi \in \mathcal{E}$,

$$\int_{\mathcal{E}'} \exp(i\langle F,\phi\rangle)d\mu_C(F) = C(\phi).$$

We check condition 1 of the theorem. The characteristic functional given by (8.12) is defined for all $f \in L^\alpha(\mathbf{R}\times\mathbf{R}^n)$. Therefore, by Proposition Proposition 21 the functional

$$C(\phi) = E\left(e^{i\langle W,\Psi_g\phi\rangle}\right) = \exp\left(\int_{\mathbf{R}\times\mathbf{R}^n}\left(e^{-|\Psi_g\phi(u,\vec{v})|^\alpha} - 1\right)dud\vec{v}\right) \qquad (8.13)$$

is defined for all $\phi \in \mathcal{D}(\mathbf{R}^n)$ when $0 < H < 1$ and for all $\phi \in \mathcal{D}_0(\mathbf{R}^n)$ when $-1 < H \leq 0$. When $0 < H < 1$, Ψ_g is a continuous mapping from $\mathcal{D}(\mathbf{R}^n)$ to $L^\alpha(\mathbf{R} \times \mathbf{R}^n)$. Since

$$0 \leq 1 - e^{-|\Psi_g \phi(u,\vec{v})|^\alpha} \leq |\Psi_g \phi(u,\vec{v})|^\alpha,$$

then by dominant convergence theorem, $C(\phi)$ is continuous on $\mathcal{D}(\mathbf{R}^n)$.

Clearly, $C(0) = 1$. Therefore condition 3 is satisfied by $C(\phi)$. It remains to check condition 2. By (8.13),

$$\sum_{j,k} \xi_j \bar{\xi}_k C(\phi_j - \phi_k) = \sum_{j,k} \xi_j \bar{\xi}_k E\left(e^{i\langle W, \Psi_g(\phi_j - \phi_k)\rangle} \right).$$

Because $\langle W, \Psi_g \phi \rangle$ is real, the right hand side of the equation equals $E\left(|\sum_i \xi_i e^{i\langle W, \Psi_g \phi_i \rangle}|^2 \right)$ which is non-negative. Therefore $C(\phi)$ also satisfies condition 2.

To see why the distribution is scale and translation invariant, note that the Jacobi of the transform

$$\begin{aligned} T: \quad & u \to u + u_0 \\ & \vec{v} \to \vec{v} + e^u \vec{v}_0 \end{aligned}$$

is one. Then

$$\begin{aligned} C(U_{u_0,\vec{v}_0}\phi) &= \exp\left(\int_{\mathbf{R}\times\mathbf{R}^n} \left(e^{-|\Psi_g \phi(u+u_0, \vec{v}+e^u \vec{v}_0)|^\alpha} - 1 \right) du d\vec{v} \right) \\ &= \exp\left(\int_{\mathbf{R}\times\mathbf{R}^n} \left(e^{-|\Psi_g \phi(u,\vec{v})|^\alpha} - 1 \right) du d\vec{v} \right) \\ &= C(\phi). \end{aligned}$$

The case $\phi \in \mathcal{D}_0(\mathbf{R}^n)$, $-1 < H \leq 0$ can be similarly treated.

8.5 Lévy Type Scale and Translation Invariant Distributions

The construction given by (8.10) is an example of using Poisson point process to generate stationary stable processes, which is a well-known method in the study of Lévy processes. This suggests constructions using Lévy processes. See [2] for a reference to this topic.

For $\alpha = H + 1 \geq 1$, $1 \leq \alpha < 2$, there are the well-known Cauchy noises on $\mathcal{D}(\mathbf{R}^n)$ which are given by the following characteristic functionals

$$C(\phi) = \exp\left(-\int_{\mathbf{R}^n} |\phi|^\alpha \right).$$

From the above formula, it is seen that Cauchy noises are scale and translation invariant. When $\alpha = 1$ and $H = 0$, the corresponding Cauchy noise is a "true" scale and translation invariant distribution in that $F \sim S^*_{e^u} F$, $\forall u$, without having to put a factor in front of $S^*_{e^u} F$. This distribution has infinite covariance. Indeed, we have the following well-known result.

Proposition 24. If a probability distribution μ on $\mathcal{D}'$ is scale and translation invariant with index $H = 0$ and has finite covariance, then with probability one, any F from μ is a random constant.

To make the paper more self-contained, we put a proof of Proposition Proposition 24 in Appendix.

We will construct scale and translation invariant distributions with $\alpha \neq H + 1$ in the remaining part of this section. Let $\mathcal{F} = C^\infty(\mathbf{R} \times \mathbf{R}^n) \cap L^\alpha(\mathbf{R} \times \mathbf{R}^n)$. As the construction at the end of §8.4, we will first make an intuitive argument to construct a Lévy type distribution on $\mathcal{F}'$ and then resort to Minlos' theorem to verify the construction.

Let ν on $\mathbf{R} \setminus \{0\}$ be a Lévy measure

$$\nu(dy) = \frac{dy}{|y|^{1+\alpha}} 1_{\mathbf{R}\setminus\{0\}},$$

where $1 \leq \alpha < 2$. Then

$$\int_{\mathbf{R}} \left(e^{isx} - 1 \right) \nu(dx) = -|s|^\alpha.$$

Fix a function $f \in \mathcal{F}$. In the space $\mathbf{R} \times \mathbf{R}^n \times (\mathbf{R} \setminus \{0\})$, define a Poisson process with density $dud\vec{v}d\nu$. If $(u, \vec{v}, y)$ occurs in the process, then we say a jump with size y happens at $(u, \vec{v})$. Given $(u, \vec{v}) \in \mathbf{R} \times \mathbf{R}^n$, the weighted total jump at $(u, \vec{v})$ is the sum of all the jumps happened at $(u, \vec{v})$ multiplied by $f(u, \vec{v})$. Since the probability that a jump with size y happens at $(u, \vec{v})$ is $dud\vec{v}d\nu(y)$ and each jump at $(u, \vec{v})$ is independent from each other, it is seen that the characteristic function of the weighted total jump at $(u, \vec{v})$ is then

$$\begin{aligned} & \prod_y \left(1 - dud\vec{v}d\nu(y) + e^{isf(u,\vec{v})y} dud\vec{v}d\nu(y) \right) \\ = \; & \exp\left(\sum_y \log(1 + (e^{isf(u,\vec{v})y} - 1)dud\vec{v}d\nu(y)) \right) \\ = \; & \exp\left(dud\vec{v} \int_{\mathbf{R}\setminus\{0\}} (e^{isf(u,\vec{v})y} - 1)d\nu(y) \right) \\ = \; & \exp\left(- |sf(u, \vec{v})|^\alpha dud\vec{v} \right). \end{aligned}$$

Let the value of the functional W at f be the sum of the weighted jumps at all $(u, \vec{v})$. Then $\langle W, f \rangle$ has the characteristic function

$$\prod_{(u,\vec{v})} \exp\left(- |sf(u, \vec{v})|^\alpha dud\vec{v} \right) = \exp\left(- \int_{\mathbf{R}\times\mathbf{R}^n} |sf(u, \vec{v})|^\alpha dud\vec{v} \right), \tag{8.14}$$

which is convergent. So the sum of weighted jumps converges with probability one.

Having got the expression (8.14), it is not hard to check that

$$C(\phi) = \exp\left(- \int_{\mathbf{R}\times\mathbf{R}^n} |\Psi_g \phi(u, \vec{v})|^\alpha dud\vec{v} \right)$$

is a characteristic functional. Therefore, by Minlos' theorem, there is a unique (scale and translation) invariant distribution with characteristic function C.

8.6 Gaussian Type Scale and Translation Invariant Distributions

The distribution defined via the characteristic functional

$$C(\phi) = \exp\Big(-\frac{1}{2}(\Psi_g\phi, \Psi_g\phi)\Big)$$

is a Gaussian distribution on $\mathcal{D}_0(\mathbf{R}^n)$, where the inner product $(\cdot,\cdot)$ on $L^2(\mathbf{R}\times\mathbf{R}^n)$ is defined as

$$(f,h) = \int_{\mathbf{R}\times\mathbf{R}^n} f(u,\vec{v})\overline{h(u,\vec{v})}dud\vec{v}. \tag{8.15}$$

The transform $U^*_{u_0,\vec{v}_0}$, for any $u_0 \in \mathbf{R}$ and $\vec{v}_0 \in \mathbf{R}^n$, is unitary under the above defined inner product. Therefore, the Gaussian distribution defined by C is scale and translation invariant.

