Image Compression and Data Clustering: New Takes on Some Old Problems

by

Wei-Ying Wang

B.A., National Taiwan University; Taiwan, 2004

M.Sc., National Taiwan University; Taiwan, 2006

M.Sc., Brown University; Providence, RI, 2012

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Division of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2017

This dissertation by Wei-Ying Wang is accepted in its present form
by the Division of Applied Mathematics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Stuart Geman, Ph.D., Advisor

Recommended to the Graduate Council

Date_____ _____

Matthew Harrison, Ph.D., Reader

Date_____ _____

Elie Bienenstock, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Andrew G. Campbell, Dean of the Graduate School

Vita

In 2000, Wei-Ying Wang started his undergraduate study at National Taiwan University, and he received his Bachelor of Arts degree in Economics in 2004. After that he entered the graduate school at National Taiwan University, and received his Master of Science degree in Mathematics in 2006. Later he joined the military until 2008. He then worked for image denoising and signal processing in Academia Sinica for two years.

Wei-Ying Wang came to the United States in 2010, and has been attending the Ph.D. program in the Division of Applied Mathematics at Brown University. He received a Master of Science degree in Applied Mathematics in 2012. This dissertation was defended on May 11th, 2017.

Acknowledgments

I would like to thank all the people who have helped me to complete my thesis.

Most of all, I want to express my deepest gratitude to Prof. Geman, who helps me on so many aspects. This thesis is only possible because of him. When he saw my compression algorithm, he related it to a classical theorem so that we have an image compression algorithm that has analytic guarantee. He also came up with the idea of summarizing data with a concave distance function, rendering a very robust estimator. There are a lot of other examples and I can't tell it all here. I am very fortunate to be his student.

I would also like to thank my other committee members, Prof. Harrison and Prof. Bienenstock, for carefully reading my dissertation and giving precious comments.

I wish to acknowledge Prof. Chii-Ruey Hwang, Prof. Ting-Li Chen, Prof. Lo-Bin Change, Dr. Hong Zhang, Dr. Jackson Looper, and Wilson Mckerrow, who are extremely helpful during my research.

I need to mention about my wife, Yi-Wen Lai, who faithfully walks all the way with me, and see me through highs and lows of my days. I am thankful to have my first son, Chris, for providing so much fun and meaning in my life to keep me going. Also, I owe my thanks to my parents in Taiwan, who have been very supportive and encouraging.

Last but not least, thank God for putting me here, experiencing this much, and carrying me out during countless stay-up night.

Contents

List of Figures

Chapter One

Overview

An important job of a statistician is to summarize data. To this end, many statistical methods were developed in the 18th century, when mathematicians like Laplace and Gaussian introduced the least square method and the least absolute value regression. These "classical" statistical tools become the foundation of modern statistics, even though people was only able to deal with a very small amount of data. Then, in the early 20th century, computers were invented and these once considered "impractical" methods finally shined. However, as computer technology advanced, it brought more challenges. For example, as data being digitized, we started to consider the most efficient way to store and transmit them. In 1940s, Claude Shannon published topics about data compression and entropy. The idea of compression is to use probabilities to "filter out" the unnecessary information in the data. Just like many statistical tools—it summarizes the data. A related field is called the "lossless compression," which compresses the data in a way that people can even reconstruct the original data from the summarized information. Another example is a field called data clustering, which emerged with the increasing amount of data. In 1950s, K-means algorithm was invented to cluster, or say, to summarize, data points into several representative points. About the same time, people started to realize that some problems cannot be solved with polynomial computational time, like the problem of minimizing the loss function of K-means. These are called NP-hard problems, where it might need another technology revolution to deal with. We look into these two old topics, data compression and data clustering, in this thesis, and gives some new ideas upon it.

In Chapter 2, we will examine theories behind lossless image compression algorithms. It is surprising that people didn't explore the theories of image compression a lot. There were many literature in information theory, but only a few of them were applied for describing information on images. In this chapter we will try to be as

self-contained as possible so that people who have some knowledge in probability theory should be able to appreciate it. We will expand existing theories for information on images, and come up with an optimal compression algorithm from it. Just like those statisticians invented the classical methods far before computers, the algorithm is very impractical in view of modern computers. However, during experiments we see that it triumphs a state of the art image compression scheme and we can further push our algorithm more to achieve the theoretical optimal compression rate of an image.

In Chapter 3, we generalize the loss function of K-means to hyperplanes, where it was defined on points. The problem is still NP-hard and the original K-means algorithm is inapt for this generalization, since it gets stuck on local optima easily. The data concerned require rotation equivariant methods to summarize it, rendering most regression methods useless. We utilize a concave distance function, called the *sparse distance*, who helps us gain huge robustness. One will see the effect of using such a distance, where it successfully summarizes the data with high amount of noise and outliers. Moreover, the sparse distance makes the solution of our loss function "*interpolating*"—an important properties to allow us to solve the minimizer. While it is still difficult to solve if there are many inherent clusters, we designed a well-suited algorithm that can approximate the exact solution very well.

Chapter Two

Entropy Rate on Images

x_1	x_2	x_3	x_4	x_5
x_6	x_7	x_8	x_9	x_{10}
x_{11}	x_{12}	y		

Figure 2.1.1: A typical context for 3×5 black-and-white image patch, where $y \in \{0, 1, \cdots, 255\}$ is the value of a target pixel, and its context value is represented as $(x_1, x_2, \cdots, x_{12}) \in \{0, 1, \cdots, 255\}^{12}$.

2.1 Introduction

After the arithmetic coding was invented, most research on lossless compression algorithms focus on building models. Most of the image compression algorithms, and most successful ones, fall into the category of *predictive context modeling*, which build a statistical model on the value of a target pixel given the surrounding pixels, or say, the context. A better prediction of target pixels from their context rendering a better compression rate. One of the hardest problem on image compression is the unmanageable number of contexts. For example, in a typical context in a small 3×5 image patches as shown in Figure 2.1.1 , where each pixel has 256 depth of intensities will result in 256^{12} of possible contexts. This size grow fast when one consider a larger sized context. Of course, most of them are not seen in images, and lots of them are similar to each other. To find suitable and representative contexts is a generally done by data clustering.

Here are some successful algorithms that use predictive context modeling. A Famous algorithm LOCO-I/Jpeg-LS[25] uses a very simple predictor with different residual models on several clusters of contexts. The advantage of it is the minimum

memory and computation resources requirement. CALIC [41] algorithm uses a simple predictor and ideas such as the error feedback to better classifying contexts. It's a successful image compression and usually has 0.5 to 1 less bits per pixel [1] than that of LOCO. Practically, most of the time CALIC is among the best when comparing different compression schemes. The reason it is not as widely-used as LOCO-I is the requirement of relatively expensive computation resources. There are some compression scheme that is comparable to the results of CALIC. For example, TMW [29] introduces multiple linear predictors and utilize the idea of blending on different class of context. Another example is Golchin *et. al.* [15] who build several linear predictors and assigns blocks of images into its best predictor. While these algorithm uses local information, there are also attempts of using global attributes of image, [35] using segmentation method to reduce the number of the contexts to area like edges and smooth textures in the image.

These algorithms achieve some success on practical end, and often time the best and the second best have marginal differences of compression rates. However, theoretical supports are rarely found. One of the algorithm that has theoretical backup is called the universal modeling, which is a variation of the Ziv-Lemple coding. The core of this algorithm is to exploit repetitions of the context inside of the source, and build an empirical distribution on that. The application of the universal modeling on image compression can be found in [26]. Due to the nature of huge amount of contexts in images, its not easy to find similar patches on sources with small size. The result is often poor and cannot compete to some of the best compression schemes, like CALIC.

[1] the average number of bits required to code an image

The motivation of this article is to get some idea about how much room was left for improvement in lossless compression. Thus, we look for the notion of "minimum coding rate" in the literature, the most related article is written by Ornstein and Weiss [34], where it extends the Shannon-McMillian-Breimen theorem on higher dimensions. It indicates, under stationary and ergodic condition, the minimum coding rate (or say, bits per pixel) for an image [2] will converge to a constant, which happens to be the entropy rate. However, it is hard to believe that all images have the same theoretical minimum bits per pixel. We often see the compression results differ from one another. For example, the image with higher amount of noise will require higher amount of bits to code it, since there is more "surprises" in the image. So what is going wrong with the theory and the real world? Apparently, the theory is valid and the stationary condition is very nature. The inconsistency, we believe, is due to the ergodic assumption. In Section 2.2 we will eventually drop the ergodic condition, and the minimum coding rate, under the stationary condition, converged[3] to a random variable depends on the image it self. Furthermore, the random variable can be characterized with an ergodic components of the underlying stationary probability.

So, we have the idea of the theoretical bound for every image; the next step to to understand how the state of the art lossless image compression algorithm performed. Specifically, we aim at predicative context modeling. In Section 2.3 we give a theoretical result for predictive context modeling method. With only stationary condition and some small assumptions, the coding rate, when the context size grows, will converge to the minimum bits per pixel for every image, as we mentioned in the previous

[2]The "image" refers to an infinite sized image. Mathematically speaking, an image is considered as a finite alphabets sequence with $\mathbb{Z}^2$ indices.

[3]The convergence is in the sense of "almost everywhere," means it converges for almost all images except those have probability 0.

paragraph. In the end of this section, we will mention about an idea to utilize the theorem we got, to build an algorithm that can be "push to the limit," i.e. achieve the minimum coding rate.

The algorithm we built to test the idea is described in Section 2.4, where we call it the comparison-based image compression algorithm (CBIC). The key idea is to use an empirical distribution from external source. With the property that a empirical distribution will converge to the true one with larger amount external library[4] and larger context, we can have the proposed algorithm converged to the true entropy, or say, the minimum bits per pixel. Through experiments we see that it is able to defeat most of the state of the art algorithms, as we would expect. Note that CBIC is not a practical compression algorithm, since it requires lots of memory usage and expensive computation resources. However, the ability to scale up and provably optimal make it possible to examine the best algorithm we have so far. We think that our result suggests that there is very little room left for compression algorithms, since we only do a tiny bit better than theirs and the fact that the gain becomes marginal when we scale up the context size.

This article is arranged as follows: Section 2.2 aims on the theory behind the minimum coding rate of an image. In the end of the section, we proved that, under the stationary condition, the minimum coding rate of an image converge to the entropy rate of the ergodic component with respect to the image. Section 2.3 gives theoretical

[4]The use of the external library for compression can be found in the lossy compression schemes. Algorithms like VQ, the vector quantization [30], codes blocks of images into indices of library. Their work includes reducing the redundancy and dependency of the library so as to minimize the size of the library to gain a better compression ratio. However, in our algorithm, the redundancy of the library is one of the important elements, which forms a nice statistical property for empirical distributions.

background of image compression algorithms with predictive context modeling. Section 2.4 introduce CBIC, which is able to scale up the context size so as to achieve the optimal coding rate. It helps examine the performance of some existing algorithms like JPEG-LS and CALIC.

2.2 The Entropy of Images

In this section, we would like to give a theoretical background of the optimal compression rate. There are abundant theories on the entropy rate on 1D sequence space, see [17, 16]. To extend them to images, i.e. 2D lattice, is rarely seen in the literature. Although they have many similarities, it is not clear that we can directly extend the 1D result. We aimed to fill this void and give a better understanding of the theory behind image compression algorithms. In this section, most of theories will be extended from that of 1D, follows the idea of Gray's book [16, 17].

Before we describe the theoretical result, we will give several definitions. We first define a probability space on $\mathbb{Z}^2$-lattice, say $(\Omega, \mathcal{F}, P)$. Let $\mathcal{X}$ be the finite space of pixel values, for example, $\mathcal{X} = \{0, 1, \cdots, 255\}$, which are possible intensities in a pixel. Define $\Omega \triangleq \mathcal{X}^{\mathbb{Z}^2}$ and $\mathcal{F} \triangleq \sigma(\{\chi^{\Lambda_n^\ell}\}_{n\in\mathbb{N}\cup 0, \ell\in\mathbb{Z}^2})$, a σ-algebra generated by generating set $\mathcal{C} \triangleq \{\chi^{\Lambda_n^\ell}\}_{n\in\mathbb{N}\cup 0, \ell\in\mathbb{Z}^2}$, where

$$\Lambda_n \triangleq \{(i,j) \in \mathbb{Z}^2 | -n \le i, j \le n\} \text{ and}$$
$$\Lambda_n^\ell \triangleq \Lambda_n + \ell,$$

which is an $(2n+1) \times (2n+1)$ square lattices of $\mathbb{Z}^2$ centered at ℓ, for $\ell \in \mathbb{Z}^2$. Λ_n is shown in Figure 2.2.1. Note that $\Lambda_n \subset \Lambda_{n+1}$ with $\cup_n \Lambda_n = \mathbb{Z}^2$. We then define P to be a probability measure on $(\Omega, \mathcal{F})$. For any $\omega \in \Omega$, define the the *transformation*, $T^\ell : \Omega \to \Omega$ for $\ell \in \mathbb{Z}^2$ as

$$(T^\ell \omega)_i = \omega_{\ell+i}$$

for $\omega \in \Omega$, $\ell, i \in \mathbb{Z}^2$. Also, we define $X_0(\omega) = X(\omega) \triangleq \omega_0$, $X_\ell(\omega) \triangleq X(T^\ell\omega) = \omega_\ell$, and

$X_A \triangleq \omega_A$ for any $\ell \in \mathbb{Z}^2$ and $A \subset \mathbb{Z}^2$. We call that $(\Omega, \mathcal{F}, P, \mathcal{T})$ a dynamical system, where $\mathcal{T} = \{T^\ell\}_{\ell \in \mathbb{Z}^2}$.

We will use the notation

$$(< b) \triangleq \{a \in \mathbb{Z}^2 | a < b\}$$

for $b \in \mathbb{Z}^2$, where "$<$" is the lexicographic order on $\mathbb{Z}^2$. That is, $a < b$ if and only if either $a_1 < b_1$ or $a_1 = b_1, a_2 < b_2$.

The entropy of a random variable Y is defined as $H_P(Y) \triangleq E_P(-\log P(Y))$, and the *entropy rate* is defined as:

$$h_P \triangleq \lim_{n\to\infty} \frac{1}{|\Lambda_n|} H_P(X_{\Lambda_n}) \tag{2.2.1}$$

with $|A|$ being the number of elements in A for $A \subset \Omega$. Here we have $|\Lambda_n| = (2n+1)^2$. We sometimes omit subscript of h and H if the underlying probability is obvious. The logarithms throughout the article are based on 2.

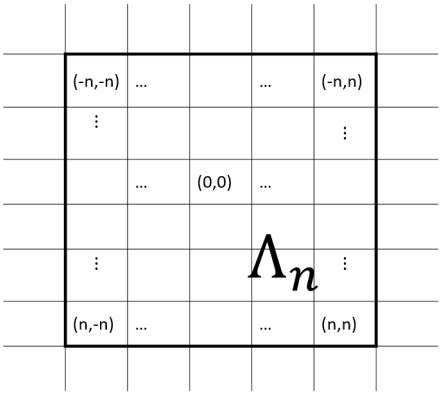

Figure 2.2.1: Λ_n on $\mathbb{Z}^2$ plane, note that the coordinator we use is different from Cartesian coordinator system.