The distribution given by (8.15) can be generalized. For example, instead of considering $L^2(\mathbf{R}\times\mathbf{R}^n)$, we consider the Hilbert space

$$H = \{f(u,\vec{v}) : f(u,\vec{v}) \in L^2(\mathbf{R}\times\mathbf{R}^n),\ \partial_{\vec{v}}f(u,\vec{v}) \in L^2(\mathbf{R}\times\mathbf{R}^n)\}.$$

Define

$$(f,h) = \int_{\mathbf{R}\times\mathbf{R}^n}\Big(f(u,\vec{v})\overline{h(u,\vec{v})} + M\partial_{\vec{v}}f(u,\vec{v})\cdot\overline{\partial_{\vec{v}}h(u,\vec{v})}\Big)dud\vec{v}, \tag{8.16}$$

where $M \geq 0$ is a constant. Then $U^*_{u_0,\vec{v}_0}$ is again unitary. This can be checked by showing

$$\partial_{\vec{v}}(U^*_{u_0,\vec{v}_0}\phi) = U^*_{u_0,\vec{v}_0}(\partial_{\vec{v}}\phi).$$

However, this apparent generalization does not shed much new light on the construction. Indeed, suppose F_0 is defined by

$$\langle F_0,\phi\rangle = \langle W_0, \Psi_g\phi\rangle,$$

and F_i, $1 \leq i \leq n$ are defined by

$$\langle F_i,\phi\rangle = \langle W_i, \sqrt{M}\Psi_{g_i}\phi\rangle,$$

where

$$g_i = \frac{\partial g}{\partial v_i},$$

and where W_0, W_1, $\ldots W_n$ are i.i.d. $\sim$ the law defined by (8.15). Then $F_0 + F_1 + \ldots + F_n$ is distributed by the law induced by Ψ_g and the measure given by (8.16).

8.7 A Scale and Translation Invariant Distribution that is not Infinitely Divisible

The scale and translation invariant distributions we have constructed so far are infinitely divisible. In this section we will show that the class of scale and translation invariant distributions is very rich by constructing a scale and translation invariant distribution which is not infinitely divisible.

The construction is an application of a scheme which we will present in the next paragraph. We will present the scheme in a way which is far from being mathematically rigorous. We adopt this less rigorous approach because we want the scheme to serve as a guideline to construct various scale and translation distributions rather than to be a rigorous mathematical result.

Let $\mathbf{X}_\lambda(\vec{v})$, $\lambda \in \Lambda$ be a family of stationary processes on $\mathbf{R}^n$ which take real values. That is, for each $\lambda \in \Lambda$, for each sample $X_\lambda(\vec{v})$ from $\mathbf{X}_\lambda(\vec{v})$, and for each $\vec{v} \in \mathbf{R}^n$, $X_\lambda(\vec{v}) \in \mathbf{R}$. Let $\mathbf{Y}(u)$ be a stationary process on $\mathbf{R}$ which takes values in Λ, i.e., for each sample $Y(u)$ from $\mathbf{Y}(u)$ and for each $u \in \mathbf{R}$, $Y(u) \in \Lambda$. Construct a process $\mathbf{Z}(u, \vec{v})$ as follows. First, choose a sample $Y(u)$ from $\mathbf{Y}(u)$. Then at each u, independently choose a sample $X_\lambda(\vec{v})$ from the process $\mathbf{X}_\lambda(\vec{v})$, where $\lambda = Y(u)$. Let $Z(u, \vec{v}) = X_\lambda(\vec{v}) = X_{Y(u)}(\vec{v})$. We then have a functional W on $\mathcal{F}$ (see Proposition Proposition 20), such that for any $f \in \mathcal{F}$,

$$W : f \mapsto \int_{\mathbf{R}\times\mathbf{R}^n} Z(u, \vec{v}) f(u, \vec{v}) du d\vec{v}. \tag{8.17}$$

We then define F on $\mathcal{E}$ by $\langle F, \phi\rangle = \langle W, \Psi_g \phi\rangle$.

To see why the distribution of F is scale and translation invariant, we only need to check the distribution of W is invariant under $U^*_{u_0,\vec{v}_0}$. The $U^*_{u_0,\vec{v}_0}$ can be decomposed into two transforms (1) $u \mapsto u - u_0$ and (2) $\vec{v} \mapsto \vec{v} - e^{u-u_0} v_0$. Under transform (1), since $\mathbf{Y}(u)$ is stationary, and for each u, the sample $X_{Y(u)}(\vec{v})$ is selected independently, the distribution of $\mathbf{Z}$ is invariant. Under transform (2), since for each u, the sample $X_{Y(u)}(\vec{v})$ is selected independently to other u's, and each $\mathbf{X}_{Y(u)}(\vec{v})$, given u, is stationary, then $\mathbf{Z}$ is again invariant. Therefore, W defined by (8.17) is invariant under $U^*_{u_0,\vec{v}_0}$.

We now apply the scheme to construct a scale and translation invariant distribution which is not infinitely divisible. Let $\Lambda = \{0, 1\}$. Let $\mathbf{X}_0(\vec{v}) \equiv 0$. Let $\mathbf{X}_1(\vec{v})$ be a white noise process with the characteristic functional

$$E\left(e^{i\langle X_1,\phi\rangle}\right) = \exp\left(-\int_{\mathbf{R}^n} |\phi(\vec{v})|^2 d\vec{v}\right).$$

Let $\mathbf{Y}(u)$ be a random process such that for each sample $Y(u)$, there is a unique $u_0 \in [0, 1)$, $Y(u_0) = 1$. For any u, if $u = u_0 + k$, for some integer k, then $Y(u) = 1$. Otherwise $Y(u) = 0$. Furthermore, u_0 is uniformly distributed in $[0, 1)$. It is then easy to see that $\mathbf{Y}(u)$ is stationary.

Given a sample $Y(u)$, assume $u_0 \in [0, 1)$ satisfies $Y(u_0) = 1$. Define F on $\mathcal{D}_0(\mathbf{R}^n)$ by

$$F : \phi \mapsto \sum_{u\in\mathbf{R}} \langle X_{Y(u)}, \Psi_g \phi\rangle.$$

According to our assumptions, the sum in the above formula equals

$$\sum_{k=-\infty}^{\infty} \langle B_k, \Psi_g \phi(u_0 + k, \cdot) \rangle,$$

where B_k are independent white noises. It is easy to get the characteristic functional of F equal to

$$E\left(e^{i\langle F,\phi\rangle}\right) = \int_0^1 \exp\left(-\sum_{k=-\infty}^{\infty} \int_{\mathbf{R}^n} |\Psi_g \phi(u+k, \vec{v})|^2 d\vec{v}\right) du. \tag{8.18}$$

Proposition 25. The distribution given by the characteristic functional (8.18) is not infinitely divisible.

The proof of Proposition Proposition 25 is quite long, although not difficult. So we put it in the Appendix.

8.8 Scale and Translation Invariant Distributions On $\mathbf{Z}^d$

For $r = \{r_1, \dots, r_d\}$ and $s = \{s_1, \dots, s_d\}$, let $r \le s$ mean $r_i \le s_i$, $1 \le i \le d$. For any scalar c, let it also denote $\{c, c, \dots, c\}$. Denote the cube $[r_1, s_1] \times \cdots [r_d, s_d]$ as $[r, s]$.

The following definition is from Sinai [10]. Let $\mathbf{X}^d$ be the space of realizations of a d-dimensional random field $x = \{x_s,\ s = (s_1, \cdots, s_d) \in \mathbf{Z}^d\}$, where $\mathbf{Z}^d$ is the integer lattice. Each random variable x_s takes on real values, and the space $\mathbf{X}^d$ is a vector space. There is a group $\{T_s,\ s \in \mathbf{Z}^d\}$ of translations acting naturally on the space $\mathbf{X}^d$. The symbol $\mathcal{M}(\mathcal{M}^{\text{st}})$ denotes the space of all probability distribution on $\mathbf{X}^d$ (all stationary distributions on $\mathbf{X}^d$, i.e., distributions invariant with respect to the group $\{T_s^*,\ s \in \mathbf{Z}^d\}$ of translations, where $\{T_s^*,\ s \in \mathbf{Z}^d\}$ is the group adjoint to $\{T_s,\ s \in \mathbf{Z}^d\}$ which acts on $\mathcal{M}$).

For each $0 < \alpha < 2$, introduce the multiplicative semi-group $A_k(\alpha) = A_k$, $k \ge 1$ an integer, of linear endmomorphisms of $\mathbf{X}^d$ whose action is given by the formula

$$\tilde{x}_s = (A_k x)_s = \frac{1}{k^{d\alpha/2}} \sum_{sk \le r < (s+1)k} x_{r_1,\dots,r_d}, \quad s = (s_1, \dots, s_d). \tag{8.19}$$

Let $\{A_k^*,\ k \in \mathbf{N}\}$ denote the adjoint semi-group acting on the space $\mathcal{M}$, i.e.,

$$(A_k^* P)(C) = P(A_k^{-1} C), \qquad C \subset \mathbf{X}^d,\ P \in \mathcal{M}.$$

Definition 9. A probability distribution $P \in \mathcal{M}$ is called a scale and translation invariant distribution, if

$$A_k^* P = P, \quad \forall\, k \ge 1$$
$$T_s^* P = P, \quad \forall\, s \in \mathbf{Z}^d.$$

The basic idea to construct scale and translation invariant distribution on $\mathbf{X}^d$ is as follows. Define linear mapping $G : \mathcal{D}'(\mathbf{R}^d) \to \mathbf{X}^d$, such that for any $I \in \mathcal{D}'(\mathbf{R}^d)$, $G(I) = x = \{x_s,\ s = (s_1, \dots, s_d) \in \mathbf{Z}^d\}$, where

$$x_s = \langle I, T_s \phi \rangle$$

and $\phi = \chi_{[0,1]^d}$. Then by (8.19),

$$\begin{aligned}
\tilde{x}_s &= \frac{1}{k^{d\alpha/2}} \sum_{sk \le r < (s+1)k} x_{r_1,\ldots,r_d} \\
&= \frac{1}{k^{d\alpha/2}} \sum_{sk \le r < (s+1)k} \langle I, T_r\phi \rangle \\
&= \frac{1}{k^{d\alpha/2}} \langle I, \chi_{[sk,(s+1)k]} \rangle \\
&= k^{d(1-\alpha/2)} \langle I, S_k T_s \phi \rangle
\end{aligned}$$

Therefore, if the distribution on I is scale and translation invariant with index $H = d(1 - \alpha/2)$, then G induces a scale and translation invariant distribution defined by (8.19).

The problem is that $\phi = \chi_{[0,1]^d} \notin \mathcal{D}$.

Proposition 26. Suppose $g_1, \ldots, g_d \in C_0^\infty(\mathbf{R})$ such that

$$\int_{\mathbf{R}} g_i = 0, \quad i = 1, \ldots, d.$$

Let $g = g_1 \otimes g_2 \cdots \otimes g_d$. Then $\Psi_g\phi \in C^\infty(\mathbf{R}^d)$ and for any $p \ge 1$, $\Psi_g\phi \in L^p(\mathbf{R}^d)$.

Proof. Without loss of generality, assume $\text{supp}(g_i) \subset [0, e^{u_0}]$, $i = 1, \ldots, d$. Let $I = [0,1]^d$, $U = [0, e^{u_0}]^d$, $H = d(1 - \alpha/2)$, and $\psi = \Psi_g\phi$.

$$\begin{aligned}
& \psi(u, \vec{v}) \\
= \;& e^{Hu} \int_I g(e^u x + \vec{v}) dx \\
= \;& e^{(H-d)u} \int_{e^u I + \vec{v}} g(x) dx \\
= \;& e^{(H-d)u} \int_{(e^u I + \vec{v}) \cap U} g(x) dx \\
= \;& e^{(H-d)u} \prod_{i=1}^{d} \int_{(e^u I + v_i) \cap [0,\ e^{u_0}]} g_i(x) dx
\end{aligned}$$

If $(e^u I + \vec{v}) \cap U = \emptyset$ or $(e^u I + \vec{v}) \supset U$, then $\psi(u, \vec{v}) = 0$.

If $u \ge u_0$, then only when $\vec{v} \in A_u$, where

$$A_u = ([-e^u, e^{u_0}] \setminus [-(e^u - e^{u_0}), 0])^d,$$

can $\psi(u, \vec{v})$ be non-zero, and in this case $|\psi(u, \vec{v})| \le e^{(H-d)u} \sup |g|$. Note that $m(A_u)$ is a constant $2^d m \Big(\text{supp}(g) \Big) = 2^d e^{du_0}$.

Therefore, there is a constant C only depending on p, such that for any $p \ge 1$,

$$\begin{aligned}
\int_{\substack{u \ge u_0 \\ \vec{v} \in \mathbf{R}^d}} |\psi(u, \vec{v})|^p du \; d\vec{v} \;&\le\; \int_{u_0}^{\infty} du \int_{\vec{v} \in A_u} e^{p(H-d)u} (\sup |g|)^p d\vec{v} \\
&\le\; 2^d C \; m \Big(\text{supp}(g) \Big) (\sup |g|)^p < \infty.
\end{aligned}$$

If $u \leq u_0$, then only when $\vec{v} \in B_u$, where

$$B_u = [-e^u, e^{u_0}]^d$$

can $\psi(u, \vec{v})$ be non-zero, and in this case $|\psi(u, \vec{v})| \leq e^{Hu} \sup |g|$. Note that $m(B_u)$ is bounded $2^d m\Big(\operatorname{supp}(g)\Big) = 2^d e^{du_0}$.

Therefore, there is a constant D only depending on p, such that for any $p \geq 1$,

$$\begin{aligned} \int_{\substack{u \leq u_0 \\ \vec{v} \in \mathbf{R}^d}} |\psi(u, \vec{v})|^p du\ d\vec{v} &\leq \int_{u_0}^{\infty} du \int_{\vec{v} \in B_u} e^{pHu} (\sup |g|)^p d\vec{v} \\ &\leq 2^d D\ m\Big(\operatorname{supp}(g)\Big) (\sup |g|)^p < \infty. \end{aligned}$$

This completes the proof. □

We will show how to construct Gaussian scale and translation invariant distributions on $\mathbf{Z}^d$.

Example 4. Choose a sequence of functions $\{f_n\}_n$ such that $f_n \in C_0^\infty(\mathbf{R})$ and $f_n \uparrow 1$. Let

$$\phi_n = \underbrace{f_n \otimes f_n \otimes \cdots f_n}_{d+1\ f_n\text{'s}}.$$

Then we get

$$\phi_n \psi \in \mathcal{D}(\mathbf{R} \times \mathbf{R}^d).$$

Recall that $\psi = \Psi_g \phi$ and $\phi = \chi_{[0,1]^d}$.

Let the distribution on $\mathcal{D}(\mathbf{R} \times \mathbf{R}^d)$ be a Gaussian distribution, such that

$$E\Big(e^{i\langle W,\ h\rangle}\Big) = \exp\Big(-\int_{\mathbf{R} \times \mathbf{R}^d} |h|^2\Big), \quad \forall h \in \mathcal{D}(\mathbf{R} \times \mathbf{R}^d).$$

Let

$$C_{nm}(t) = E\left(e^{it\Big(\langle W,\ \phi_n \Psi_g T_s \phi\rangle - \langle W,\ \phi_m \Psi_g T_s \phi\rangle\Big)}\right).$$

Then we get, for any $s \in \mathbf{Z}^d$,

$$C_{nm}(t) \xrightarrow{\text{u.c.}} 1, \quad n, m \to \infty.$$

Therefore

$$\langle W, \phi_n \Psi_g T_s \phi\rangle - \langle W, \phi_m \Psi_g T_s \phi\rangle \xrightarrow{\text{P}} 0.$$

We then can use diagonal argument to get a subsequence $\phi_{n_m} \in C_0^\infty(\mathbf{R})$ such that w.p. 1,

$$\langle W, \phi_{n_m} \Psi_g T_s \phi\rangle \ \text{ converges, } \forall\ s \in \mathbf{Z}^d. \tag{8.20}$$

For each $W \in \mathcal{D}(\mathbf{R} \times \mathbf{R}^d)$ satisfying (8.20), let

$$\{x_s\}_{s \in \mathbf{Z}^d} = \lim_{m\to\infty} \langle W, \phi_{n_m} \Psi_g T_s \phi \rangle.$$

To show $\{x_s\}$ is translation invariant, pick $s_1, \ldots, s_N \in \mathbf{Z}^d$ and $t_1, \ldots, t_N \in \mathbf{R}$. Then

$$\begin{aligned} E\left[\exp\left(i \sum_{l=1}^{N} t_l x_{s_l}\right)\right] &= \lim_{m\to\infty} E\left[\exp\left(i \sum_{l=1}^{N} t_l \langle W,\ \phi_{n_m} \Psi_g T_{s_l} \phi \rangle\right)\right] \\ &= \exp\left(-\int |\sum_{l=1}^{N} t_l \Psi_g T_{s_l} \phi|^2\right). \end{aligned}$$

On the other hand, for any $s \in \mathbf{Z}^d$,

$$\begin{aligned} E\left[\exp\left(i \sum_{l=1}^{N} t_l x_{s+s_l}\right)\right] &= \exp\left(-\int |\sum_{l=1}^{N} t_l \Psi_g T_{s+s_l} \phi|^2\right) \\ &= \exp\left(-\int |\sum_{l=1}^{N} t_l \Psi_g T_{s_l} \phi|^2\right). \end{aligned}$$

Therefore, $\{x_r\} \sim \{x_{r+s}\}$.