Another similar definition is the ideal coding rate, which describes the *coding rate* when image size Λ_n is large:

$$\lim_{n\to\infty} \frac{-\log P(X_{\Lambda_n}(\omega))}{|\Lambda_n|}, \tag{2.2.2}$$

given $\omega \in \Omega$. Note that we haven't established whether the entropy rate or the coding rate are existed yet. The necessary condition for their existence will be proved in Subsection 2.2.1.

Subsection 2.2.1 describes the theorem saying that the coding rate is actually the entropy rate under stationary and ergodic conditions. Subsection 2.2.2 depicts the entropy rate under stationary condition, in particular, we can view the entropy rate as the conditional entropy given infinitely many history, which will be described in Corollary 2.2.4. Subsection 2.2.3 gives a solid background (which is hard to find in literature) of the ergodic decomposition for 2D lattice, which is essential after we drop the ergodic condition. Subsection 2.2.4 defines the Markov approximation on 2D, and discusses some related properties of it. The Markov approximation, P^m for $m \in \mathbb{N}$, will have the following property

$$P^m(X_{\Lambda_n}) = \prod_{\ell\in\Lambda_n} P(X_\ell | X_{B_m^\ell \cap \Lambda_n})$$

for $B_m^\ell \subset \Omega$, is the m^{th} order history for location $\ell \in \mathbb{Z}^2$, depicted in Figure 2.2.2. The Markov approximation is an important idea in practical compression scheme where it uses local information approaching the global one. Note that the coding rate (Equation 2.2.2) is defined on the global information. Subsection 2.2.5 proves the the Theorem 2.2.34, which generalizes Shannon-McMillian-Breimen theorem on

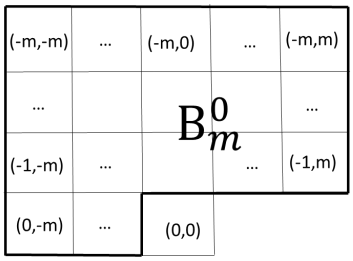

Figure 2.2.2: B_m^0 on $\mathbb{Z}^2$ plane.

2D non-ergodic source. Briefly speaking, given an $\omega \in \Omega$ the coding rate with probability P will be the entropy rate of P's ergodic components with respect to ω, which corresponds to the lowest possible compression rate of an image.

2.2.1 Coding Rate on a Stationary and Ergodic Source

First let's introduce the following assumptions:

1. (Stationary condition) A dynamical system $(\Omega', \mathcal{F}', \mu, \mathcal{S})$ is stationary if the underlying probability measure, μ, is stationary. i.e. $\mu(F) = \mu(S^{-1}F)$ for $F \in \mathcal{F}'$ and $S \in \mathcal{S}$.

2. (Ergodic condition) A dynamical system $(\Omega', \mathcal{F}', \mu, \mathcal{S})$ is ergodic if $\mu(A) \in \{0, 1\}$ for $A \in \mathcal{I}_{\mathcal{S}}$, the class of $\mathcal{S}$-invariant sets[5]. i.e. $\mathcal{I}_{\mathcal{S}} = \{A \in \mathcal{F} | S^{-1}A = A\}$ for $S \in \mathcal{S}$.

Remark. The reader might wondering why we use S^{-1} in the definition instead of simply S. This is because, for example, in space $(\Omega', \mathcal{F}')$, given $F \in \mathcal{F}'$ and S is a transformation, $SF = \{\omega : \omega \in SF\} = \{\omega : S^{-1}\omega \in F\}$, and $S^{-1}\omega$ might not

[5]The class of $\mathcal{S}$ invariance sets is actually a σ-algebra. See Ch. 6 of [13].

be well-defined when $\Omega' = \chi^{\{0,1,2,\cdots\}}$ and S is the left shift: $(S\omega)_i = \omega_{i+1}$. However, if $\Omega' = \chi^{\mathbb{Z}}$ then $S^{-1}\omega$ is well-defined and the stationary condition can be written as $P(F) = P(SF)$ in this case. Of course, in our setting $\Omega = \chi^{\mathbb{Z}^2}$ we can use $P(F) = P(TF)$ for all $T \in \mathcal{T}$ and $F \in \mathcal{F}$ as the definition of P being stationary without trouble.

Note that we sometimes say a sequence of random variable is stationary or ergodic means the underlying dynamical system is.

If the dynamical system is 1D (i.e. when $\Omega' = \mathcal{X}^{\mathbb{Z}}$), and has these two conditions, both entropy rate and coding rate are existed and they agree to each other. This result is well-known in 1D setup, which was originally proved by Breiman [6]. Fortunately, with similar arguments, we can establish them in 2D (i.e. $\Omega = \mathcal{X}^{\mathbb{Z}^2}$ as our original definition of Ω), which will be described in the rest of this subsection.

For the existence of coding rate in 2D, an important result is done by Ornstein and Weiss [34] and Lindenstrauss [24], who proved the Shannon-McMillan-Breiman theorem (SMB) for d-dimensional lattice for stationary ergodic sequences (another more comprehensive and easier to understand reference is [37]). More exactly, if we assume $\{X_{\mathbf{i}}\}_{\mathbf{i}\in\mathbb{Z}^2}$ is an ergodic and stationary sequence on $(\Omega, \mathcal{F}, P, \mathcal{T})$, then for almost every ω, we have the coding rate

$$\lim_{n\to\infty} \frac{-\log P(X_{\Lambda_n})}{|\Lambda_n|} \text{ exist and constant a.e.} \tag{2.2.3}$$

In other words,

$$\lim_{n\to\infty} -\frac{1}{(2n+1)^2}\sum_{\ell\in\Lambda_n}\log P(X_\ell|X_{(<\ell)\cap\Lambda_n}) \text{ exist and constant a.e.} \tag{2.2.4}$$

One can see that the coding rate is actually the compression rate for infinite-sized image using the correct underlying probability, which is assumed to be stationary and ergodic. In other words, the coding rate is the best compression rate we can achieve for an infinite-sized image from a stationary ergodic source. Note that we will later show that the convergence is actually the limit value of the entropy rate, h_P. This conclusion can be shown with uniformly integrability. The next lemma, which is modified from Billingsley [3] who proved it in 1D, demonstrates this in our 2D setup.

Lemma 2.2.1. *Let* $g_n = \frac{-\log P(X_{\Lambda_n})}{|\Lambda_n|}$*, then* $\{g_n\}_{n\in\mathbb{N}}$ *is uniformly integrable.*

Proof. Let $a_n \triangleq |\Lambda_n|$. Given $\omega \in \Omega$, if $g_n(\omega) \in [k, k+1)$, we have $P(\omega_{\Lambda_n}) \in (2^{(-k-1)a_n}, 2^{-ka_n}]$, so

$$\begin{aligned}
\int_{g_n\in[k,k+1)} g_n dP &\leq (k+1)P(g_n \in [k,k+1)) \\
&\leq (k+1)P(\omega : P(\omega_{\Lambda_n}) \in (2^{(-k-1)a_n}, 2^{-ka_n}]) \\
&\leq (k+1)2^{-ka_n}\cdot|\chi|^{a_n} \\
&\leq (k+1)2^{-ka_n+a_n\log|\chi|} \\
&= (k+1)2^{-a_n(k-\log|\chi|)}.
\end{aligned}$$

Thus

$$\begin{aligned}\int_{g_n>k} g_n dP &= \sum_{i=1}^{\infty} \int_{g_n\in[k+i,k+i+1)} g_n dP \\ &\leq \sum_{i=1}^{\infty} (k+i+1) e^{-(k+i-\log|\chi|)a_n},\end{aligned}$$

which converges to 0 as $k \to \infty$. □

This leads to the following corollary.

Corollary 2.2.2. *Under the stationary and ergodic conditions, the entropy rate h_P,exist, and is the same as the coding rate a.e., i.e.*

$$\lim_{n\to\infty} \frac{1}{|\Lambda_n|} H_P(X_{\Lambda_n}) = \lim_{n\to\infty} \frac{-\log P(X_{\Lambda_n})}{|\Lambda_n|}$$

a.e.

Proof. Let g_n be defined as in Lemma 2.2.1. Since we have g_n converges a.e. to a constant, say h (Equation 2.2.3) and $g_n \in L^1(P)$ is uniformly integrable, the theory about uniform integrability tell us:

$$\begin{aligned} h &= E(h) \\ &= E(\lim_{n\to\infty} g_n) \\ &= \lim_{n\to\infty} E(g_n) \\ &= \lim_{n\to\infty} \frac{1}{|\Lambda_n|} E(-\log P(X_{\Lambda_n})) \\ &= \lim_{n\to\infty} H(X_{\Lambda_n}) \\ &= h_P. \end{aligned}$$

□

2.2.2 Entropy Rate on a Stationary Source

Here we want to show that if the dynamic system is stationary, $h \triangleq \lim_{n\to\infty} \frac{1}{|\Lambda_n|} H(X_{\Lambda_n})$ exists and equals to $H(X_0|X_{(<0)})$, where

$$H(X_0|X_{(<0)}) \triangleq \lim_{n\to\infty} H(X_0|X_{(<0)\cap\Lambda_n}).$$

Note that it always exists since $\{H(X_0|X_{(<0)\cap\Lambda_n})\}_{n\in\mathbb{N}}$ is a decreasing non-negative sequence in $\mathbb{R}$. We assume that $\{X_i\}_{i\in\mathbb{Z}^2}$ is a stationary but not necessary a ergodic sequence.

First, note that we can calculate $H(X_{\Lambda_n})$ as follows:

$$H(X_{\Lambda_n}) \;=\; \sum_{\ell\in\Lambda_n} H(X_\ell|X_{(<\ell)\cap\Lambda_n}).$$

The following lemma prepare the way to prove $h_P = H(X_0|X_{(<0)})$.

Lemma 2.2.3. *Under stationary condition, we have* $H(X_\ell|X_{(<\ell)\cap\Lambda_n}) \downarrow H(X_0|X_{(<0)})$ *as* $n \to \infty$.

Proof. Since$\{H(X_0|X_{(<0)\cap\Lambda_n})\}_{n\in\mathbb{N}}$ is a decreasing non-negative sequence in $\mathbb{R}$, and

$$\begin{aligned} H(X_\ell|X_{(<\ell)\cap\Lambda_n}) &\downarrow H(X_\ell|X_{(<\ell)}) \\ &= H(X_0|X_{(<0)}) \end{aligned}$$

for all $\ell \in \mathbb{Z}^2$, where the last equation is because of the stationary condition. $\square$

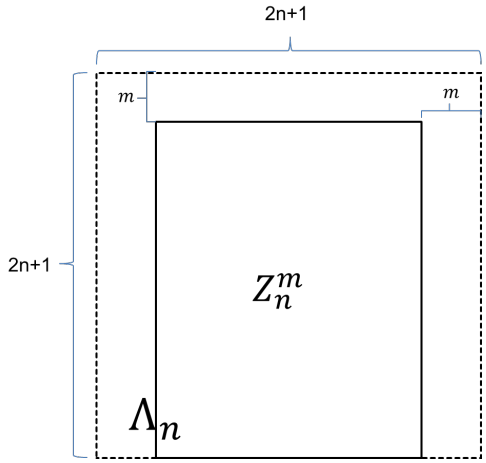

Figure 2.2.3: The plot of Z_n^m on $\mathbb{Z}^2$.

Corollary 2.2.4. *Under the stationary assumption, the entropy rate, defined as*

$$\lim_{n\to\infty} \frac{1}{|\Lambda_n|} H(X_{\Lambda_n})$$

exists and equals $H(X_0|X_{(<0)})$.

Proof. Let $\hat{h} \triangleq H(X_0|X_{(<0)})$. First note that it is easy to see that $\lim_{n\to\infty} \frac{1}{|\Lambda_n|} H(X_{\Lambda_n}) \geq \hat{h}$, since for all $n \in \mathbb{N}$:

$$\begin{aligned}
\frac{1}{|\Lambda_n|} H(X_{\Lambda_n}) &= \frac{1}{|\Lambda_n|} \sum_{\ell\in\Lambda_n} H(X_\ell|X_{(<\ell)\cap\Lambda_n)}) \\
&\geq \frac{1}{|\Lambda_n|} \sum_{\ell\in\Lambda_n} H(X_\ell|X_{(<\ell)}) \\
&\overset{\text{stationary}}{=} \frac{1}{|\Lambda_n|} \sum_{\ell\in\Lambda_n} H(X_0|X_{(<0)}) \\
&= \hat{h}.
\end{aligned}$$

So the rest is to prove $\lim_{n\to\infty} \frac{1}{|\Lambda_n|} H(X_{\Lambda_n}) \leq \hat{h}$. Because of Lemma 2.2.3, given an

arbitrary $\epsilon > 0$, there is an $M \in \mathbb{N}$ s.t. for all $m > M$, $H(X_0|X_{(<0)\cap\Lambda_m}) - \hat{h} < \epsilon$. Set $Z_n^m \triangleq \{\ell \in \Lambda_n | B_m^\ell \subset \Lambda_n\}$, where $B_m^\ell = B_m^0 + \ell$, $B_m^0 = (<0) \cap \Lambda_m$ for $\ell \in \mathbb{Z}^2$ (Figure 2.2.2 and Figure 2.2.3 depict B_m^0 and Z_n^m, respectively). Calculate

$$\begin{aligned}
\frac{1}{|\Lambda_n|}H(X_{\Lambda_n}) &= \frac{1}{|\Lambda_n|}\sum_{\ell\in\Lambda_n} H(X_\ell|X_{(<\ell)\cap\Lambda_n)}) \\
&= \frac{1}{|\Lambda_n|}[\sum_{\ell\in Z_n^m} H(X_\ell|X_{(<\ell)\cap\Lambda_n)}) + \sum_{\ell\in\Lambda_n\setminus Z_n^m} H(X_\ell|X_{(<\ell)\cap\Lambda_n)})] \\
&\leq \frac{1}{|\Lambda_n|}[\sum_{\ell\in Z_n^m} H(X_\ell|X_{B_m^\ell}) + \sum_{\ell\in\Lambda_n\setminus Z_n^m} H(X_\ell|X_{(<\ell)\cap\Lambda_n)})] \\
&= \frac{1}{|\Lambda_n|}[\sum_{\ell\in Z_n^m} H(X_0|X_{B_m^0}) + \sum_{\ell\in\Lambda_n\setminus Z_n^m} H(X_\ell|X_{(<\ell)\cap\Lambda_n)})]
\end{aligned}$$

because of stationary condition. Note there $|Z_n^m| = (2n+1-m)(2n+1-2m)$ and all the term in the summation are uniformly bounded. So

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{|\Lambda_n|}H(X_{\Lambda_n}) &\leq H(X_0|X_{B_m^0}) \\
&< \hat{h} + \epsilon.
\end{aligned}$$

Since ϵ is arbitrary, we have $\lim_{n\to\infty}\frac{1}{|\Lambda_n|}H(X_{\Lambda_n}) \leq \hat{h}$. □

Thus, if we allow ergodic assumption, we can combine this with Corollary 2.2.2 to get:

Corollary 2.2.5. *Under the stationary and ergodic assumptions, the entropy rate, h, exist and can be view as* $\lim_{n\to\infty}\frac{1}{|\Lambda_n|}H(X_{\Lambda_n})$, $H(X_0|X_{(<0)})$ *,or* $\lim_{n\to\infty}\frac{-\log P(X_{\Lambda_n})}{|\Lambda_n|}$.