To show $\{x_s\}$ is scale invariant, define $\tilde{x}_s$ by (8.19) and calculate

$$\begin{aligned} E\left[\exp\left(i \sum_{l=1}^{N} t_l \tilde{x}_{s_l}\right)\right] &= \lim_{m\to\infty} E\left[\exp\left(i \sum_{l=1}^{N} \frac{1}{k^{d\alpha/2}} \sum_{s_l k \le r < (s_l+1)k} x_r\right)\right] \\ &= \lim_{m\to\infty} E\left[\exp\left(i \sum_{l=1}^{N} \frac{1}{k^{d\alpha/2}} \sum_{s_l k \le r < (s_l+1)k} \langle W, \phi_{n_m} \Psi_g T_r \phi \rangle\right)\right] \\ &= \lim_{m\to\infty} E\left[\exp\left(i \sum_{l=1}^{N} \langle W, \phi_{n_m} k^{d(1-\alpha/2)} \Psi_g S_k T_{s_l} \phi \rangle\right)\right] \\ &= \exp\left(-\int |\sum_{l=1}^{N} t_l k^{d(1-\alpha/2)} \Psi_g S_k T_{s_l} \phi|^2\right) \\ &= \exp\left(-\int |\sum_{l=1}^{N} t_l \Psi_g T_{s_l} \phi|^2\right) \end{aligned}$$

Therefore, we get $\{x_r\} \sim \{\tilde{x}_r\}$.

8.9 Discussion

In this paper we established a method to construct scale and translation invariant distributions on continuous linear functionals. We can introduce other invariances into the distributions. For example, invariance under orthogonal transforms is quite easy to establish. This is because the group of orthogonal transforms $SO(n)$ is compact. Therefore the product distribution of a scale and translation invariant distribution and the uniform distribution on $SO(n)$ will be scale and translation invariant as well as rotation invariant.

Another generalization is as follows. For simplicity, only consider the case $H = 0$ and $\mathcal{E} = \mathcal{D}_0(\mathbf{R}^n)$. Suppose G is a Lie group of invertible linear transforms on $\mathbf{R}^n$. Suppose that G has a finite number of generators $A_1, \dots A_k$ which commute. Then elements in G can be represented by

$$e^{\sum t_i A_i} = \prod e^{t_i A_i}.$$

Define the operator $S_{\vec{t}}$ on $\mathcal{D}(\mathbf{R}^n)$ by

$$S_{\vec{t}}\,\phi(x) = J(e^{\sum t_i A_i})\phi(e^{\sum t_i A_i} x),$$

where $J(\cdot)$ is the Jacobi of a transform. Let $S^*_{\vec{t}}$ be the adjoint of $S_{\vec{t}}$. We then define a wavelet transform Ψ_g from $\mathcal{D}(\mathbf{R}^n)$ to space of functions on $\mathbf{R}^k \times \mathbf{R}^n$ such that

$$\Psi_g \phi(t, \vec{v}) = (S^*_{\vec{t}} T^*_{\vec{v}} g, \phi).$$

We then get, for any $\vec{t_0} \in \mathbf{R}^k$ and $\vec{v_0} \in \mathbf{R}^n$,

$$\Psi_g T_{\vec{v_0}} S_{\vec{t_0}} \phi(t, \vec{v}) = \Psi_g \phi(\vec{t} + \vec{t_0}, \vec{v} + e^{\sum t_i A_i} \vec{v_0})$$

Introduce operator $U_{\vec{t_0}, \vec{v_0}}$ such that

$$U_{\vec{t_0}, \vec{v_0}} f(\vec{t}, \vec{v}) = f(\vec{t} + \vec{t_0}, \vec{v} + \vec{v_0})$$

As before, we can first construct distribution invariant under $U^*_{\vec{t}, \vec{v}}$ and then distribution invariant under $S_{\vec{t}}$ and $T_{\vec{v}}$. Again, the key observation is that the Jacobi of U is one.

Bibliography

[1] G. Samorodnitsky, M. S. Taqqu. *Stable Non-Gaussian Random Processes, Stochastic Models with Infinite Variance, Stochastic Modeling.* Chapman & Hall. 1994.

[2] M. F. Shlesinger, G. M. Zaslavsky, U. Frisch (editors). *Lévy Flights and Related Topics in Physics*, *Proceedings of the International Workshop Held at Nice, France, 27–30, June, 1994*. Springer-Verlag, Berlin Heidelberg. 1995.

[3] D. C. Knill, D. Field, D. Kersten. Human discrimination of fractal images. *Journal of the Optical Society of America A, Optics and Image Science*, Vol. 7, No. 6. June, 1990.

[4] D. L. Ruderman. Origins of Scaling in Natural images. *Vision Research.* To appear.

[5] S. C. Zhu, D. Mumford. Prior Learning and Gibbs Reaction-Diffusion. *IEEE Trans. on PAMI.* To appear.

[6] S. Geman, Z. Chi. Scale-Invariance in Natural Images and Compositional Systems. In preparation.

[7] D. Mumford, S. C. Zhu, B. Gidas. Stochastic Models for Generic Images. In preparation.

[8] D. Mumford. The Statistical Description of Visual Signals. K. Kirchgassner, O. Mahrenholtz, and R. Mennicken (editors). *ICIAM 95.* Akademie Verlag. 1995.

[9] C. E. Heil, D. F. Walnut. Continuous and Discrete Wavelet Transforms. *SIAM Review*, Vol. 31, No. 4. December, 1994.

[10] Y. G. Sinai. Self-similar Probability Distributions, *Theory of Probability and its Applications*, Vol XXI, Number 1. 1976.

Appendix

Proof of (8.2). Define modulation operator $E_{\vec{v}}$ for each $\vec{v} \in \mathbf{R}^n$ and function $f \in L^2(\mathbf{R}^n)$ as

$$E_{\vec{v}} f(\vec{x}) = e^{2\pi i \vec{v}\cdot x} f(\vec{x}),$$

then

$$\int_{\mathbf{R}} \int_{\mathbf{R}^n} |(f, V(u, \vec{v})g)|^2 du d\vec{v}$$

$$
\begin{aligned}
&= \int_{\mathbf{R}}\int_{\mathbf{R}^n} |(f, D_{e^u}T_{\vec{v}}g)|^2 dud\vec{v} \\
&= \int_{\mathbf{R}}\int_{\mathbf{R}^n} |(D_{e^{-u}}f, T_{\vec{v}}g)|^2 dud\vec{v} \\
&= \int_{\mathbf{R}}\int_{\mathbf{R}^n} |(D_{e^u}\hat{f}, E_{-\vec{v}}\hat{g})|^2 dud\vec{v} \qquad \text{(by Parseval's formula)} \\
&= \int_{\mathbf{R}}\int_{\mathbf{R}^n} \left|\int_{\mathbf{R}^n} e^{-nu/2}\hat{f}(e^{-u}\vec{\gamma})\overline{\hat{g}(\vec{\gamma})}e^{2\pi i\vec{\gamma}\cdot\vec{v}}d\vec{\gamma}\right|^2 dud\vec{v} \\
&= \int_{\mathbf{R}}\int_{\mathbf{R}^n} \left|(D_{e^u}\hat{f}\cdot\overline{\hat{g}})^{\vee}(\vec{v})\right|^2 dud\vec{v} \\
&= \int_{\mathbf{R}}\int_{\mathbf{R}^n} \left|D_{e^u}\hat{f}\cdot\overline{\hat{g}}(\vec{\gamma})\right|^2 d\vec{\gamma}du \qquad \text{(by Plancheral's formula)} \\
&= \int_{\mathbf{R}}\int_{\mathbf{R}^n} e^{-u}|\hat{f}(e^{-u}\vec{\gamma})|^2|\hat{g}(\vec{\gamma})|^2 d\vec{\gamma}du \\
&= \int_{\mathbf{R}^n} |\hat{f}(\xi)|^2 \int_{\mathbf{R}} |\hat{g}(e^u\xi)|^2 dud\xi \quad (\xi = e^u\vec{\gamma}).
\end{aligned}
$$

For each $\xi \neq 0$, let

$$\omega = \frac{\xi}{|\xi|},$$

and make substitution

$$t = e^u|\xi|,$$

then

$$\int_{\mathbf{R}} |\hat{g}(e^u\xi)|^2 du = \int_0^\infty \frac{|\hat{g}(t\omega)|^2}{t}dt = c_{g,\omega}.$$

Hence (8.2) is proved. □

We now give an explanation of (8.3), which is

$$f(x) = \int_{\mathbf{R}}\int_{\mathbf{R}^n} \Phi_g Af(u,\vec{v})AD_{e^u}T_{\vec{v}}g(x)dud\vec{v}.$$