Remark. Note that this is the Shannon-McMillian-Breiman theorem on 2D-lattice version.

2.2.3 Ergodic Theory on Images

When looking at Corollary 2.2.5, it is nature to ask a question: What happens if one drops the ergodic assumption? To answer this, one has to understand the ergodic decomposition of a stationary probability. Briefly speaking, the entropy rate $\lim_{n\to\infty}\frac{1}{|\Lambda_n|}H(X_{\Lambda_n})$ or $H(X_0|X_{(<0)})$ will still converge, due to Corollary 2.2.4, but the convergence will be the mixture of entropy of ergodic components. We will see that in Theorem 2.2.20. And the coding rate $\lim_{n\to\infty}\frac{-\log P(X_{\Lambda_n}(\omega))}{|\Lambda_n|}$ will still converge a.s., but the convergence will be a random variable depending on which ergodic component the ω belongs to. This part will be proved in Theorem 2.2.34. Before proving these theorems, one needs to make sure that some famous ergodic theorems for 1D sequences is valid for generalizing to 2D images. Subsection 2.2.3.1 introduces Birkhoff ergodic theorem on 2D, which says

$$\lim_{n\to\infty}\frac{1}{|\Lambda_n|}\sum_{\ell\in\Lambda_n}f(T^\ell\omega)=E_P(f|\mathcal{I})\ P\text{-a.e},$$

for $\mathcal{I}=\mathcal{I}_{\mathcal{T}}$ is the $\mathcal{T}$-invariant σ-algebra[6]. In Subsection 2.2.3.2, we generalize the ergodic decomposition theorem to 2D, which basically says that every stationary distribution can be decomposed as mixture of ergodic stationary distributions. In the end of this section, Subsection 2.2.3.3 states that the entropy rate can be thought to be a mixture of entropies of ergodic components, which is relatively easy to generalized from 1D setup.

[6]Remember that $\mathcal{T}=\{T^\ell\}_{\ell\in\mathbb{Z}^2}$.

2.2.3.1 Birkhoff Ergodic Theorem on Images

Let's first introduce the pointwise ergodic theorem on $\mathbb{Z}^2$ lattice. It is well-known that in 1D sequence: For $f \in L^1(\mu)$ and μ is a stationary probability on 1D sequence with respect to a transformation S, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(S^i\omega) = E_\mu(f|\mathcal{I}_S) \ \mu - a.e.,$$

where $\mathcal{I}_S$ is an S-invariant σ-algebra (see [13]). A similar statement is valid in 2D, which can be found in [31]: for any $f \in L^1(P)$,

$$\lim_{n\to\infty} \frac{1}{|\Lambda_n|} \sum_{\ell\in\Lambda_n} f(T^\ell\omega) = E_P(f|\mathcal{I}) \ P\text{-a.e}, \tag{2.2.5}$$

where $\mathcal{I} = \mathcal{I}_\mathcal{T}$ is the $\mathcal{T}$-invariant σ-algebra. i.e. the class of $\mathcal{T}$ invariant events, $\mathcal{I} \triangleq \{B \in \mathcal{F}|T^{-\ell}B = B\}$ for any $\ell \in \mathbb{Z}^2$. However, the explanation of this statement is hard to find (in [31], it is quoted from [33], where it might need more mathematical background to see it). The reader can assume this is right and proceed to the next subsection. For those who are interested about this, we would like to give a complete explanation to it. In the following content we assume a stationary dynamical system $(\Omega, \mathcal{F}, P, \mathcal{T})$ as we had on $\mathbb{Z}^2$ lattice defined in the beginning Section 2.2.

First let's introduce the *special averaging sequence* defined in [32].

Definition 2.2.6. (Special averaging sequence) $\{A_n\}_{n\in\mathbb{N}}$ is a special averaging sequence on group $G = (\Omega', \circ)$ if:

(a) $\lim |(gA_n)\triangle A_n|/|A_n| = 0$ for all $g \in G$,

(b) $A_1 \subset A_2 \subset \cdots$, and

(c) $|A_n^{-1} A_n| \leq M \cdot |A_n|$ for some constant M and for all n.

The symbol "$\triangle$" is the symmetric difference of two sets and $AB \triangleq \{a \circ b : a \in A, b \in B\}$. Apparently, $\{\Lambda_n\}_{n\in\mathbb{N}}$ is a special averaging sequence on group $(\mathbb{Z}^2, +)$. Another example is $\{B_m\}_{m\in\mathbb{N}}$, defined as $\Lambda_m \cap (< 0)$, (we sometimes denoted as B_m^0) which is shown in Figure 2.2.2. Note that the sequence that just run on one direction, e.g.. $\{(0, i) \in \mathbb{Z}^2 : i \leq n\}_{n\in\mathbb{N}}$, is not a special averaging sequence on $(\mathbb{Z}^2, +)$ as it fails on (a) with $g = (1, 0)$. With the fact that $(\mathbb{Z}^2, +)$ is an amenable group, we have the following theorem as shown in [32]. (Note that we assume P is a stationary distribution.)

Theorem 2.2.7. *If $\{A_n\}$ is a special averaging sequence on $(\mathbb{Z}^2, +)$, then for any $f \in L^1(P)$,*

$$\lim_{n\to\infty} \frac{1}{|A_n|} \sum_{\ell\in A_n} f(T^\ell \omega) \quad = \quad f^\star(\omega)$$

exist for P-a.e. $\omega \in \Omega$ and is a $\mathcal{T}$-invariant function, i.e. for any $T \in \mathcal{T}$, $f^\star(T\omega) = f^\star(\omega)$ for P-a.e. $\omega \in \Omega$.

We would like to characterize $f^\star$ by saying $f^\star = E(f|\mathcal{I})$, where $\mathcal{I}$ is the σ-field of $\mathcal{T}$-invariant events. We will prove it with uniform integrable property in Theorem 2.2.10. Let's first prepares it with following two lemmas.

Lemma 2.2.8. *For any $f \in L^1(P)$ and any transformation $S = T^\ell$, $\ell \in \mathbb{Z}^2$, we have*

$$E(f(S)) \quad = \quad E(f)$$

Proof. We know that if f is an indicator function on Ω, say 1_A, $A \in \mathcal{F}$, we have

$$\begin{aligned} E(1_A(S)) &= P(S^{-1}A) \\ &= P(A) = E(1_A(\omega)). \end{aligned}$$

So the equation is true for indicator function. Thus, the result is true when f is a simple function (finite linear combination of indicator functions).

Assume $f \geq 0$. Since there is a sequence of simple functions $\{f_n\}_{n\in\mathbb{N}}$, s.t. $f_n \uparrow f$ a.e. as $n \to \infty$, we have

$$\begin{aligned} \int f(S\omega) &= \int \lim_{n\to\infty} f_n(S\omega) \\ &\underset{=}{\text{MCT}} \lim_{n\to\infty} \int f_n(S\omega). \end{aligned}$$

(MCT stands for the monotone converge theorem.) Thus, the equation is true for $f \geq 0$. We can now finish the proof by writing $f = f^+ - f^-$, where f^+ and f^+ are both non-negative functions. □

Lemma 2.2.9. *If $\{A_n\}$ is a special averaging sequence on $\mathbb{Z}^2$, then for any $f \in L^1(\Omega, \mathcal{F}, P)$, the sequence $\{g_n\}_{n\in\mathbb{N}}$ defined as*

$$g_n(\omega) \triangleq \frac{1}{|A_n|} \sum_{\ell\in A_n} f(T^\ell\omega)$$

is uniformly integrable.

Proof. First, set $\int |f| dP = M < \infty$, then calculate

$$\begin{aligned} \int |g_n| dP &\leq \frac{1}{|A_n|} \sum_{\ell \in A_n} \int |f(T^\ell \omega)| dP(\omega) \\ &= \frac{1}{|A_n|} \sum_{\ell \in A_n} \int |f(\omega)| dP(\omega) \\ &= M. \end{aligned}$$

Thus g_n is uniformly L^1-bounded.

Second, for any $E \in \mathcal{F}$, calculate

$$\begin{aligned} \int_E |g_n| &\leq \frac{1}{|A_n|} \sum_{\ell \in A_n} \int_E |f(T^\ell \omega)| \\ &= \frac{1}{|A_n|} \sum_{\ell \in A_n} \int_{T^{-\ell} E} |f(\omega)|. \end{aligned}$$

We know that since $f \in L^1$, for any $\epsilon > 0$, there is a $\delta > 0$, s.t. for any event $E' \in \mathcal{F}$ with $P(E') < \delta$, we have

$$\int_{E'} |f| < \epsilon.$$

Note that because of stationary, $P(E) = P(T^{-\ell} E)$ for any $\ell \in A_n$. Thus, for any $\epsilon > 0$, there is a $\delta > 0$, s.t. for any event $E \in \mathcal{F}$ with $P(E) < \delta$, we have

$$\int_E |g_n| < \epsilon,$$

and is independent of n. □

Now we have prepared to characterize the convergence, $f^\star$, in Theorem 2.2.7.

Theorem 2.2.10. *For $f \in L^1(P)$, we have $f^\star(\omega) = E(f|\mathcal{I})(\omega)$ P-a.e, where $\mathcal{I}$ is the σ-field of $\mathcal{T}$-invariant events.*

Proof. We need to show that:

(a) $f^\star \in \mathcal{I}$.

(b) For any $B \in \mathcal{I}$, $\int_B f^\star = \int_B f$.

Note that (a) is satisfied because $f^\star$ is $\mathcal{T}$-invariant function. As for the property (b), given any $B \in \mathcal{I}$, calculate

$$\begin{aligned} E(f^\star 1_B) &= E(\lim_{n\to\infty} \frac{1}{|A_n|} \sum_{\ell \in A_n} f(T^\ell) 1_B) \\ &\underset{=}{\text{u.i.}} \lim_{n\to\infty} E(\frac{1}{|A_n|} \sum_{\ell \in A_n} f(T^\ell) 1_B) \\ &= \lim_{n\to\infty} \frac{1}{|A_n|} \sum_{\ell \in A_n} E(f(T^\ell) 1_B). \end{aligned}$$

(u.i means uniformly integrable.) And calculate

$$\begin{aligned} E(f(T^\ell) 1_B) &= \int_B f(T^\ell \omega) dP(\omega) \\ &= \int_{T^{-\ell} B} f(\omega) dP(\omega) \\ &\underset{=}{B \in \mathcal{I}} \int_B f(\omega) dP(\omega) \\ &= E(f 1_B). \end{aligned}$$

So, we have proved (b). □

Thus we have proved Equation 2.2.5 on more general averaging sequence case. Note that $E(f|\mathcal{I})$ doesn't depend on the choice of the averaging set$\{A_n\}$, means that the limit of $\frac{1}{|A_n|}\sum_{\ell\in A_n} f(T^\ell)$ is the same for any averaging sequence $\{A_n\}$.

2.2.3.2 Ergodic Decomposition on Images

The main purpose of this subsection is to provide a theoretical background on ergodic decomposition for $\mathbb{Z}^2$ lattice. Again, we assume our dynamic system $(\Omega, \mathcal{F}, T, P)$ on $\mathbb{Z}^2$ is stationary. To have a little understand of the ergodic decomposition theorem, the most simple form of the theorem is

$$P(F) = \int P_\delta(F) dQ_\psi(\delta)$$

for $F \in \mathcal{F}$, where P_δ is a (stationary and ergodic) measure on $(\Omega, \mathcal{F})$, $Q_\psi(\cdot) = P(\psi^{-1}(\cdot))$ is a measure on the measures of $(\Omega, \mathcal{F})$, and ψ is the ergodic component function that assigns $\omega \in \Omega$ to its ergodic component ψ_ω, a stationary ergodic probability on $(\Omega, \mathcal{F})$. As one can see, the stationary probability P is *decomposed* to ergodic components $\{P_\delta\}_{\delta\in\psi(\Omega)}$. Furthermore, the probability distribution $\psi_\omega(\cdot)$ can be characterized as $P(\cdot|\psi)(\omega)$. Let's give an simple example to describe this:

Example. Set $\Omega' = \{\omega_{01}, \omega_{10}, \omega_{001}, \omega_{010}, \omega_{100}\}$, where

$$\omega_{01} = \{0, 1, 0, 1, \cdots\};$$
$$\omega_{10} = \{1, 0, 1, 0, \cdots\};$$
$$\omega_{001} = \{0, 0, 1, 0, 0, 1, \cdots\};$$
$$\omega_{010} = \{0, 0, 1, 0, 0, 1, \cdots\};$$
$$\omega_{100} = \{1, 0, 0, 1, 0, 0, \cdots\}.$$

Define $\mathcal{F}' = 2^{\Omega'}$ and $\mu(\omega) \triangleq \frac{1}{2}P_1(\omega) + \frac{1}{2}P_2(\omega)$, where

$$P_1(\omega) = \frac{1}{2}\mathbb{I}_{\{\omega_{01},\omega_{10}\}}(\omega), \text{ and}$$
$$P_2(\omega) = \frac{1}{3}\mathbb{I}_{\{\omega_{001},\omega_{010},\omega_{001}\}}(\omega),$$

where $\mathbb{I}$ is an indicator function. Define the Transformation S as $(S\omega)_i = \omega_{i+1}$.

In this example we have a non-ergodic stationary dynamical system $(\Omega', \mathcal{F}', \mu, S)$. One can verify it is not ergodic by the following fact: Set $A \triangleq \{\omega_{10}, \omega_{01}\}$, then $A \in \mathcal{I}_S$, the T'-invariant σ-algebra., and we have $\mu(A) = \frac{1}{2} \notin \{0, 1\}$. On the other hand, $(\Omega', \mathcal{F}', \mu, S)$ is an ergodic and stationary system for either $i = 1, 2$. In this case, μ can be decomposed to P_1, P_2, the ergodic components. One can also see the ergodic component function ψ_μ for μ can be defined as $\psi_\mu(\omega) = P_1$ if $\omega \in \{\omega_{01}, \omega_{10}\}$ and $\psi_\mu(\omega) = P_2$ if $\omega \in \{\omega_{001}, \omega_{010}, \omega_{001}\}$.

The verification of the ergodic decomposition theorem for 2D is similar to that in 1D. However, it is hard to find literature for this. To prove the ergodic decomposition theorem on images, we will follow the idea of Ch.8 of Gray's book [16], where

everything is proved for 1D and *asymptotic mean stationary* (AMS) case, which is a weaker form of stationary property. Note that as we study through Gray's book, there are some proofs that we find difficult to understand, so we will try to complete them as much as we can.