We first define an approximate identity $\{\rho_k\}_{k=1}^{\infty}$. For each k, $\rho_k(x) = k^{n/2}\rho(kx)$, where $\rho \in L^1(\mathbf{R}^n) \cup L^2(\mathbf{R}^n)$ such that $\int \rho(x)dx = 1$ and $\rho(x) = \rho(-x)$. Then we say (8.3) holds in the sense that $\lim_{k\to\infty} \|f - f_k\| = 0$, where

$$f_k(x) = \int_{\mathbf{R}}\int_{\mathbf{R}^n} \Phi_g Af(u,\vec{v})(\rho_k * AD_{e^u}T_{\vec{v}}g)(x)dud\vec{v}.$$

The proof of the limit is as follows.

$$
\begin{aligned}
& (f * \rho_k)(x) \\
&= \int_{\mathbf{R}^n} f(y)\rho_k(x-y)dy \\
&= (f, \overline{T_x\rho_k}) \qquad \text{(since } \rho_k \text{ is even)}
\end{aligned}
$$

$$
\begin{aligned}
&= (\Phi_g Af, \Phi_g A(\overline{T_x\rho_k})) \qquad\qquad \text{(since } \Phi_g A \text{ is an isometry)}\\
&= \int_{\mathbf{R}}\int_{\mathbf{R}^n} \Phi_g Af(u,\vec{v})(D_{e^u}T_{\vec{v}}g, A\overline{T_x\rho_k})dud\vec{v}\\
&= \int_{\mathbf{R}}\int_{\mathbf{R}^n} \Phi_g Af(u,\vec{v})(AD_{e^u}T_{\vec{v}}g, \overline{T_x\rho_k})dud\vec{v} \qquad\qquad \text{(since } A \text{ is self adjoint)}\\
&= \int_{\mathbf{R}}\int_{\mathbf{R}^n} \Phi_g Af(u,\vec{v})(\rho_k * AD_{e^u}T_{\vec{v}}g)(x)dud\vec{v}.
\end{aligned}
$$

But $\{\rho_k\}$ is an approximate identity, so $\lim_{k\to\infty}\|f*\rho_k - f\|_{L^2} = 0$. □

We now prove the three propositions given in §8.4. First, we need to prove several lemmas.

Lemma 11. If $g \in \mathcal{D}(\mathbf{R}^n)$, then for any $\phi \in \mathcal{D}(\mathbf{R}^n)$, $\Psi_g\phi \in C^\infty(\mathbf{R}\times\mathbf{R}^n)$.

The proof of Lemma Lemma 11 is left to the reader.

Lemma 12. If $g \in \mathcal{D}(\mathbf{R}^n)$ has vanishing first $k-1$ moments

$$\int_{\mathbf{R}^n} x^\alpha g(x)dx = 0, \quad \alpha \geq 0,\ |\alpha| \leq k-1,$$

and compact support

$$\text{supp}(g) \subset B_r(0) = U,$$

then for $H \geq 0$ and $p > 0$ with $p(n+k-H) > n$, there is a positive integer m_0, such that for any $m \geq m_0$,

$$|\Psi_g\phi(u,\vec{v})| \leq \frac{e^{u(H-n-1)}}{\Big(|\vec{v}/e^u|+1\Big)^m} Kb_m(\phi), \qquad \text{for } u \geq 0, \vec{v} \in \mathbf{R}^n, \tag{A8.1}$$

$$\int_{\substack{u\geq a\\ \vec{v}\in\mathbf{R}^n}} |\Psi_g\phi(u,\vec{v})|^p dud\vec{v} < C(p,n,H,m)\Big(Kb_m(\phi)\Big)^p, \tag{A8.2}$$

$$|\Psi_g\phi(u,\vec{v})| \leq \frac{e^{u(m+H-n-1)}}{|\vec{v}|^m} Kb_m(\phi), \qquad \text{for } u < 0, |\vec{v}| \geq 2r, \tag{A8.3}$$

$$\int_{\substack{u\leq a\\ |\vec{v}|\geq 2r}} |\Psi_g\phi(u,\vec{v})|^p dud\vec{v} < D(p,n,H,m)\Big(Kb_m(\phi)\Big)^p, \tag{A8.4}$$

where $C(p,n,H,m)$ and $D(p,n,H,m)$ are constants only depending on p,n,H,m, K is a constant only depending of g and

$$b_m(\phi) = 2^m \max_{\substack{\alpha\geq 0,|\alpha|\leq m\\ \beta\geq 0,|\beta|\leq m}} \max_x \Big(|x|+R\Big)^{|\beta|} |\partial_\alpha\phi(x)|, \tag{A8.5}$$

where $R \geq 1$ is a constant only depending on g.

Proof. For simplicity, only consider the case $k = 1$. For $k > 1$, the result is similarly proved.

If $u \geq 0$, then by Taylor's expansion

$$\begin{aligned}
& \Psi_g \phi(u, \vec{v}) \\
= \;& e^{Hu}(S^*_{e^u} T^*_{\vec{v}}, \phi) \\
= \;& e^{Hu} \int_{\mathbf{R}^n} g(e^u x + \vec{v}) \phi(x) dx \\
= \;& \frac{e^{Hu}}{e^{un}} \int_U \phi\left(e^{-u}(x - \vec{v})\right) g(x) dx \\
= \;& e^{(H-n)u} \int_U \left[\phi(-e^{-u}\vec{v}) + e^{-u} x \cdot \phi'\left(e^{-u}(\theta x - \vec{v})\right)\right] g(x) dx, \quad \theta = \theta(x) \in [0,1] \\
= \;& e^{(H-n-1)u} \int_U x \cdot \phi'\left(e^{-u}(\theta x - \vec{v})\right) g(x) dx,
\end{aligned}$$

Hence

$$|\Psi_g \phi(u, \vec{v})| \leq e^{(H-n-1)u} \int_U |x| \cdot |g(x)| dx \cdot \max_{x \in U} \left|\phi'\left(e^{-u}(\theta x - \vec{v})\right)\right|.$$

Letting

$$K = \int_U |x| \cdot |g(x)| dx,$$

we get

$$|\Psi_g \phi(u, \vec{v})| \leq e^{(H-n-1)u} K \max_{x \in U} \left|\phi'\left(e^{-u}(\theta x - \vec{v})\right)\right|. \tag{A8.6}$$

Since $e^u \geq 1$, $0 \leq \theta \leq 1$ and $|x| \leq r$ on U, then

$$|e^{-u}\vec{v}| \leq |e^{-u}(\theta x - \vec{v})| + r.$$

Choose $R \geq r + 1$. Then for any $x \in U$,

$$\begin{aligned}
& \left(|e^{-u}\vec{v}| + 1\right)^m \left|\phi'\left(e^{-u}(\theta x - \vec{v})\right)\right| \\
\leq \;& \left(|e^{-u}(\theta x - \vec{v})| + r + 1\right)^m \left|\phi'\left(e^{-u}(\theta x - \vec{v})\right)\right| \\
\leq \;& \left(|e^{-u}(\theta x - \vec{v})| + R\right)^m \left|\phi'\left(e^{-u}(\theta x - \vec{v})\right)\right| \\
\leq \;& b_m(\phi).
\end{aligned}$$

The above inequalities then imply

$$|\Psi_g \phi(u, \vec{v})| \leq \frac{e^{Hu}}{e^{u(n+1)}} \frac{K b_m(\phi)}{\left(|\vec{v}/e^u| + 1\right)^m},$$

This proves (A8.1). Note that, since $p(n + 1 - H) > n$, $H < n + 1$. Hence from the above inequality, $\Psi_g \phi(u, \vec{v})$ is bounded on $u \geq 0$, $\vec{v} \in \mathbf{R}^n$. Let

$$m_0 > \max(n + k - H, n/p, n + 1). \tag{A8.7}$$

Then for $m \geq m_0$,

$$\begin{aligned}
& \int_{[0,\infty)\times\mathbf{R}^n} |\Psi_g\phi(u,\vec{v})|^p du d\vec{v} \\
\leq & \int_{[0,\infty)\times\mathbf{R}^n} \left(\frac{e^{(H-n-1)u}}{|\vec{v}/e^u|^m+1}\right)^p \Big(Kb_m(\phi)\Big)^p du d\vec{v}, \quad (\vec{v}\leftarrow e^u\vec{v}) \\
= & \int_{[0,\infty)\times\mathbf{R}^n} e^{nu}\left(\frac{e^{(H-n-1)u}}{|\vec{v}|^m+1}\right)^p \Big(Kb_m(\phi)\Big)^p du d\vec{v} \\
= & \int_{[0,\infty)} e^{p(H-n-1)u+nu} du \int_{\mathbf{R}^n} \frac{1}{(|\vec{v}|^m+1)^p} d\vec{v} \cdot \Big(Kb_m(\phi)\Big)^p < \infty
\end{aligned}$$

since $p(H-n-1)+n<0$. This proves (A8.2).

To prove (A8.3), consider again (A8.6). This time, we have, $u<0$ and, for any $x \in U = B_r(0)$,

$$|\theta x + \vec{v}| \geq |\vec{v}| - |x| \geq |\vec{v}| - r \geq \frac{|\vec{v}|}{2}.$$

Therefore

$$|e^{-u}\vec{v}|^m |\phi'\Big(e^{-u}(\theta x+\vec{v})\Big)| \leq \; 2^m |e^{-u}\Big(\theta x+\vec{v}\Big)|^m \cdot |\phi'\Big(e^{-u}(\theta x+\vec{v})\Big)| \leq \; b_m(\phi),$$

which implies

$$\max_{x\in U} |\phi'\Big(e^{-u}(\theta x+\vec{v})\Big)| \leq \frac{e^{mu}b_m(\phi)}{|\vec{v}|^m}.$$

This together with (A8.6) proves (A8.3). (A8.4) can be proved similarly as (A8.2). □

Lemma 13. (The case $H>0$) Suppose $g \in \mathcal{D}(\mathbf{R}^n)$ has vanishing first $k-1$ moments

$$\int_{\mathbf{R}^n} x^\beta g(x)dx = 0, \quad \beta \geq 0, \; |\beta| \leq k-1,$$

and compact support

$$\mathrm{supp}(g) \subset B_r(0) = U.$$

If $H>0$ in the definition of Ψ_g, then for any $0<p\leq\infty$ with $p(n+k-H)>n$, for any integer $s \geq 0$, Ψ_g is a continuous mapping from $\mathcal{D}$ into $C^\infty(\mathbf{R}\times\mathbf{R}^n)$ and into $W^{p,s}(\mathbf{R}\times\mathbf{R}^n)$, where

$$W^{p,s} = \{f \in C^s(\mathbf{R}\times\mathbf{R}^n) : \; \partial_\beta f \in L^p(\mathbf{R}\times\mathbf{R}^n), \text{ for all } \beta \geq 0 \text{ s.t. } |\beta| \leq s\}.$$

Proof. Consider the values of $\Psi_g\phi$ on the region $\{u<0, |\vec{v}| \leq 2r\}$. If $u<0$ and $|\vec{v}| \leq 2r$, then if $m>n$,

$$\begin{aligned}
|\Psi_g\phi(u,\vec{v})| &\leq e^{Hu}\int_{\mathbf{R}^n} |g(e^u x - \vec{v})\phi(x)|dx \\
&\leq e^{Hu}\int_{\mathbf{R}^n} \frac{|g(e^u x - \vec{v})|}{(|x|+R)^m} b_m(\phi)dx \\
&\leq e^{Hu}Kb_m(\phi),
\end{aligned}$$

where

$$K = \sup|g| \int_{\mathbf{R}^n} \frac{dx}{(|x|+R)^m},$$

and $m > m_0$ defined by (A8.7) and b_m is defined by (A8.5).

Together with Lemma Lemma 12, this shows that Ψ_g is a continuous map to $L^p(\mathbf{R}\times\mathbf{R}^n)$ and to $C^\infty(\mathbf{R}\times\mathbf{R}^n)$. The case $p=\infty$ is implied by the continuity of the mapping into $C^\infty(\mathbf{R}\times\mathbf{R}^n)$.

That Ψ_g is a continuous mapping from $\mathcal{D}$ into $W^{p,s}(\mathbf{R}\times\mathbf{R}^n)$, $s>0$, can be similarly proved. For example, if $D^\beta = \partial/\partial u$, then

$$\begin{aligned}
&\frac{\partial}{\partial u}\Psi_g\phi(u,\vec{v}) \\
= \;& H\Psi_g\phi(u,\vec{v}) + e^{Hu}\int_{\mathbf{R}^n} e^u \sum_{i=1}^n x_i \frac{\partial}{\partial x_i} g(e^u x - \vec{v})\phi(x)dx \\
= \;& H\Psi_g\phi(u,\vec{v}) + e^{Hu}\int_{\mathbf{R}^n} g(e^u x - \vec{v}) \sum_{i=1}^n \frac{\partial}{\partial x_i}(x_i\phi(x))dx \\
= \;& H\Psi_g\phi(u,\vec{v}) + \sum_{i=1}^n \Psi_g\phi_i(u,\vec{v}),
\end{aligned}$$

where

$$\phi_i(x) = \frac{\partial}{\partial x_i}(x_i\phi(x)).$$

Since $\phi_i \in \mathcal{D}(\mathbf{R}^n)$, then $\Psi_g\phi_i(u,\vec{v}) \in L^p(\mathbf{R}\times\mathbf{R}^n)$ for any $p \geq 0$. Hence (A8.1) – (A8.4) holds for $D^\beta = \partial/\partial_u$. □

Lemma 14. (The case $H \leq 0$) Suppose $g \in \mathcal{D}(\mathbf{R}^n)$ has vanishing integral and compact support

$$\operatorname{supp}(g) \subset B_r(0) = U.$$

If $H \leq 0$ in the definition of Ψ_g, then for any $0 < p \leq \infty$ with $p(n+1-H) > n$, for any integer $s \geq 0$, Ψ_g is a continuous mapping from $\mathcal{D}_{k-1}(\mathbf{R}^n)$, $k > -H$, into $C^\infty(\mathbf{R}\times\mathbf{R}^n)$ and into $W^{p,s}(\mathbf{R}\times\mathbf{R}^n)$.

Proof. Again, we only need to consider the values of $\Psi_g\phi$ on the region $\{u<0, |\vec{v}| \leq 2r\}$. This time, if $\phi \in \mathcal{D}_{k-1}(\mathbf{R}^n)$, then by Taylor's expansion,

$$\begin{aligned}
&\Psi_g\phi(u,\vec{v}) \\
= \;& e^{Hu}\int_{\mathbf{R}^n} g(e^u x - \vec{v})\phi(x)dx \\
= \;& e^{Hu}\int_{\mathbf{R}^n}\Bigg(\sum_{|\alpha|\leq k-1}\frac{1}{|\alpha|!}\partial_\alpha g(-\vec{v}) + \sum_{|\alpha|=k} e^{ku}x^\alpha \cdot \partial_\alpha g(\theta e^u x - \vec{v})\Bigg)\phi(x)dx, \quad 0 \leq \theta \leq 1 \\
= \;& \sum_{|\alpha|=k} e^{(k+H)u}\int_{\mathbf{R}^n} x^\alpha \cdot \partial_\alpha g(\theta e^u x - \vec{v})\phi(x)dx, \quad 0 \leq \theta \leq 1 \\
\leq \;& e^{(k+H)u}Kb_m(\phi), \quad u \to -\infty,
\end{aligned}$$

where $m > m_0$ defined by (A8.7) and where K is a constant only depending on g. Together with Lemma Lemma 12, this shows that Ψ_g is a continuous mapping from $\mathcal{D}_{k-1}(\mathbf{R}^n)$ into $L^p(\mathbf{R} \times \mathbf{R}^n)$ and to $C^\infty(\mathbf{R} \times \mathbf{R}^n)$, the case of $p = \infty$ is implied by the latter case.

That Ψ_g is a continuous mapping from $\mathcal{D}_{k-1}$ into $W^{p,s}(\mathbf{R} \times \mathbf{R}^n)$, $s > 0$ is similarly treated by the method described at the end of the proof of Lemma Lemma 13. The only one more thing that needs to be noted is for any $\beta \geq 0$, $D^\beta(x^\beta \phi(x))$ also has integral zero. □

Proof of Proposition Proposition 21: Directly from Lemma Lemma 13 and Lemma 14, where $s = 0$ and $p \geq 1$. □

Proof of Proposition Proposition 22: Fix m large enough, depending on H (this time $p = 1$), define function

$$h(u, \vec{v}) = \begin{cases} e^{u(H-n-1)}(|e^{-u}\vec{v}|^m + 1)^{-1}, & u \geq 0, \vec{v} \in \mathbf{R}^n \\ e^{u(m+H-n-1)}|\vec{v}|^{-m}, & u < 0, |\vec{v}| \geq 2r \\ e^{Hu}, & u < 0, |\vec{v}| < 2r, \text{ if } 0 < H < 1 \\ e^{(1+H)u}, & u < 0, |\vec{v}| < 2r, \text{ if } -1 < H \leq 0. \end{cases}$$