Before introducing the ergodic decomposition, let's first introduce the concept of *standard measurable space.* Some terminology will be used: First, a *field* under a sample space Ω' is a collection of subsets of Ω' that (1) has Ω' in it, (2) is closed under complement, and (3) is closed under finite unions[7]. Second, a field *generated* by a collection of subsets, say A, means it is the smallest field contains A. Last, a *countable generating field* means the field is generated by countable subsets.

Definition 2.2.11. A measurable space $(\Omega', \mathcal{F}')$ is *standard* if it has the *countable extension property.* That is, there is a countable generating field, $\mathcal{G}'$ s.t. $\sigma(\mathcal{G}') = \mathcal{F}'$ and any set function μ on $\mathcal{G}$ can be uniquely extended to a probability measure on $\mathcal{F}'$ if μ has the following properties:

(a) (*normalization*) $\mu(\Omega') = 1$,

(b) (*nonnegativity*) $\mu(G) \geq 0$ for all $G \in \mathcal{G}'$, and

(c) (*finite additivity*) $\mu(\cup_{i=1}^{n} G_i) = \sum_{i=1}^{n} \mu(G_i)$ for $G_i \in \mathcal{G}'$, $G_i \cap G_j = \phi$ with $i \neq j = 1, 2, \cdots, n$.

There are many different definitions of the standard measure space, see Ch.2 of [16]. We define it like this because it is ready to use. In our setting, as we have seen

[7]Note that σ-algebra is closed under countable unions

in the beginning of this section, the measurable space $(\Omega, \mathcal{F})$ has a nature countable generating set $\mathcal{C} \triangleq \{\chi^{\Lambda_n^\ell}\}_{n\in\mathbb{N}\cup 0, \ell\in\mathbb{Z}^2}$. Thus if $(\Omega, \mathcal{F})$ is a standard space then we can define a set function on the field that generated by $\mathcal{C}$, say $\mathcal{G}$, then extend to $\mathcal{F}$. The verification of $(\Omega, \mathcal{F})$ being a standard space involves the Carathéodory extension theorem and the Kolmogorov extension theorem, which can be found in Ch2 of [16]. We will simply put the theorem here.

Theorem 2.2.12. $(\Omega, \mathcal{F})$ *is a standard space.*

In the following, we will go through the formation of the *ergodic component*s. Let's define the set of the *ergodic points*

$$\mathcal{E} \triangleq \cap_{G\in\mathcal{G}}\{\omega : \lim_{n\to\infty} \frac{1}{|A_n|} \sum_{\ell\in A_n} 1_G(T^\ell\omega) \text{ exists}\}. \tag{2.2.6}$$

Note that $\mathcal{G}$ is countable, so we know that $\mathcal{E}$ is a measurable (because $1_G(T^\ell\omega)$ is measurable function for any $\ell \in \mathbb{Z}^2$) and $P(\mathcal{E}) = 1$. Consider a set function for $G \in \mathcal{G}$

$$\mu_\omega(G) \triangleq \begin{cases} E_P(1_G|\mathcal{I})(\omega) & , \ \omega \in \mathcal{E} \\ \eta(G) & , \ \omega \in \mathcal{E}^c \end{cases}$$

for every $\omega \in \Omega$ and some stationary ergodic distributions η on $(\Omega, \mathcal{F})$. Since μ_ω is normalized, nonnegative, finitely additive set function on $\mathcal{G}$, it can be extend to a probability measure on $(\Omega, \mathcal{F})$ (by Definition 2.2.11). So we can now write $\mu_\omega(F) = E_{\mu_\omega}(1_F)$ for any $F \in \mathcal{F}$ and $\omega \in \Omega$. Also, it is easy to verify that μ_ω is a stationary ergodic probability distribution on $(\Omega, \mathcal{F})$. In fact,$\{\mu_\omega\}_{\omega\in\Omega}$ is referred as the e*rgodic decomposition* of the stationary distribution P.

It will be worth mentioning that it will be difficult to claim measurability if we define $\mathcal{E}$ as the intersection of sets on $\mathcal{F}$, which is an uncountable intersection. Also, the reader might be tempted (like the author did!) to define $\mu_\omega(G) = E_P(1_G|\mathcal{I})(\omega)$ whenever $\lim_{n\to\infty} \frac{1}{|A_n|} \sum_{\ell\in A_n} 1_G(T^\ell\omega)$ converges for every $G \in \mathcal{G}$, and $\mu_\omega(G) = \eta(G)$ otherwise. However, it will be hard to claim the finite additivity of the set function, since there might be an ω (although in the measure 0 set) that the limit exists for set F_1 but not set F_2.

An important property of μ_ω is that it is invariant.

Lemma 2.2.13. *μ_ω, as a function of $\omega \in \Omega$, is $\mathcal{T}$-invariant. i.e. $\mu_\omega(F) = \mu_{T^k\omega}(F)$ for all $\omega \in \Omega$, $F \in \mathcal{F}$, and $k \in \mathbb{Z}$.*

Proof. Because of unique extendability of the standard space, we only need to consider events in $\mathcal{G}$. For $\omega \in \mathcal{E}$ and $G \in \mathcal{G}$, we first need to show that $T^k\omega \in \mathcal{E}$, for any $k \in \mathbb{Z}^2$. Define $Z_n \triangleq \frac{1}{|A_n|} \sum_{\ell\in A_n} 1_G(T^\ell\omega)$ and $Z \triangleq \lim_{n\to\infty} Z_n(\omega)$, then calculate for any $k \in \mathbb{Z}^2$:

$$|Z_nT^k - Z| \;\leq\; |Z_nT^k - Z_n| + |Z_n - Z|.$$

The limit of the second term is 0. And the first term

$$\begin{aligned} |Z_n T^k - Z_n| &= |\frac{1}{|A_n|}\sum_{\ell\in A_n} 1_G(T^\ell T^k\omega) - \frac{1}{|A_n|}\sum_{\ell\in A_n} 1_G(T^\ell\omega)| \\ &= \frac{1}{|A_n|}|\sum_{\ell\in T^{-k}A_n} 1_G(T^\ell\omega) - \sum_{\ell\in A_n} 1_G(T^\ell\omega)| \\ &\leq \frac{1}{|A_n|}\sum_{\ell\in T^{-k}A_n\triangle A_n} 1_G(T^\ell\omega) \\ &\leq \frac{|T^{-k}A_n\triangle A_n|}{|A_n|} \\ &\to 0 \end{aligned}$$

as $n\to\infty$ from the definition of special averaging set. Thus $|Z_nT^k - Z|\to 0$. So we showed that if $\omega\in\mathcal{E}$, then $T^k\omega\in\mathcal{E}$, and moreover,

$$\mu_{T\omega}(G) = \lim_{n\to\infty} Z_nT^k = Z = \lim_{n\to\infty}\frac{1}{|A_n|}\sum_{\ell\in A_n} 1_G(T^\ell\omega) = \mu_\omega(G).$$

In the last, for those $\omega\notin\mathcal{E}$, it is easy to imply that $T\omega\notin\mathcal{E}$ for any $T=T^k$, $k\in\mathbb{Z}^2$. So $\mu_\omega=\eta=\mu_{T\omega}$.

□

With the definition of μ_ω, we can further characterize Theorem 2.2.10 by the following Lemma.

Lemma 2.2.14. *For $f\in L^1(P)$, we have*

$$E(f|\mathcal{I})(\omega) = E_{\mu_\omega}(f)$$

for P-a.e. ω.

Proof. We only need to consider $\omega \in \mathcal{E}$. To prove the equation, it is easy to see that the result is true if f is a simple function. Next, suppose $f \geq 0$, then we know there is a sequence of simple function $\{f_k\}$, s.t. $f_k \uparrow f$ P-a.e., Thus

$$\begin{aligned} E(f|\mathcal{I})(\omega) &= E(\lim_k f_k|\mathcal{I})(\omega) \\ &\underset{=}{MCT} \lim_k E(f_k|\mathcal{I}) \\ &= \lim_k E_{\mu_\omega}(f_k) \\ &\underset{=}{\text{MCT}} E_{\mu_\omega}(f). \end{aligned}$$

Thus the equality is true if $f \geq 0$ and is integrable. For integrable f, we can then write $f = f^+ - f^-$ to complete the proof. □

Next, we introduce the *ergodic component function*. Let $\mathcal{P} = \mathcal{P}((\Omega, \mathcal{F}))$ be the space of probabilities on $(\Omega, \mathcal{F})$. And the ergodic component function is defined as $\psi : \Omega \to \mathcal{P}$ with

$$\psi(\omega) = \psi_\omega \triangleq \mu_\omega,$$

which assigns ω to its own stationary ergodic distribution. It will be convenient to have the following characterization: $E_{\mu_\omega}(f) = E_P(f|\psi(\omega))$ for $f \in L^1(P)$, which will require a σ-algebra on $\mathcal{P}$. Also we would like to define a measure $Q_\psi(\cdot) = P(\psi^{-1}(\cdot))$ to measure the size of the partition of ω that have the same ergodic component, which means we need to have a measure space for $\mathcal{P}$. The σ-algebra for the measure space can be induced by defining a metric on $\mathcal{P}$, where we use the *distributional distance* here:

$$d_{\mathcal{G}}(\mu, \nu) \triangleq \sum_{i=1}^{\infty} 2^{-i} |\mu(G_i) - \nu(G_i)|,$$

where $\{G_i\}_{i=1}^{\infty} = \mathcal{G}$. (Remember that $\mathcal{G}$ is the field countably generated by $\mathcal{C} = \{\chi^{\Lambda_n^\ell}\}_{n\in\mathbb{N}\cup 0,\ell\in\mathbb{Z}^2}$.) With this metric we can define the σ-algebra, say $\mathcal{L}$, to be the Borel σ-algebra:

$$\mathcal{L} \triangleq \sigma(\{\mu : d_{\mathcal{G}}(\mu,\nu) < r\}_{v\in\mathcal{P},r\geq 0}).$$

Thus we have the following lemma, which is from Lemma 8.5.1 of [16].

Lemma 2.2.15. *There is a measurable space $(\mathcal{P},\mathcal{L})$ s.t. the map $\psi : \Omega \to \mathcal{P}$ defined as $\psi(\omega) \triangleq \mu_\omega$ is measurable.*

Proof. We need to show that $\psi^{-1}(A) \in \mathcal{F}$ for any $A \in \mathcal{L}$. Define $\mathcal{S} \triangleq \{A \in \mathcal{L} : \psi^{-1}(A) \in \mathcal{F}\}$ to be the collection of sets in $\mathcal{L}$ that make ψ measurable. It's easy to see that $\mathcal{S}$ is a σ-algebra because ψ^{-1} preserve unions and complements (a general property for a mapping). We first show that for any $\nu \in \mathcal{P}$, and $r \geq 0$, we have $B_r(v) \in \mathcal{S}$, where $B_r(v) \triangleq \{\mu \in \mathcal{P} : d_{\mathcal{G}}(\mu,\nu) < r\}$. i.e. $\psi^{-1}(B_r(v)) \in \mathcal{F}$.

To see this, one can write

$$\begin{aligned}
\psi^{-1}(B_r(v)) &= \{\omega \in \Omega : \psi(\omega) \in B_r(v)\} \\
&= \{\omega : d_{\mathcal{G}}(\psi(\omega),\nu) < r\} \\
&= \{\omega : \sum_{i=1}^{\infty} 2^{-i}|\mu_\omega(G_i) - \nu(G_i)| < r\}.
\end{aligned}$$

Then we have $\psi^{-1}(B_r(v)) \in \mathcal{F}$ because $\mu_\omega(G_i) \in \{E(1_{G_i}|\mathcal{I})(\omega), \eta(G_i)\}$, as a function of ω, is measurable. So $B_r(v) \in \mathcal{S}$, which means $\sigma(B_r(v)) \subset \mathcal{S}$ and we have $\psi^{-1}(A) \in \mathcal{F}$ for any $A \in \mathcal{L}$. □

Since ψ actually maps Ω into stationary and ergodic distributions, say $\mathcal{P}_e$, a subspace of $\mathcal{P}$. We can further restrict the measure space $(\mathcal{P}, \mathcal{L})$ to $(\mathcal{P}_e, \mathcal{L}_e)$, where $\mathcal{L}_e$ is the Borel σ-algebra of $\mathcal{P}_e$. However, it will require that $\mathcal{P}_e$ is a measurable set in $(\mathcal{P}, \mathcal{L})$, since we would like ψ to be measurable in $\mathcal{P}_e$, too. The following lemma (slightly modified from Lemma 6.7.4 in [16], where it was proved in 1D and AMS case) is necessary for us to demonstrate this fact in Theorem 2.2.18.

Lemma 2.2.16. *A stationary dynamical system $(\Omega, \mathcal{F}, Q, T)$ is ergodic if and only if*

$$\lim_{n\to\infty} \frac{1}{|A_n|} \sum_{\ell \in A_n} Q(T^{-\ell} G \cap G) \quad = \quad Q^2(G) \tag{2.2.7}$$

for any $G \in \mathcal{G}$, and a special averaging sequence $\{A_n\}$.

Note that $\mathcal{G}$ is the (countable) generating field defined as the smallest field contains $\mathcal{C}$.

Proof. Note that for any $f \in L^1$, $E_Q(f|\mathcal{I}) = E_Q(f)$ if Q is ergodic, since $Q(A) \in \{0, 1\}$ for any $A \in \mathcal{I}$ from the definition of ergodicity.

1. ($\Rightarrow$) For any $G \in \mathcal{G}$, calculate

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{|A_n|}\sum_{\ell\in A_n}Q(T^{-\ell}G\cap G) &= \lim_{n\to\infty}\frac{1}{|A_n|}\sum_{\ell\in A_n}E_Q(1_{T^{-\ell}G\cap G}) \\
&= \lim_{n\to\infty}\frac{1}{|A_n|}\sum_{\ell\in A_n}\int_G 1_{T^{-\ell}G}(\omega)dQ(\omega) \\
&\overset{\text{u.i.}}{=} \int_G \lim_{n\to\infty}\frac{1}{|A_n|}\sum_{\ell\in A_n}1_{T^{-\ell}G}(\omega)dQ(\omega) \\
&\overset{\text{ergodic}}{=} \int_G E(1_G)dQ(\omega) \\
&= Q^2(G).
\end{aligned}$$

Note that the uniformly integrability (u.i.) is from Lemma 2.2.1, which requires Q to be a stationary measure.

2. ($\Leftarrow$) First, understand that for any $A\in\mathcal{I}$, $\lim_{n\to\infty}\frac{1}{|A_n|}\sum_{\ell\in A_n}Q(T^{-\ell}A\cap A) = Q(A)$. So we have $Q(A)=Q^2(A)$, and this implies $Q(A)\in\{0,1\}$, which proves the system is ergodic. The hard part is that the Equation 2.2.7 is valid only for events on the generating field $\mathcal{G}$, and $\mathcal{I}$ is not necessary contained in $\mathcal{G}$. So we have to argue that Equation 2.2.7 holds for any event $F\in\mathcal{F}$. We know that for any $\epsilon\in(0,1)$ and $F\in\mathcal{F}$ there is an $F_0\in\mathcal{G}$ s.t. $Q(F\triangle F_0)<\epsilon$ (see Corollary 1.5.3 in [16]). Set

$$(\star) \;=\; |\frac{1}{|A_n|}\sum_{\ell\in A_n}Q(T^{-\ell}F\cap G)-Q^2(F)|,$$

and from the triangular inequality we can calculate

$$\begin{aligned}(\star) \quad \leq \quad & |\frac{1}{|A_n|}\sum_{\ell\in A_n} Q(T^{-\ell}F\cap F) - \frac{1}{|A_n|}\sum_{\ell\in A_n} Q(T^{-\ell}F_0\cap F_0)| \\ & +|\frac{1}{|A_n|}\sum_{\ell\in A_n} Q(T^{-\ell}F_0\cap F_0) - Q^2(F_0)| \\ & +|Q^2(F_0) - Q^2(F)|.\end{aligned}$$

The second term will converge to 0 when $n \to \infty$ by the condition we have. The third term will be less than ϵ^2, because $|Q(C) - Q(D)| \leq Q(C\triangle D)$ for any event C and D. And the first term will be less than

$$\frac{1}{|A_n|}\sum_{\ell\in A_n} |Q(T^{-\ell}F\cap F) - Q(T^{-\ell}F_0\cap F_0)|.$$

Note that we have

$$\begin{aligned}|Q(T^{-\ell}F\cap F) - Q(T^{-\ell}F_0\cap F_0)| &\leq \frac{1}{2}Q(T^{-\ell}F\cap F\triangle T^{-\ell}F_0\cap F_0) \\ &= (\diamond)\end{aligned}$$

We can do a further estimate of $(\diamond)$ by $Q(C\triangle D) \leq Q(C\triangle E) + Q(D\triangle E)$, so

$$\begin{aligned}(\diamond) \quad \leq \quad & Q(T^{-\ell}F\cap F\triangle T^{-\ell}F_0\cap F) + Q(T^{-\ell}F_0\cap F_0\triangle T^{-\ell}F_0\cap F) \\ \leq \quad & Q(T^{-\ell}F\triangle T^{-\ell}F_0) + Q(F_0\triangle F) \\ = \quad & Q(T^{-\ell}(F\triangle F_0)) + Q(F_0\triangle F) \\ = \quad & 2Q(F_0\triangle F).\end{aligned}$$

Thus the first term is bounded by 2ϵ, this implies

$$\lim_{n\to\infty} |\frac{1}{|A_n|} \sum_{\ell\in A_n} Q(T^{-\ell}F \cap G) - Q^2(F)| \leq 2\epsilon.$$

We can then let $\epsilon \to 0$ and concludes that Equation 2.2.7 holds for any event $F \in \mathcal{F}$.

□

We are now prepared to claim that $\mathcal{P}_e$ is a measurable set. Define $\mathcal{P}_s$ to be the collection of stationary probability distributions in $\mathcal{P}$. The next lemma is modified from Lemma 8.5.2 in [16].

Lemma 2.2.17. *We have*

(a) *$\mathcal{P}_s$, the collection of stationary distributions in $\mathcal{P}$, is a close set in $(\mathcal{P}, \mathcal{L})$, and*

(b) *$\mathcal{P}_e$, the collection of stationary ergodic distributions in $\mathcal{P}$, is a measurable set in $(\mathcal{P}, \mathcal{L})$.*

Proof.

(a) Given any sequence $\{m_n\}_{n\in\mathbb{N}} \subset \mathcal{P}_s$, s.t. $m_n(F) \to m(F)$ as $n \to \infty$ for any $F \in \mathcal{F}$. Given any $G \in \mathcal{G}$, from the definition of $\mathcal{G}$ ($\mathcal{G}$ is a field generated by $\mathcal{C}$) we know that $T^\ell G \in \mathcal{G} \subset \mathcal{F}$, for any $\ell \in \mathbb{Z}^2$. So $m(G) = m(T^\ell G)$ for $G \in \mathcal{G}$ since

$$m(G) = \lim_{n\to\infty} m_n(T^\ell G) = \lim_{n\to\infty} m_n(G) = m(T^\ell G).$$

By the uniqueness of the measure extension property (i.e. Definition 2.2.11 on standard space $(\otimes, \mathcal{F})$), we conclude that $m(F) = m(T^{\ell}F)$ for any $F \in \mathcal{F}$, i.e. $m \in \mathcal{P}_s$. Hence it is closed.

(b) From Lemma 2.2.16, we know that

$$\mathcal{P}_e \;=\; \mathcal{P}_s \cap \{m \in \mathcal{P} : \lim_{n\to\infty} \frac{1}{|A_n|} \sum_{\ell \in A_n} m(T^{-\ell}G \cap G) = m^2(G) \text{ for any } G \in \mathcal{G}\}$$

(note that $\mathcal{P}_e \subset \mathcal{P}_s$ by definition). From (a) we know that $\mathcal{P}_s$ is closed, hence measurable. So we only need to show the second set is measurable. Furthermore, since the second set equals to

$$\cap_{G\in\mathcal{G}}\{m : \lim_{n\to\infty} \frac{1}{|A_n|} \sum_{\ell \in A_n} m(T^{-\ell}G \cap G) = m^2(G)\}$$

and the fact that $\mathcal{G}$ is countable and $T^{-\ell}G \in \mathcal{G}$, we only need to show that for any $G \in \mathcal{G}$, $m(G)$, as a function of m, is a measurable function on $(\mathcal{P}, \mathcal{L})$. We can show this by proving $C_G \triangleq \{m' \in \mathcal{P} : m'(G) < r\}$ is an open set on the metric space the $(\mathcal{P}, d_{\mathcal{G}})$ for any $r > 0$. The proof follows: Given any $m' \in C_G$, there is an $\epsilon > 0$ s.t. $m'(G) < r - \epsilon$. Given $\delta < 2^{-k}\epsilon$, for any $v \in \{v' \in \mathcal{P} : d_{\mathcal{G}}(m', v') < \delta\}$, we have

$$\sum_{G_i \in \mathcal{G}} 2^{-i} |m'(G_i) - v(G_i)| < \delta,$$

which implies $|m'(G) - v(G)| < 2^k\delta$ for some $k \in \mathbb{N}$. If $v(G) < m'(G)$ then $v \in C_G$. If $v(G) > m'(G)$, we have

$$v(G) < m'(G) + 2^k\delta < r - \epsilon + 2^k\delta < r.$$

Thus we have proved that for any $m' \in C_G$, there is a δ s.t. for any $v \in \{v' \in \mathcal{P} : d_{\mathcal{G}}(m', v') < \delta\}$, $v \in C_G$, i.e. C_G is open. This implies $C_G \in \mathcal{L}$. Hence $m(G)$ is a measurable function on $(\mathcal{P}, \mathcal{L})$ for any $G \in \mathcal{G}$.

□

We have prepared ourselves all the background that needed for the ergodic decomposition theorem on images, i.e. $(\Omega, \mathcal{F})$, with respect to ergodic component ψ. The following is modified from Theorem 8.5.1 in [16].

Theorem 2.2.18. *(Ergodic Decomposition) Under $(\mathcal{P}_e, \mathcal{L}_e)$, the measurable space defined on stationary and ergodic distributions on $(\Omega, \mathcal{F})$, a function $\psi : \Omega \to \mathcal{P}_e$ is defined as*

$$\psi(\omega)(F) \triangleq \psi_\omega(F) \triangleq \begin{cases} E(1_F|\mathcal{I})(\omega) & , \omega \in \mathcal{E} \\ \eta(F) & , \omega \in \mathcal{E}^c \end{cases}$$

for any $F \in \mathcal{F}$ and for some *$\eta \in \mathcal{P}_e$, where $\mathcal{E}$ is defined in Equation 2.2.6 with $P(\mathcal{E}) = 1$. Then we have the following properties:*

(a) *ψ is measurable and $\mathcal{T}$-invariant, i.e. for any $T = T^k$, $k \in \mathbb{Z}^2$, $\psi(T\omega) = \psi(\omega)$, or say $\psi_{T\omega}(F) = \psi_\omega(F)$ for any $F \in \mathcal{F}$.*

(b) *On $(\mathcal{P}_e, \mathcal{L}_e)$, there is a well-defined probability distribution $Q_\psi(C) = P(\psi^{-1}(C))$ for $C \in \mathcal{L}_e$. And for any $f \in L^1(P)$, we have*

$$E_P f = \int_{\omega \in \Omega} E_{\psi_\omega} f dP(\omega) = \int E_{P_\delta}(f) dQ_\psi(\delta),$$

which implies $P(F) = \int_{\omega \in \Omega} \psi_\omega(F) dP(\omega) = \int P_\delta(F) dQ_\psi(\delta)$ for $F \in \mathcal{F}$.

(c) *For any $f \in L^1(P)$, we have $E_\psi(f) = E_P(f|\psi)$, which implies $\psi(F) = P(F|\psi)$ for $F \in \mathcal{F}$.*

Note that we wrote $P_\delta = \delta$ for $\delta \in \mathcal{P}_e$ in this theorem.

Proof. We will prove these statements on generating fields $\mathcal{G}$, since we can extend the result to $\mathcal{F}$ because $(\Omega, \mathcal{F})$ is standard space.

(a) (Measurabilility) From Lemma 2.2.15, we know that for any $A \in \mathcal{L}$ we have $\psi^{-1}(A) \in \mathcal{F}$. Since $\mathcal{P}_e$ is a measurable subset of $\mathcal{P}$ from Lemma 2.2.17, any open set in $\mathcal{P}_e$ is measurable in $(\mathcal{P}, \mathcal{L})$. This implies $\mathcal{L}_e \subset \mathcal{L}$. (Note that both of them are Borel σ-algebra.) So given any $A \in \mathcal{L}_e$, we have $A \in \mathcal{L}$. Thus we can conclude ψ is measurable.

(Invariance) The invariance simply follows from Lemma 2.2.13.

(b) (Well-definedness) First, by (a) ψ is measurable, so $P(\psi^{-1}(C))$ has meaning for any $C \in \mathcal{L}_e$. It remains to show that it is a probability measure. Obviously it is non-negative. Next let's check if it have normalization property: $Q_\psi(\mathcal{P}_e) = P(\psi^{-1}(\mathcal{P}_e)) = 1$ since for every ω, $\psi(\omega) \in \mathcal{P}_e$. The countable additivity can be easily proved because ψ^{-1} preserves unions, and complements. Thus we conclude Q_ψ is a well-defined probability measure on $(\mathcal{P}_e, \mathcal{L}_e)$.

(First equality) For any $f \in L^1(P)$ calculate

$$\begin{aligned} E_P(f) &= E_P(E_P(f|\mathcal{I})) \\ &\overset{\text{Lem } 2.2.14}{=} E_P(E_{\psi_\omega}(f)) \\ &= \int_{\omega\in\Omega} E_{\psi_\omega}(f) dP(\omega). \end{aligned}$$

(Second equality) Calculate

$$\begin{aligned}
\int_{\omega\in\Omega} E_{\psi_\omega}(f)dP(\omega) &= \int_{\omega\in\Omega}\int_{v\in\Omega} f(v)d\psi_\omega(v)dP(\omega) \\
&\underset{\delta=\psi(\omega)}{=} \int_{\delta\in\psi(\Omega)}\int_{v\in\Omega} f(v)dP_\delta(v)dP(\psi^{-1}(\delta)) \\
&\underset{(\star)}{=} \int_{\delta\in\mathcal{P}_e}\int_{v\in\Omega} f(v)dP_\delta(v)dP(\psi^{-1}(\delta)) \\
&= \int E_{P_\delta}(f)dQ_\psi(\delta),
\end{aligned}$$

where in $(\star)$, we used $Q_\psi(\mathcal{P}_e) = Q_\psi(\psi(\Omega)) = 1$, so $Q_\psi(\mathcal{P}_e\backslash\psi(\Omega)) = 0$.

(c) We are going to show: (1) $E_\psi(f) \in \sigma(\psi)$, and (2) $\int_B E_\psi(f) = \int_B f$ for $B \in \sigma(\psi)$.

(1) For any $f \in L^1$, it suffices to show that a function $q_f : \mathcal{P}_e \to \mathbb{R}$ defined as $q(\mu) \triangleq E_\mu(f)$ is a continuous function. Since for any $a \in \mathbb{R}$ if q_f is continuous then $q_f^{-1}((-\infty, a))$ is an open set in $(\mathcal{P}_e, \mathcal{L}_e)$, thus

$$\begin{aligned}
\{\omega : E_{\psi_\omega} f < a\} &= \psi^{-1}(\{\mu \in \mathcal{P}_e : E_\mu(f) < a\}) \\
&= \psi^{-1}q_f^{-1}((-\infty, a)) \\
&\in \psi^{-1}(\mathcal{L}_e) \\
&\triangleq \sigma(\psi)
\end{aligned}$$

Let's first show that when $f = 1_G$, $G \in \mathcal{G}$, the generating field, q_{1_G} is continuous. So, we have $q(\mu) = \mu(G)$ for $\mu \in \mathcal{P}_e$. Set $\delta > 0$, for any $\mu,\ \nu \in \mathcal{P}_e$ with $d_{\mathcal{G}}(\mu, v) < \delta$. Thus, by definition of $d_{\mathcal{G}}$, we have for $\mathcal{G} = \{G_i\}_{i=1}^n$

$$\sum_{i=1}^{\infty} 2^{-i}|\mu(G_i) - \nu(G_i)| < \delta,$$

which implies

$$|\mu(G) - \nu(G)| < 2^k \delta$$

for some $k \in \mathbb{N}$. So, for any $\mu \in \mathcal{P}_e$ and for any $\epsilon > 0$, we can choose $\delta < 2^{-k}\epsilon$, s.t. for any $\nu \in \mathcal{P}_e$ with $d_{\mathcal{G}}(\mu, v) < \delta$, we have

$$|q_{1_G}(\mu) - q_{1_G}(\nu)| = |\mu(G) - \nu(G)| < \epsilon,$$

i.e. q_{1_G} is continuous for $G \in \mathcal{G}$.

Next, to see that q_{1_F} is continuous for $F \in \mathcal{F}$, define

$$\mathcal{S} \triangleq \{F \in \mathcal{F} : q_{1_F} \text{ is continuous.}\}.$$

It is easy to see that $\mathcal{S}$ is a σ-algebra. (Since, for any $\mu \in \mathcal{P}_e$, first, given any $A \in \mathcal{S}$, $q_{1_{A^c}}(\mu) = 1 - \mu(A) = 1 - q_{1_A}(\mu) \in \mathcal{S}$; second, given any $\{A_i\}_{i=1}^{\infty} \subset \mathcal{S}$ that is pairwise disjointed, we have $q_{1_{\cup A_i}}(\mu) = \sum_i \mu(A_i) = \sum_i q_{1_{A_i}} \in \mathcal{S}$.) And, by previous statement, $G \in \mathcal{S}$ for $G \in \mathcal{G}$, so $\mathcal{F} = \sigma(\mathcal{G}) \subset \mathcal{S}$. So we have proved q_{1_F} is continuous for $F \in \mathcal{F}$.