First we show that for almost all samples $\{(u_i, \vec{v}_i)\}$ from the Poisson process,

$$\sum_i (|w_{u_i,\vec{v}_i}|, h) < \infty.$$

Indeed, it is necessary and sufficient to show that

$$\int_{\mathbf{R}\times\mathbf{R}^n} (|w_{u,\vec{v}}|, h) du d\vec{v} < \infty. \tag{A8.8}$$

This can be done as follows

$$\begin{aligned} & \int_{\mathbf{R}\times\mathbf{R}^n} |(w_{u,\vec{v}}, h)| du d\vec{v} \\ = & \int_{\mathbf{R}\times\mathbf{R}^n} \left| \int_{\mathbf{R}\times\mathbf{R}^n} \mathcal{W}(x-u, \vec{y} - e^{x-u}\vec{v}) h(x, \vec{y}) dx d\vec{y} \right| du d\vec{v} \\ = & \int_{\mathbf{R}\times\mathbf{R}^n} \left| \int_{\mathbf{R}\times\mathbf{R}^n} \mathcal{W}(x, \vec{y}) h(x+u, \vec{y} + e^x \vec{v}) dx d\vec{y} \right| du d\vec{v} \\ \leq & \int_{\mathbf{R}\times\mathbf{R}^n} |\mathcal{W}(x, \vec{y})| dx d\vec{y} \int_{\mathbf{R}\times\mathbf{R}^n} |h(x+u, \vec{y} + e^x \vec{v})| du d\vec{v} \\ = & \int_{\mathbf{R}\times\mathbf{R}^n} \frac{|\mathcal{W}(x, \vec{y})|}{e^{nx}} dx d\vec{y} \cdot \|h\|_1 < \infty. \end{aligned}$$

When $0 < H < 1$, for any $\phi \in \mathcal{D}(\mathbf{R}^n)$, for any sample $\{(u_i, \vec{v}_i)\}$ which satisfies (A8.8),

$$|(w_{u_i,\vec{v}_i}, \Psi_g \phi)| \leq (|w_{u_i,\vec{v}_i}|, |\Psi_g \phi|) \leq K b_m(\phi)(|w_{u_i,\vec{v}_i}|, h),$$

where K is a constant only depending on g and b_m is defined as (A8.5). This shows that

$$\sum_i (w_{u_i,\vec{v}_i}, \Psi_g \phi)$$

converges unconditionally and the sum is continuous in ϕ. Hence

$$F = \sum_i \Psi_g^* w_{u_i,\vec{v}_i}$$

is a well-defined continuous linear functional on $\mathcal{D}(\mathbf{R}^n)$.

To prove that the distribution has finite covariance, we compute

$$\begin{aligned}
&\int_{\mathbf{R}\times\mathbf{R}^n} (|w_{u,\vec{v}}|, h)^2 du d\vec{v} \\
= \;& \int_{\mathbf{R}\times\mathbf{R}^n} \left(\int_{\mathbf{R}\times\mathbf{R}^n} |\mathcal{W}(x-u, \vec{y} - e^{x-u}\vec{v})| h(x,\vec{y}) dx d\vec{y} \right)^2 du d\vec{v} \\
= \;& \int_{\mathbf{R}\times\mathbf{R}^n} \left(\int_{\mathbf{R}\times\mathbf{R}^n} |\mathcal{W}(x,\vec{y})| h(x+u, \vec{y} + e^{x}\vec{v}) dx d\vec{y} \right)^2 du d\vec{v} \\
\leq \;& \|\mathcal{W}\|_{L^1} \cdot \int_{\mathbf{R}\times\mathbf{R}^n} |\mathcal{W}(x,\vec{y})| dx d\vec{y} \int_{\mathbf{R}\times\mathbf{R}^n} |h(x+u, \vec{y} + e^{x}\vec{v})|^2 du d\vec{v} \\
= \;& \|\mathcal{W}\|_{L^1} \cdot \int_{\mathbf{R}\times\mathbf{R}^n} \frac{|\mathcal{W}(x,\vec{y})|}{e^{nx}} dx d\vec{y} \cdot \|h\|_2^2 < \infty,
\end{aligned}$$

proving that $\sum(|w_{u_i,\vec{v}_i}|, h)$ has finite second moment. Then it is easy to show for $\phi, \psi \in \mathcal{D}(\mathbf{R}^n)$, $\sum(w_{u_i,\vec{v}_i}, \Psi_g \phi)$ and $\sum(w_{u_i,\vec{v}_i}, \Psi_g \psi)$ have finite covariance.

The case $-1 < H \leq 0$ is similarly treated. □

The proof of Proposition Proposition 23 is similar to Proposition Proposition 22. Hence we omit the proof.

We now give a proof for Proposition Proposition 24.

Proposition Proposition 24. If the random continuous linear functional f is defined on $\mathcal{D}$ and is distributed by a scale and translation invariant law with finite covariance, then f is a random variable.

Proof. For simplicity, only consider f with one variable. Translation $T_x : \mathcal{D} \to \mathcal{D}$, $x \in \mathbf{R}$ and and scaling $S_t : \mathcal{D} \to \mathcal{D}$, $t \in \mathbf{R}^+$ are defined as

$$\begin{aligned}
T_x : \;& \phi(u) \mapsto \phi(u-x), \\
S_t : \;& \phi(u) \mapsto \frac{\phi(t^{-1}u)}{t},
\end{aligned}$$

for $\phi(u) \in \mathcal{D}$, respectively. Then

$$\langle f, T_x\phi \rangle \sim \langle f, \phi \rangle \Rightarrow E(\langle f, T_x\phi \rangle) = E(\langle f, \phi \rangle).$$

Let $\bar{f}$ be the mean of f. Then $\bar{f}$ is a distribution and

$$\langle \bar{f}, T_x\phi \rangle = \langle \bar{f}, \phi \rangle \Rightarrow \frac{d}{dx}\langle \bar{f}, T_x\phi \rangle = 0.$$

It can be shown that

$$\frac{d}{dx}\langle f, T_x\phi \rangle = \langle f, \frac{d}{dx}(T_x\phi) \rangle.$$

Since

$$\frac{d}{dx}(T_x\phi(u))\Big|_{x=0} = -\phi'(u),$$

then

$$\langle \bar{f}, \phi' \rangle = 0 \tag{A8.9}$$

Suppose $\psi \in \mathcal{D}$. Consider

$$B(\phi,\psi) = E(\langle f,\phi\rangle \cdot \langle f,\psi\rangle).$$

Without loss of generality, we assume B is real. Also note that B is symmetric, i.e., $B(\phi,\psi) = B(\psi,\phi)$.

By translation invariance,

$$B(\phi,\psi) = B(T_x\phi, T_x\psi), \qquad \text{for any } x \in \mathbf{R}.$$

Differentiate the right hand side with respect to x at $x = 0$,

$$B(\phi',\psi) + B(\phi,\psi') = 0, . \tag{A8.10}$$

By scale invariance,

$$B(\phi,\psi) = B(S_t\phi, S_t\psi), \qquad \text{for any } t \in \mathbf{R}^+. \tag{A8.11}$$

Since

$$\frac{d}{dt}(S_t\phi(u))\Big|_{t=1} = \frac{d}{dt}\left(\frac{\phi(t^{-1}u)}{t}\right)\Big|_{t=1} = -\phi(u) - u\phi'(u) = -(u\phi(u))',$$

then, by differentiating the right hand side of (A8.11) at $t = 1$,

$$B((u\phi)',\psi) + B(\phi,(u\psi)') = 0.$$

Note that since $\phi(u) \in \mathcal{D}$, $u\phi(u) \in \mathcal{D}$, hence the above equation makes sense. By (A8.10),

$$B(u\phi,\psi') + B(\phi',u\psi) = 0. \tag{A8.12}$$

Take $\psi = \phi$, then by (A8.10) and (A8.12)

$$\left\{\begin{array}{l} B(\phi,\ \phi') = 0 \\ B(u\phi,\phi') = 0. \end{array}\right. \tag{A8.13}$$

Fix $x \in \mathbf{R}$ and $t \in \mathbf{R}^+$, take

$$\phi(u) = te^{-(u+x)^2t^2/2} = T_xS_tg(u),$$

where g is the Gaussian function

$$g(u) = e^{-\frac{u^2}{2}},$$

which is in $\mathcal{D}$. Then $\phi'(u) = -t^2(u+x)\phi(u)$. Then by (A8.13),

$$B(\phi', \phi') = B(-t^2(u+x)\phi, \phi') = -t^2 B(u\phi, \phi') - xt^2 B(\phi, \phi') = 0.$$

This equation, together with (A8.9), implies for each $x \in \mathbf{R}$ and $t \in \mathbf{R}^+$, with probability one,

$$\langle f, \phi' \rangle = 0,$$

in other words,

$$\langle f, \Big(T_x S_t g \Big)' \rangle = 0.$$

The probability that the above equation holds for all $x \in \mathbf{Q}$ and $t \in \mathbf{Q}^+$ is one. By continuity, with probability one, the above equation holds for all real x and positive t. Since

$$\Big(T_x S_t g \Big)' = \frac{\partial}{\partial x} \Big(T_x S_t g \Big)$$

and the action of f and $\partial / \partial x$ commute, we get

$$\frac{\partial}{\partial x} \langle f, T_x S_t g \rangle = 0,$$

which implies $\langle f, T_x S_t g \rangle$ does not depend on x. Since $S_t g$ with their translations form a basis for $\mathcal{D}$, this implies that with probability one, for any $\phi \in \mathcal{D}$, $\langle f, T_x \phi \rangle$ does not depend on x. More "precisely", for any $\phi \in \mathcal{D}$,

$$\phi = \sum_{y,t} c_{y,t} T_y S_t g,$$

then with probability one, for any x,

$$\langle f, T_x \phi \rangle = \sum_{y,t} c_{y,t} \langle f, T_x T_y S_t g \rangle = \sum_{y,t} c_{y,t} \langle f, T_y S_t g \rangle = \langle f, \phi \rangle.$$

Hence f is a random variable with probability one. □

Finally, we prove Proposition Proposition 25.