Thus it is easy to see that q_{1_f} is continuous if f is any simple function. For $f \geq 0$, we can then define $\{f_k\}$ to be the sequence of simple functions s.t. $f_k \uparrow f$ P-a.e., and using monotone convergence theorem to conclude q_{1_f} is continuous for $f \geq 0$. In the last, for general $f \in L^1$ we have q_{1_f} is continuous since we can write $f = f^+ - f^-$.

(2) Since ψ is invariant, we have $\psi^{-1}(A) \in \mathcal{I}$, for any $A \in \mathcal{L}_e$. This implies

$\sigma(\psi) \subset \mathcal{I}$. So for any $B \in \sigma(\psi)$, we have $B \in \mathcal{I}$, and

$$\begin{aligned} \int_B f(\omega)dP(\omega) &= E_P(E_P(f1_B|\mathcal{I})) \\ &\underset{B \in \mathcal{I}}{=} E_P(1_B E_P(f|\mathcal{I})) \\ &\underset{\text{Lem } 2.2.14}{=} \int_B E_{\mu_\omega}(f)dP(\omega) \\ &= \int_B E_{\psi_\omega}(f)dP(\omega). \end{aligned}$$

So we have the result that $E_\psi(f) = E_P(f|\psi)$.

□

The theorem gives many useful conclusions. First, a stationary probability distribution can be viewed as weighted sum of stationary ergodic distributions, where the weights is indexed by elements in $\mathcal{P}_e$ and not by that in Ω. Second, We can divide Ω into several *ergodic components*, the ergodic components about ω can be defined as the set $\psi^{-1}(\psi(\omega))$ which tells you about which ergodic stationary distribution that ω belong to, with size $Q_\psi(\psi(\omega))$. Third, ψ is actually the probability of P condition on the ergodic component ψ. Also note that ψ and $\mathcal{P}_e$ doesn't depend on averaging sequences, so we can say that ψ and $\mathcal{P}_e$ are *induced* by the stationary dynamical system $(\Omega, \mathcal{F}, P, \mathcal{T})$.

Let's write down the useful corollary we got so far, for a future use.

Corollary 2.2.19. *Assume $(\Omega, \mathcal{F}, P, \mathcal{T})$ we defined on 2D-lattice is stationary with induced ergodic component function ψ. For any $f \in L^1(P)$ and ant special averaging*

sequence $\{A_n\}$

$$\lim_{n\to\infty} \frac{1}{|A_n|} \sum_{\ell \in A_n} f(T^\ell \omega) \quad = \quad E(f|\mathcal{I})(\omega) = E_\psi(f)$$

for P*-a.e.* $\omega \in \Omega$.

2.2.3.3 Ergodic Decomposition of the Entropy Rate

In this subsection, we will introduce one of the important results regarding the ergodic decomposition. Since, from Theorem 2.2.18, we saw that a stationary P can be viewed as the combination of stationary ergodic distributions $\{P_\delta\}_{\delta \in \mathcal{P}_e}$, a nature question is that: Can we view the entropy rate h_P as a combination of $\{h_{P_\delta}\}_{\delta \in \mathcal{P}_e}$? The answer is yes. The following fundamental theorem helps us understand the relationship between entropy rate and ergodic decomposition, which is due to Jacobs [21], and is well-described in Section 2.4 in [17]. We will not give a complete proof here, since most of them is the same as in 1D setting. Instead, we will briefly describe the proof and give a complementary argument for properties that are not clearly extendable to 2D-lattice.

Theorem 2.2.20. *(The Ergodic Decomposition of the Entropy Rate) We have*

$$h_P \quad = \quad \int_{\omega \in \Omega} h_{\psi_\omega} dP(\omega)$$

To prove this, the reader might be tempted to use property (b) on Theorem 2.2.18 and Lebesgue dominate convergence theorem to exchange integral and limit, however, since $\frac{1}{|\Lambda_n|} H_{\psi_\omega}(X_{\Lambda_n}(\omega))$ converge almost surely in ψ (not in P), we cannot apply the

theorem here. It's actually a non-trivial result. The basic idea is based on Theorem 8.9.1 of [16], which says if a function $D : \mathcal{P}_s \to \mathbb{R} \geq 0$, where $\mathcal{P}_s$ is the set of stationary distributions on $(\Omega, \mathcal{F})$, satisfies

(a) $D(\psi_\omega) \in L^1(P)$ for any $\omega \in \Omega$,

(b) $D(\lambda\mu_1 + (1-\lambda)\mu_2) = \lambda D\mu_1 + (1-\lambda)D(\mu_2)$ for $\lambda \in (0,1)$, $\mu_1, \mu_2 \in \mathcal{P}_s$, and

(c) D is upper semi-continuous, then

$$D(P) = \int D(\psi_\omega) dP(\omega).$$

So we can prove Theorem 2.2.20 by choosing $D(P) = h_P \triangleq \lim_{n\to\infty} \frac{1}{|\Lambda_n|} H_P(X_{\Lambda_n})$. Note that D satisfies (b) (this is from Lemma 2.3.4 of [17]): First, note that we have

$$\begin{aligned} I = H_{\lambda\mu_1+(1-\lambda)\mu_2}(X) &= -\lambda \sum_x \mu_1(x) \log(\lambda\mu_1(x) + (1-\lambda)\mu_2(x)) \\ &\quad -(1-\lambda) \sum_x \mu_2(x) \log(\lambda\mu_1(x) + (1-\lambda)\mu_2(x)) \\ &= \lambda E_{\mu_1}(-\log(\lambda\mu_1(X) + (1-\lambda)\mu_2(X))) \\ &\quad +(1-\lambda) E_{\mu_2}(-\log(\lambda\mu_1(X) + (1-\lambda)\mu_2(X))). \end{aligned}$$

So the the expectation in the first term of I can be estimated:

$$\begin{aligned}
&E_{\mu_1}(-\log(\lambda\mu_1(X)+(1-\lambda)\mu_2(X)))\\
=&E_{\mu_1}(-\log(\mu_1(X))-\log(\lambda+(1-\lambda)\frac{\mu_2(X)}{\mu_1(X)}))\\
=&H_{\mu_1}(X)+E_{\mu_1}(-\log(\lambda+(1-\lambda)\frac{\mu_2(X)}{\mu_1(X)}))\\
\overset{(\star)}{\geq}&H_{\mu_1}(X)+E_{\mu_1}(1-\lambda-(1-\lambda)\frac{\mu_2(X)}{\mu_1(X)}))\\
=&H_{\mu_1}(X)+1-\lambda-(1-\lambda)\\
=&H_{\mu_1}(X)
\end{aligned}$$

Note that at $(\star)$ we use the inequality$-\log x \geq 1-x$. With the same reason we have the expectation in the second term of I large or equal to $H_{\mu_2}(X)$. Thus we have the lower bound of I as

$$I \;\geq\; \lambda H_{\mu_1}(X)+(1-\lambda)H_{\mu_2}(X).$$

Second, the upper bound of I can be estimated with

$$\begin{aligned}
I \;\leq\;& \lambda E_{\mu_1}(-\log(\lambda\mu_1(x))\\
&(1-\lambda)E_{\mu_2}(-\log((1-\lambda)\mu_2(x))\\
=\;& \lambda H_{\mu_1}(X)+(1-\lambda)H_{\mu_2}(X)+g(\lambda),
\end{aligned}$$

where $g(\lambda)=-\lambda\log(\lambda)+(1-\lambda)\log(1-\lambda)$. Thus, by dividing $|\Lambda_n|$ and let $n\to\infty$ on both bounds of I, we have the property (b).

Since most of the proof is the same as in Section 2.4 of [17], we will not repeat

here. In particular, when proving (c), the upper semi-continuity, it requires $h_P = \inf_{n\in\mathbb{N}} \frac{1}{|\Lambda_n|} H_P(X_{\Lambda_n})$, which can be proved by

$$\begin{aligned}
\frac{1}{|\Lambda_n|} H_P(X_{\Lambda_n}) &= \frac{1}{|\Lambda_n|} \sum_{\ell\in\Lambda_n} H_P(X_\ell | X_{(<\ell)\cap\Lambda_n}) \\
&\geq \frac{1}{|\Lambda_n|} \sum_{\ell\in\Lambda_n} H_P(X_\ell | X_{(<\ell)}) \\
&\underset{=}{\text{stationary}} \frac{1}{|\Lambda_n|} \sum_{\ell\in\Lambda_n} H_p(X_0 | X_{(<0)}) \\
&= H(X_0 | X_{(<0)}) \\
&\underset{=}{\text{Cor. 2.2.4}} h_P
\end{aligned}$$

for all $n \in \mathbb{N}$.

2.2.4 Markov Approximation on Images

Here we introduce the Markov approximation, for stochastic process on $\mathbb{Z}^2$. Markov approximation will play a crucial role as proving Theorem 2.2.34. Note that in 1D, the kth-order Markov approximation is defined as

$$p^k(Y_{0:n-1}) = p(Y_{0:k-1}) \prod_{i=k}^{n-1} p(Y_i | Y_{i-k:i-1})$$

(see ch16.8 of [9], where the indices are defined on $\{0, 1, \cdots, \infty\}$ and the notation $a : b$ means $\{a, a+1, \cdots, b\}$). This is an vital part when proving the original Shannon-McMillian-Breiman (SMB) theorem. And we will extend the concept to 2D-lattice. Specifically, the properties we will develop in this section will help Lemma 2.2.32, the upper bound of the non-ergodic SMB theorem in the next section. The reader

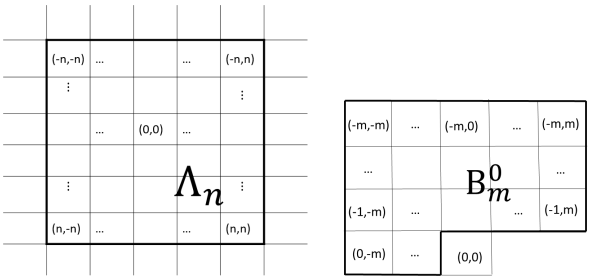

Figure 2.2.4: Example of rectangle sets.

can first proceed to the next subsection, where we will prove the non-ergodic SMB theorem on 2D, and come back to this section when one needs supports on Lemma 2.2.32.

First, we define the set that can be approximated.

Definition 2.2.21. Set $A \subset \mathbb{Z}^2$ is a *rectangle set* if for some $\ell, i \in \mathbb{Z}^2$ and $n \in \mathbb{N} \cup \{0\}$,

$$A = (< i) \cap \Lambda_n^\ell.$$

Also, we define $\mathcal{A}$ to be the collection of all possible rectangle sets.

A quick examples are Λ_n^ℓ and B_m^ℓ for $\ell \in \mathbb{Z}^2$ and $n, m \in \mathbb{N} \cup \{0\}$, where $B_m^\ell \triangleq \ell + B_m^0 \subset \mathbb{Z}^2$ and $\Lambda_n^\ell = \ell + \Lambda_n^0$. Remember that $B_m^0 = B_m \triangleq \Lambda_m^0 \cap (< 0)$ and $\Lambda_n^0 = \Lambda_n$. (as depicted as in Figure 2.2.4).

Definition 2.2.22. (Markov approximation) A set function $P^m : \mathcal{X}^{\mathcal{A}} \to \mathbb{R}$ is the mth-order Markov approximation P^m to the probability distribution P s.t.

$$P^m(x_A) \;=\; \prod_{\ell \in A} P(X_\ell = x_\ell | X_{B_m^\ell \cap A} = x_{B_m^\ell \cap A})$$

for all $A \in \mathcal{A}$. Moreover, define

$$P^m(x_A|x_B) \triangleq \begin{cases} \frac{P^m(x_{A\cup B})}{P^m(x_B)} & \text{, if } P^m(x_B) > 0 \\ 0 & \text{, o.w.} \end{cases}$$

if $A \cup B \in \mathcal{A}$ and $B \in \mathcal{A}$.

We will use the abbreviation $P^m(X_A)$ as a random variable $P^m(X_A(\omega))$. Same abbreviation rule applies to P. In the next lemma, we introduce some property of the Markov approximation.

Lemma 2.2.23. *(P^m properties) P^m has the following properties:*

1. $\sum_{x_A \in \mathcal{X}_A} P^m(x_A) = 1$ *for all* $A \in \mathcal{A}$.
2. $P^m(x_A) = P^m(x_{A+\ell})$ *for* $A \in \mathcal{A}$, $\ell \in \mathbb{Z}^2$.
3. *The conditional probability on mth-order area is the same as the original one, i.e.*

$$P^m(x_0|x_{B_m^0}) = P(x_0|x_{B_m^0}).$$

4. $P^m(x_0|x_A) = P^m(x_0|x_{B_m})$ *if for some* $k \in \mathbb{Z}^2$, $A = (< 0) \cap \Lambda_n^k \supset B_m^0$.
5. $P \ll P^m$ *on* $\mathcal{X}^{\mathcal{A}}$. *That is,* $P^m(x_A) = 0$ *implies* $P(x_A) = 0$ *for all* $A \in \mathcal{A}$.
6. $P_\delta \ll P^m$ *on* $\mathcal{X}^{\mathcal{A}}$ *for all* $P_\delta \in \mathcal{P}_e$, *where* $\mathcal{P}_e$ *is the collection of all the stationary ergodic distribution induced by the dynamical system* $(\Omega, \mathcal{F}, P, \mathcal{T})$.

Proof. All the proofs are straightforward:

1. Given any rectangle set A, for some $\ell \in \mathbb{Z}^2$, and denote $\mathcal{X}_A$ as the set of all the configuration on A. Set $\{\ell^1 > \ell^2 > \cdots > \ell^{|A|}\} = A$ Then

$$\begin{aligned}
\sum_{x_A \in \mathcal{X}_A} P^m(x_A) &= \sum_{x_A \in \mathcal{X}_A} \prod_{i=1}^{|A|} P(x_{\ell^i} | x_{B_m^{\ell^i} \cap A}) \\
&= \sum_{x_{\ell^1} \in \mathcal{X}} P(x_{\ell^1} | x_{B_m^{\ell^1} \cap A}) \sum_{x_{A \setminus \{\ell^1\}} \in \mathcal{X}_{A \setminus \{\ell^1\}}} \prod_{i=2}^{|A|} P(x_{\ell^i} | x_{B_m^{\ell^i} \cap A}) \\
&= 1 \cdot \sum_{x_{A \setminus \{\ell^1\}} \in \mathcal{X}_{A \setminus \{\ell^1\}}} \prod_{i=2}^{|A|} P(x_{\ell^i} | x_{B_m^{\ell^i} \cap A}).
\end{aligned}$$

   Then iteratively we can get $\sum_{x_A \in \mathcal{X}_A} P^m(x_A) = 1$.

2. Trivial result since P is stationary.

3. This is because

$$\begin{aligned}
P^m(x_0 | x_{B_m^0}) &= \frac{P^m(x_{B_m^0 \cup \{0\}})}{P^m(x_{B_m^0})} \\
&= \frac{P(x_0 | x_{B_m^0}) \prod_{\ell \in B_m^0} P(x_\ell | x_{B_m^\ell \cap B_m^0})}{\prod_{\ell \in B_m^0} P(x_\ell | x_{B_m^\ell \cap B_m^0})} \\
&= P(x_0 | x_{B_m^0}).
\end{aligned}$$

4. Since

$$\begin{aligned}
P^m(x_0 | x_A) &= \frac{P^m(x_{A \cup \{0\}})}{P^m(x_A)} \\
&= \frac{P(x_0 | x_{B_m^0}) \prod_{\ell \in A} P(x_\ell | x_{B_m^\ell \cap A})}{\prod_{\ell \in A} P(x_\ell | x_{B_m^\ell \cap A})} \\
&= P(x_0 | x_{B_m^0}) \\
&= P^m(x_0 | x_{B_m^0}).
\end{aligned}$$

5. Given any rectangle set $A \subset \mathbb{Z}^2$ and $x_A \in \mathcal{B}$, s.t. $P^m(x_A) = 0$, this will imply that $P(x_\ell | x_{B^\ell_m \cap A}) = 0$ for some $\ell \in A$. Thus $P(x_\ell, x_{B^\ell_m \cap A}) = 0$ for some ℓ. And $P(x_A) = P(x_A | x_\ell, x_{B^\ell_m \cap A}) P(x_\ell, x_{B^\ell_m \cap A})$.

6. Because of $P_\delta \ll P$.

□

The reader must be aware that although P^m looks like a probability distribution in a lot of ways, it is not a probability distribution, since the definition of rectangles would not admit a σ-algebra. We will just treat P^m as a special set function on rectangle sets, which is enough for the use here. The following lemma give the coding rate for Markov approximation, which is modified to 2D from Lemma 3.1.1 of [17].

Lemma 2.2.24. *For any $m \in \mathbb{N}$, there is a $\mathcal{T}$-invariant function $h^m : \Omega \to \mathbb{R}$, s.t.*

$$\lim_{n\to\infty} \frac{-1}{|\Lambda_n|} \log P^m(X_{\Lambda_n}) \quad = \quad h^m \ P\text{-a.e.}$$

Furthermore, h^m can be specified as

$$h^m(\omega) \triangleq \begin{cases} E_{\psi(\omega)}(-\log P(X_0 | X_{B_m})) & , \ \textit{if } \lim_{n\to\infty} \frac{-1}{|\Lambda_n|} \log P^m(X_{\Lambda_n})(\omega) \textit{ converges.} \\ \log |\mathcal{X}| & , \ \textit{o.w.} \end{cases}$$

Proof. Calculate

$$\begin{aligned} \frac{-1}{|\Lambda_n|} \log P^m(X_{\Lambda_n}) &= \frac{-1}{|\Lambda_n|} \sum_{\ell \in \Lambda_n} \log P(X_\ell | X_{B^\ell_m \cap \Lambda_n}) \\ &= \frac{-1}{|\Lambda_n|} \sum_{\ell \in Z^m_n} \log P(X_\ell | X_{B^\ell_m \cap \Lambda_n}) + \frac{-1}{|\Lambda_n|} \sum_{\ell \in \Lambda_n \setminus Z^m_n} \log P(X_\ell | X_{B^\ell_m \cap \Lambda_n}), \end{aligned}$$

where $Z_n^m \triangleq \{\ell \in \Lambda_n | B_m^\ell \subset \Lambda_n\}$, as depicted in Figure 2.2.3. Since $E(-\log P(X_\ell | X_{B_m^\ell \cap \Lambda_n})$ is finite (because $\mathcal{X}$ is of finite alphabet), we know that $-\log P(X_\ell | X_{B_m^\ell \cap \Lambda_n})$ is also finite a.e. Thus the second term goes to 0 when n get sufficiently larger. With the same reason, the limit won't change if we replace the second term with

$$\frac{-1}{|\Lambda_n|} \sum_{\ell \in \Lambda_n \setminus Z_n^m} \log P(X_0 | X_{B_m^0}) T^\ell(\omega)$$

which is also goes to 0 when n goes to ∞.

And the first term:

$$\frac{-1}{|\Lambda_n|} \sum_{\ell \in Z_n^m} \log P(X_\ell | X_{B_m^\ell \cap \Lambda_n})(\omega) \quad = \quad \frac{-1}{|\Lambda_n|} \sum_{\ell \in Z_n^m} \log P(X_0 | X_{B_m^0}) T^\ell(\omega)$$

because of stationary condition. Thus

$$\lim_{n\to\infty} \frac{-1}{|\Lambda_n|} \log P^m(X_{\Lambda_n}) = \lim_{n\to\infty} \frac{-1}{|\Lambda_n|} \sum_{\ell \in \Lambda_n} \log P(X_0 | X_{B_m^0}) T^\ell(\omega)$$
$$\underline{a.e.} E_{\psi_\omega}(-\log P(X_0 | X_{B_m^0}))$$

which is from the result of ergodic theorem, i.e. Corollary 2.2.19. □

Let's give the definitions of relative entropy and relative entropy rate about P^m in the following definitions, which will be used in the next lemma.

Definition 2.2.25. Given a stationary distribution P' on $(\Omega, \mathcal{F})$ such that $P' \ll P$, the relative entropy of P' and P^m on Λ_n is defined as

$$H_{P'||P^m}(X_{\Lambda_n}) \triangleq -E_{P'}(\log \frac{P^m(X_{\Lambda_n})}{P'(X_{\Lambda_n})})$$

for $n \in \mathbb{N}$. Also, the relative entropy rate of P' and P^m on Λ_n is defined as

$$h_{P'||P^m} \triangleq \lim_{n\to\infty} \frac{1}{|\Lambda_n|} H_{P'||P^m}(X_{\Lambda_n})$$

Note that the definition of relative entropy, $-E_{P'}(\log \frac{P^m(X_{\Lambda_n})}{P'(X_{\Lambda_n})})$, is well-defined because of the absolute continuous assumption (Lemma 2.2.23 (5) and (6)). Most of the time P' will be used as P or ψ. The following lemmas (Lemma 2.2.26, 2.2.28 and 2.2.29) will be used in proving SMB theorem for stationary non-ergodic source (Theorem 2.2.34). They next lemma is modified to 2D version from Lemma 2.4.3 [17].

Lemma 2.2.26. *Given a stationary distribution P' on $(\Omega, \mathcal{F})$ such that $P' \ll P$ we have*

$$h_{P'||P^m} \quad = \quad -h_{P'} + E_{P'}(-\log P(X_0|X_{B_m^0})),$$

where $h_{P'} \triangleq \lim_{n\to\infty} \frac{1}{|\Lambda_n|} H_{P'}(X_{\Lambda_n})$.

Proof. Since

$$\frac{1}{|\Lambda_n|} H_{P'||P^m}(X_{\Lambda_n}) \quad = \quad \frac{1}{|\Lambda_n|}\{-H_{P'}(X_{\Lambda_n}) - \sum_{x_{\Lambda_n} \in \mathcal{X}^{\Lambda_n}} P'(x_{\Lambda_n}) \log P^m(x_{\Lambda_n})\}.$$

The first term will be $-h_{P'}$ as $n \to \infty$. Calculate the second term:

$$\begin{aligned}
\sum_{x_{\Lambda_n}} P'(x_{\Lambda_n}) \log P^m(x_{\Lambda_n}) &= \sum_{x_{\Lambda_n}} P'(x_{\Lambda_n}) \sum_{\ell \in \Lambda_n} \log P(x_\ell | x_{B_m^\ell \cap \Lambda_n}) \\
&= \sum_{\ell \in \Lambda_n} \sum_{x_{\Lambda_n}} P'(x_{\Lambda_n}) \log P(x_\ell | x_{B_m^\ell \cap \Lambda_n}) \\
&= \sum_{\ell \in Z_n^m} \sum_{x_{\Lambda_n}} P'(x_{\Lambda_n}) \log P(x_\ell | x_{B_m^\ell \cap \Lambda_n}) \\
&\quad + \sum_{\ell \in \Lambda_n \backslash Z_n^m} \sum_{x_{\Lambda_n}} P'(x_{\Lambda_n}) \log P(x_\ell | x_{B_m^\ell \cap \Lambda_n}) \\
&= (1) + (2)
\end{aligned}$$

where $Z_n^m \triangleq \{\ell \in \Lambda_n | B_m^\ell \subset \Lambda_n\}$. Calculate term (1):

$$\begin{aligned}
(1) &= \sum_{\ell \in Z_n^m} \sum_{x_{B_m^\ell \cup \{\ell\}}} P'(x_{B_m^\ell \cup \{\ell\}}) \log P(x_\ell | x_{B_m^\ell \cap \Lambda_n}) \\
&= \sum_{\ell \in Z_n^m} \sum_{x_{B_m^\ell \cup \{\ell\}}} P'(x_{B_m^\ell \cup \{\ell\}}) \log P(x_\ell | x_{B_m^\ell}) \\
&\overset{\text{stationary}}{=} \sum_{\ell \in Z_n^m} \sum_{x_{B_m^0 \cup \{0\}}} P'(x_{B_m^0 \cup \{0\}}) \log P(x_0 | x_{B_m^0}) \\
&= |Z_n^m| E_{P'}(\log P(x_0 | x_{B_m^0})).
\end{aligned}$$

We know that $|Z_n^m|/|\Lambda_n| \to 1$ as $n \to \infty$, so

$$\lim_{n\to\infty} \frac{1}{|\Lambda_n|} \sum_{\ell \in Z_n^m} \sum_{x_{\Lambda_n}} P'(x_{\Lambda_n}) \log P(x_\ell | x_{B_m^\ell \cap \Lambda_n}) = E_{P'}(\log P(x_0 | x_{B_m^0}))$$

As for term (2), we know that there are only finitely many values (with order $O(m^2)$) of $\sum_{x_{\Lambda_n}} P'(x_{\Lambda_n}) \log P(x_\ell | x_{B_m^\ell \cap \Lambda_n})$ for different $\ell \in \Lambda_n \backslash Z_n^m$, and each value is finite. Along with the fact that $|\Lambda_n \backslash Z_n^m|/|\Lambda_n| \to 0$ as $n \to \infty$, we know that

$$\lim_{n\to\infty} \frac{1}{|\Lambda_n|} \sum_{\ell \in \Lambda_n \backslash Z_n^m} \sum_{x_{\Lambda_n}} P'(x_{\Lambda_n}) \log P(x_\ell | x_{B_m^\ell \cap \Lambda_n}) = 0.$$

Thus we have

$$\lim_{n\to\infty}\frac{1}{|\Lambda_n|}H_{P'||P^m}(X_{\Lambda_n}) \;=\; -h_{P'}+E_{P'}(-\log P(X_0|X_{B_m^0})).$$

□

Lemma 2.2.27. *We have*

$$h_{P||P^m} \;\geq\; 0$$

for all $m\in\mathbb{N}$.

Proof. Since for any $n\in\mathbb{N}$,

$$\begin{aligned}
H_{P||P^m} &\triangleq -E_P(\log\frac{P^m(X_{\Lambda_n})}{P(X_{\Lambda_n})})\\
&= E_P(\log\frac{1}{P^m(X_{\Lambda_n})}-\log\frac{1}{P(X_{\Lambda_n})})\\
&= E_P(\sum_{\ell\in\Lambda_n}\log\frac{1}{P(X_\ell|X_{B_m^\ell\cap\Lambda_n})}-\sum_{\ell\in\Lambda_n}\log\frac{1}{P(X_\ell|X_{(<\ell)\cap\Lambda_n})})\\
&= \sum_{\ell\in\Lambda_n}E_P(\log\frac{1}{P(X_\ell|X_{B_m^\ell\cap\Lambda_n})}-\log\frac{1}{P(X_\ell|X_{(<\ell)\cap\Lambda_n})}).
\end{aligned}$$

And we have

$$E_P(\log\frac{1}{P(X_\ell|X_{B_m^\ell\cap\Lambda_n})}) \;\geq\; E_P(\log\frac{1}{P(X_\ell|X_{(<\ell)\cap\Lambda_n})}),$$

which is because $B_m^\ell\cap\Lambda_n\subset(<\ell)\cap\Lambda_n$. So $H_{P||P^m}(X_{\Lambda_n})>0$ for any $n\in\mathbb{N}$, which implies $h_{P||P^m}\triangleq\lim_{n\to\infty}\frac{1}{|\Lambda_n|}H_{P||P^m}(X_{\Lambda_n})>0$. □

The following result is intuitive, says that when m gets larger, then P^m will be closer to P. The original 1D version of this can be found in Lemma 2.6.1 of [17].

Lemma 2.2.28. $\lim_{m\to\infty} h_{P||P^m} \downarrow 0$.

Proof. From Lemma 2.2.26, since

$$\begin{aligned} h_{P||P^m}(X_{\Lambda_n}) &= -h_P + E_P(-\log P(X_0|X_{B_m^0})) \\ &= -h_P + H_P(X_0|X_{B_m^0}) \\ &\downarrow \ 0 \end{aligned}$$

when $m \to \infty$, which is from Lemma 2.2.3. □

The following lemma gives the relation between ψ and P^m. It is extracted and modified to 2D from the prove from Lemma 3.3.2 of [17]. This will be used in proving Lemma 2.2.32 on the next subsection.

Lemma 2.2.29. $\inf_{m\in\mathbb{N}} h_{\psi_\omega||P^m} = 0 \ P-a.e.$

Proof. The proof follows the idea of Lemma 2.2.28. First, calculate

$$\int \inf_{m\in\mathbb{N}} h_{\psi_\omega||P^m} dP(\omega) \leq \inf_{m\in\mathbb{N}} \int h_{\psi_\omega||P^m} dP(\omega) = (\star).$$

Then

$$
\begin{aligned}
(\star) \qquad & \underset{\text{Lem } 2.2.26}{=} && \inf_{m\in\mathbb{N}} \int dP(\omega)\{-h_{\psi_\omega} + E_{\psi_\omega}(-\log P^m(X_{\Lambda_n}(\omega))\} \\
& \underset{\text{Thm } 2.2.20}{=} && -h_P + \inf_{m\in\mathbb{N}} \int dP(\omega) E_{\psi_\omega}(-\log P^m(X_{\Lambda_n}(\omega))) \\
& = && -h_P + \inf_{m\in\mathbb{N}} \int dP(\omega) E_{\psi_\omega}(-\log P(X_0|X_{B_m^0})) \\
& \underset{\text{Thm } 2.2.18 \text{ (b)}}{=} && -h_P + \inf_{m\in\mathbb{N}} E_P(-\log P(X_0|X_{B_m^0})) \\
& \underset{\text{Lem } 2.2.26}{=} && \inf_{m\in\mathbb{N}} h_{P||P^m} \\
& \underset{\text{Lem } 2.2.28}{=} && 0
\end{aligned}
$$

Along with the fact that the integrand, $\inf_{m\in\mathbb{N}} h_{\psi_\omega||P^m}$, is non-negative, so $\inf_{m\in\mathbb{N}} h_{\psi_\omega||P^m}$ must be 0 $P-a.e.$ □

2.2.5 Coding Rate on a Stationary Source

In this subsection we extend the Shannon-McMillan-Breimen (SMB) theorem to a 2D source which is stationary but not necessary ergodic. The dynamical system $(\Omega, \mathcal{F}, P, \mathcal{T})$, as we defined it for 2D-lattice in the beginning of this section, is only assumed to be stationary. In this case, the coding rate with P will no longer be a constant but a random variable. In fact, we will see, in Theorem 2.2.34, the coding rate converges to the entropy rate of the source's ergodic component, i.e.

$$\lim_{n\to\infty} -\frac{\log P(X_{\Lambda_n})(\omega)}{|\Lambda_n|} = h_{\psi_\omega},$$

where the ergodic component function ψ is defined in Theorem 2.2.18. Note that

ψ is induced from $(\Omega, \mathcal{F}, P, \mathcal{T})$.