Proposition Proposition 25. The distribution given by the characteristic functional (8.18) is not infinitely divisible.

Proof. Assume F is distributed by a scale and translation invariant law, then for any $\phi \in \mathcal{D}_0$, the random variable $\langle F, \phi \rangle$ is distributed by a scale and translation invariant law with characteristic function

$$Ee^{it\langle F,\, \phi \rangle} = \int_0^1 e^{-t^2 f(u)} du,$$

where

$$f(u) = \sum_{k=-\infty}^{\infty} \int |\Psi_g \phi(u+k, v)|^2 dv.$$

By Lévy's theorem,

$$\log\left(\int_0^1 e^{-t^2 f(u)}du\right) = i\beta t - \frac{\sigma^2 t^2}{2} + \int \left(e^{itx} - 1 - \frac{itx}{1+x^2}\right)\frac{1+x^2}{x^2}\nu(dx),$$

where $\nu(dx)$ is a finite measure on $\mathbf{R} \setminus \{0\}$.

Replace t by $-t$. Then

$$\log\left(\int_0^1 e^{-t^2 f(u)}du\right) = -i\beta t - \frac{\sigma^2 t^2}{2} + \int \left(e^{-itx} - 1 + \frac{itx}{1+x^2}\right)\frac{1+x^2}{x^2}\nu(dx).$$

Take the average of the above two equations to get

$$\log\left(\int_0^1 e^{-t^2 f(u)}du\right) = -\frac{\sigma^2 t^2}{2} + \int \left(\cos tx - 1\right)\frac{1+x^2}{x^2}\nu(dx). \tag{A8.14}$$

It can be shown that it is mathematically correct to replace t in the above equation by it. Then

$$\log\left(\int_0^1 e^{t^2 f(u)}du\right) = \frac{\sigma^2 t^2}{2} + \int \left(\cosh tx - 1\right)\frac{1+x^2}{x^2}\nu(dx). \tag{A8.15}$$

Divide both sides of (A8.14) by t^2 and let $t \to \infty$. The left hand side converges to $-\min f(u)$ and the right hand side converges to $-\sigma^2/2$.

Divide both sides of (A8.15) by t^2 and let $t \to \infty$. The left hand side converges to $\max f(u)$ and the right hand side converges to ∞, if $\nu \neq 0$, or $\sigma^2/2$, if $\nu = 0$.

Therefore, the distribution is infinitely divisible if and only if $f(u)$ is a constant. However, we can find a $\phi \in \mathcal{D}_0$ such that the corresponding f is not a constant. Hence F can not have a scale and translation invariant distribution. □

To see why in (A8.14), t can be replaced by it, consider the function

$$K(t) = \log\left(\int_0^1 e^{tf(u)}du\right).$$

Since $f \geq 0$, $K(t)$ is an strictly increasing function, unless $f(u) \equiv 0$. From the earlier arguments, we see that $\sigma^2 = \min f(u)$. Therefore, w.l.o.g., we can assume $\sigma = 0$ and $\min f(u) = 0$. Write $G(dx) = (1 + x^{-2})\nu(dx)$. Since $1 - \cos tx$ is even in x, we can assume $G(dx)$ only has mass on $(0, +\infty)$.

$$\frac{1}{t^2}(K(0) - K(-t^2)) = \int \frac{1}{t^2}\left(1 - \cos tx\right)G(dx).$$

Letting $t \to 0$, we get

$$K'(0) \geq \int \frac{x^2}{2}G(dx).$$

Since $0 \leq t^{-2}(1-\cos tx) \leq x^2/2$, then by dominant convergence, we get

$$K'(0) = \int \frac{x^2}{2!} G(dx) \geq 0.$$

Suppose we have proved that for $n = 1, \dots k$,

$$\frac{K^{(n)}(0)}{n!} = \int \frac{x^{2n}}{(2n)!} G(dx) \geq 0.$$

Then

$$\begin{aligned} & \frac{1}{t^{2k+2}} \left(K(0) - \frac{K'(0)t^2}{1!} + \frac{K''(0)t^4}{2!} - \dots + \frac{(-1)^k K^{(k)}(0) t^{2k}}{k!} - K(-t^2) \right) \\ = & \int \frac{1}{t^{2k+2}} \left(1 - \frac{t^2x^2}{2!} + \frac{t^4x^4}{4!} - \dots + \frac{(-1)^k t^{2k} x^{2k}}{(2k)!} - \cos tx \right) G(dx). \end{aligned} \tag{A8.16}$$

Then

$$\begin{aligned} & \left| \frac{1}{t^{2k+2}} \left(K(0) - \frac{K'(0)t^2}{1!} + \frac{K''(0)t^4}{2!} - \dots + \frac{(-1)^k K^{(k)}(0) t^{2k}}{k!} - K(-t^2) \right) \right| \\ = & \int \frac{1}{t^{2k+2}} \left| 1 - \frac{t^2x^2}{2!} + \frac{t^4x^4}{4!} - \dots + \frac{(-1)^k t^{2k} x^{2k}}{(2k)!} - \cos tx \right| G(dx). \end{aligned} \tag{A8.17}$$

The last equation is because as k is even, for any $x \geq 0$,

$$D_k(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots + \frac{(-1)^k x^{2k}}{2k!} - \cos x \geq 0$$

and as k is odd, for any $x \geq 0$, $D_k(x) \leq 0$. From this we also get

$$|D_k(x)| \leq \frac{x^{2k+2}}{(2k+2)!} \tag{A8.18}$$

By letting $t \to 0$, from (A8.17) we get

$$\left| \frac{K^{(k+1)}(0)}{(k+1)!} \right| \geq \int \frac{x^{2k+2}}{(2k+2)!} G(dx).$$

From (A8.16), (A8.18) and dominant convergence,

$$\frac{K^{(k+1)}(0)}{(k+1)!} = \int \frac{x^{2k+2}}{(2k+2)!} G(dx) > 0.$$

Let

$$H(z) = \int_0^1 e^{zf(u)} du, z \in \mathbf{C}.$$

Since $H(z)$ is analytic and $H(0) = 1$, $K(z) = \log H(z)$ is analytic in a neighborhood of 0. Therefore, there is an $R > 0$, such that for $t \in (-R, R)$,

$$K(t) = \sum_{n=1}^{\infty} \frac{K^{(n)}(0)t^n}{n!}. \tag{A8.19}$$

Since $H(z) \neq 0$, for $z \in \mathbf{R}$, $K(z)$ is also analytic in a neighborhood of the real line. Because $K(t)$ is increasing, and $K^{(n)}(0) \geq 0$, for any $n \geq 0$, the series given in (A8.19) is convergent on the whole real line. Assume this is not true. Let L be the sup of the set of $t > 0$ such that the series (A8.19) converges. For any $t < L$,

$$K(t) = \sum_{n=1}^{\infty} \frac{K^{(n)}(0)t^n}{n!} < K(L+1) < \infty.$$

Let $t \uparrow L + 1$, we get

$$\sum_{n=1}^{\infty} \frac{K^{(n)}(0)(L+1)^n}{n!} < \infty,$$

which is a contradiction.

Because the series (A8.19) is convergent on the whole real line, therefore, for any t,

$$K(t^2) = \sum_{n=1}^{\infty} \frac{K^{(n)}(0)t^{2n}}{n!} = \int \sum_{n=1}^{\infty} \frac{x^{2n}t^{2n}}{(2n)!} G(dx) = \int (\cosh tx - 1) G(dx).$$

This completes the proof. □