We will prove this by the idea from Gray's book [17], where it was proved in 1D, in the end of this subsection (Theorem 2.2.34) after we prepared several lemmas. The following Lemma is Lemma 3.3.1 of [17] for 1D version.

Lemma 2.2.30. *(Coding rate for the ergodic components) Assume* $(\Omega, \mathcal{F}, P, \mathcal{T})$ *is stationary. We have*

$$P(\{v \in \Omega : \lim_{n\to\infty} -\frac{1}{|\Lambda_n|} \log \psi_\omega(X_{\Lambda_n}(\omega)) = h_{\psi_\omega}\}) \quad = \quad 1,$$

where $h_{\psi_\omega} \triangleq \lim_{n\to\infty} \frac{1}{|\Lambda_n|} H_{\psi_\omega}(X_{\Lambda_n}) = \lim_{n\to\infty} \frac{1}{|\Lambda_n|} E_{\psi_\omega}(-\log \psi_\omega(X_{\Lambda_n}))$

Remark. That this is not the result of Corollary 2.2.2, where, in terms of the notation here, it says

$$\frac{1}{|\Lambda_n|} \log \psi_\omega(X_{\Lambda_n}(v)) \to h_{\psi_\omega}$$

for $\psi_\omega - a.e.\ v \in \Omega$, instead of P-a.e. $v \in \Omega$.

Proof. Define

$$\begin{aligned} G &= \{\omega : \lim -\frac{1}{|\Lambda_n|} \log \psi_\omega(X_{\Lambda_n}(\omega)) = h_{\psi_\omega}\} \\ G_\delta &= \{\omega : \lim -\frac{1}{|\Lambda_n|} \log P_\delta(X_{\Lambda_n}(\omega)) = h_{P_\delta}\} \end{aligned}$$

for all $P_\delta \in \mathcal{P}_e$. Because of the Theorem 2.2.18(b), we know that

$$P(G) \quad = \quad \int P_\delta(G) dQ_\psi(\delta),$$

and

$$P_\delta(G) = P(G|\psi = \delta) = P(G \cap \{\psi = \delta\}|\psi = \delta) = P(G_\delta|\psi = \delta) = P_\delta(G_\delta) = 1.$$

The last equality is because of the Corollary 2.2.2 and the fact that P_δ is a stationary ergodic distribution for $X_{\Lambda_n}(\omega)$, for all ω with $\psi(\omega) = \delta$. Thus $P(G) = 1$. □

We can now prove the first part of the Theorem 2.2.34. It is similar to the first part of the proof of Lemma 3.3.2 in Gray's book [17] for 1D version. Note that Gray's argument in the proof has some defects inside, and we manage to correct that here.

Lemma 2.2.31. *(Non-ergodic SMB lower bound) Assume* $(\Omega, \mathcal{F}, P, \mathcal{T})$ *is stationary, then for* P*-a.e.* ω

$$\liminf \frac{1}{|\Lambda_n|} \log \frac{\psi_\omega(X_{\Lambda_n}(\omega))}{P(X_{\Lambda_n}(\omega))} \quad \geq \quad 0.$$

Proof. Note that for any event $F \in \mathcal{F}$, if $P(F) = 0$, we have $\psi(F) = 0$ P-a.e. from the ergodic decomposition. In the following, we will omit ω for simplicity. Given any $\epsilon > 0$, we have

$$\begin{aligned}
P(\frac{1}{|\Lambda_n|} \log \frac{P(X_{\Lambda_n})}{\psi(X_{\Lambda_n})} > \epsilon) \quad &= \quad P(\frac{P(X_{\Lambda_n})}{\psi(X_{\Lambda_n})} > e^{\epsilon|\Lambda_n|}) \\
&\overset{\text{Thm 2.2.18(b)}}{=} \int P(\frac{P(X_{\Lambda_n})}{\psi(X_{\Lambda_n})} > e^{\epsilon|\Lambda_n|}|\psi = \delta) dQ_\psi(\delta) \\
&= \int P_\delta(\frac{P(X_{\Lambda_n})}{P_\delta(X_{\Lambda_n})} > e^{\epsilon|\Lambda_n|}) dQ_\psi(\delta) \\
&\leq \int E_{P_\delta}(\frac{P(X_{\Lambda_n})}{P_\delta(X_{\Lambda_n})}) e^{-\epsilon|\Lambda_n|} dQ_\psi(\delta),
\end{aligned}$$

by Markov inequality. Note that we use the notation $\delta = P_\delta$ in equations. Next,

define $A_n^\delta \triangleq \{a_n \in \mathcal{X}^{\Lambda_n} | P_\delta(a_n) > 0\}$, we can evaluate the expectation by the following observation:

$$\begin{aligned} E_{P_\delta}(\frac{P(X_{\Lambda_n})}{P_\delta(X_{\Lambda_n})}) &= \sum_{a_n \in A_n^\delta} \frac{P(a_n)}{P_\delta(a_n)} P_\delta(a_n) \\ &\leq 1, \end{aligned}$$

which concludes

$$P(\frac{1}{|\Lambda_n|} \log \frac{P(X_{\Lambda_n})}{\psi(X_{\Lambda_n})} > \epsilon) \leq e^{-\epsilon|\Lambda_n|}.$$

So we have

$$\sum_{n=1}^{\infty} P(\tfrac{1}{|\Lambda_n|} \log \tfrac{P(X_{\Lambda_n})}{\psi(X_{\Lambda_n})} > \epsilon) < \infty.$$

Then by Borel-Cantelli Lemma, $P(\frac{1}{|\Lambda_n|} \log \frac{P(X_{\Lambda_n})}{\psi(X_{\Lambda_n})} > \epsilon, \text{i.o.}) = 0$, which means

$$\limsup \frac{1}{|\Lambda_n|} \log \frac{P(X_{\Lambda_n})}{\psi(X_{\Lambda_n})} \leq \epsilon.$$

Since ϵ is arbitrary, we have $\limsup \frac{1}{|\Lambda_n|} \log \frac{P(X_{\Lambda_n})}{\psi(X_{\Lambda_n})} \leq 0$, i.e.

$$\liminf \frac{1}{|\Lambda_n|} \log \frac{\psi(X_{\Lambda_n})}{P(X_{\Lambda_n})} \geq 0.$$

□

Next Lemma gives the other direction, which is the second part of the proof of Lemma 3.3.2 in [17] for 1D version. Note that it is tempted to use the same method as in the Lemma 2.2.31, and using the Markov inequality on $\frac{1}{|\Lambda_n|} \log \frac{\psi(X_{\Lambda_n})}{P(X_{\Lambda_n})}$,which involve

the calculation of $E_P(\frac{\psi(X_{\Lambda_n})}{P(X_{\Lambda_n})})$. This expectation is hard to bound since it is difficult to evaluate the integral $\int \psi_\omega(X_{\Lambda_n(\omega)})d\omega$. Instead, we use the Markov approximation to help estimate.

Lemma 2.2.32. *(non-ergodic SMB upper bound) Assume $(\Omega, \mathcal{F}, P, \mathcal{T})$ is stationary, then for P-a.e. ω*

$$\limsup \frac{1}{|\Lambda_n|} \log \frac{\psi_\omega(X_{\Lambda_n}(\omega))}{P(X_{\Lambda_n}(\omega))} \quad \leq \quad 0.$$

Proof. We are going to use the Markov approximation, P^m, to estimate the upper bound. First, let's establish the following claim.

Claim 2.2.33. For any $m \in \mathbb{N}$, $\limsup \frac{1}{|\Lambda_n|} \log \frac{P^m(X_{\Lambda_n}(\omega))}{P(X_{\Lambda_n}(\omega))} \leq 0$ P-a.e. ω

Proof. We will using the method similar to the way we used to prove the lower bound. For any $\epsilon > 0$, the Markov inequality shows

$$\begin{aligned} P(\frac{1}{|\Lambda_n|} \log \frac{P^m(X_{\Lambda_n})}{P(X_{\Lambda_n})} \geq \epsilon) &= P(\frac{P^m(X_{\Lambda_n})}{P(X_{\Lambda_n})} \geq e^{|\Lambda_n|\epsilon}) \\ &\leq E_P(\frac{P^m(X_{\Lambda_n})}{P(X_{\Lambda_n})})e^{-|\Lambda_n|\epsilon} \\ &= 1 \cdot e^{-|\Lambda_n|\epsilon}. \end{aligned}$$

So by Borel-Cantelli Lemma we have

$$P(\frac{1}{|\Lambda_n|} \log \frac{P^m(X_{\Lambda_n})}{P(X_{\Lambda_n})} \geq \epsilon, \text{ i.o}) = 0.$$

Thus, we have

$$\limsup \frac{1}{|\Lambda_n|} \log \frac{P^m(X_{\Lambda_n})}{P(X_{\Lambda_n})} \leq 0 \ P - a.e.,$$

since $\epsilon > 0$ is arbitrary. $\square$

Because the this claim, we can calculate

$$\begin{aligned} \limsup \frac{1}{|\Lambda_n|} \log \frac{1}{P(X_{\Lambda_n}(\omega))} &\leq \limsup \frac{1}{|\Lambda_n|} \log \frac{1}{P^m(X_{\Lambda_n}(\omega))} \\ &= E_{\psi_\omega}(-\log P(X_0|X_{B_m^0})), \end{aligned}$$

which is the result of Lemma 2.2.24 (SMB for P^m). Subtract $\lim \frac{1}{|\Lambda_n|} \log \frac{1}{\psi_\omega(X_{\Lambda_n}(\omega))}$, i.e. h_{ψ_ω} (Lemma 2.2.30), on both side, we have

$$\begin{aligned} \limsup \frac{1}{|\Lambda_n|} \log \frac{\psi_\omega(X_{\Lambda_n}(\omega))}{P(X_{\Lambda_n}(\omega))} &\leq -h_{\psi_\omega} + E_{\psi_\omega}(-\log P(X_0|X_{B_m^0})) \\ &= -h_{\psi_\omega} + E_{\psi_\omega}(-\log P^m(X_0|X_{B_m^0})) \\ &\underset{\text{Lemma 2.2.26}}{=} h_{\psi_\omega||P^m} \end{aligned}$$

for any $m \in \mathbb{N}$. So

$$\begin{aligned} \limsup \frac{1}{|\Lambda_n|} \log \frac{\psi_\omega(X_{\Lambda_n}(\omega))}{P(X_{\Lambda_n}(\omega))} &\leq \inf_{m \in \mathbb{N}} h_{\psi_\omega||P^m} \\ &\leq 0, \end{aligned}$$

which is from Lemma 2.2.29. $\square$

We are now ready to prove the non-ergodic SMB theorem. The following theorem is proved in 1D fashion in section 3.3 of [17].

Theorem 2.2.34. *(Coding Rate on Stationary Source, the generalized SMB theorem) Assume the dynamical system $(\Omega, \mathcal{F}, P, \mathcal{T})$ is stationary, we have for almost every*

$\omega \in \Omega$

$$\lim_{n\to\infty} -\frac{\log P(X_{\Lambda_n})(\omega)}{|\Lambda_n|} \quad = \quad h_{\psi_\omega},$$

for ψ is the induced ergodic component function defined in Theorem 2.2.18.

Proof. (Theorem 2.2.34) Combine Lemma 2.2.31 and Lemma 2.2.32, we have

$$\lim_{n\to\infty} \frac{1}{|\Lambda_n|} \log \frac{\psi_\omega(X_{\Lambda_n}(\omega))}{P(X_{\Lambda_n}(\omega))} \quad = \quad 0 \ P\text{-a.e.}$$

Also we have, from Lemma 2.2.30 we have

$$\lim_{n\to\infty} -\frac{1}{|\Lambda_n|} \log \psi_\omega(X_{\Lambda_n}(\omega)) = \lim_{n\to\infty} \frac{1}{|\Lambda_n|} H_{\psi_\omega}(X_{\Lambda_n}) \triangleq h_{\psi_\omega}.$$

So we have the result

$$\lim_{n\to\infty} -\frac{1}{|\Lambda_n|} \log P(X_{\Lambda_n}(\omega)) \quad = \quad h_{\psi_\omega}.$$

□

2.3 The Ideal Compression Algorithm

Like many compression algorithms, we predict the pixel value by looking at its former decoded neighboring pixels. The set $B_m^\ell = (< \ell) \cap \Lambda_m^\ell$ as we defined before can be considered as the prediction window of size m for pixel location $\ell \in \Lambda_n$ (see Figure 2.2.2). In the end of this section, Corollary 2.3.7 will tell us that a successful algorithm can be built by estimating, $P(X_\ell | X_{B_m^\ell \cap \Lambda_n})$. In other words, it uses the local information, B_m, to achieve the theoretical lower bound for an image. More exactly, we would like to show

$$\lim_{m\to\infty} \lim_{n\to\infty} \frac{-1}{|\Lambda_n|} \sum_{\ell\in\Lambda_n} \log P(X_\ell | X_{B_m^\ell \cap \Lambda_n})(\omega) \quad = \quad h_{\psi(\omega)},$$

or, in terms of the Markov approximation (see Definition 2.2.22),

$$\lim_{m\to\infty} \lim_{n\to\infty} \frac{-1}{|\Lambda_n|} \log P^m(X_{\Lambda_n})(\omega) \quad = \quad h_{\psi(\omega)},$$

under the stationary condition in the sense of

1. $L^1(P)$ and
2. P-a.e. when
   (a) $\psi(\Omega)$ is countable, or
   (b) there is a $\xi \in (0,1)$ s.t. for any $m \in \mathbb{N}$, we have $P(X_0 | X_{B_m})(\omega) > \xi$ for P-a.e. $\omega \in \Omega$.

About the condition 2(a), even thought $\mathcal{X}$ has finite alphabets, $\Omega = \mathcal{X}^{\mathbb{Z}^2}$ is not

a countable set (since $\{0,1\}^{\mathbb{N}}$ has uncountable many elements). Also, note that a easy situation to achieve condition 2(b) is when one assumes there is always certain amount of noise in the image, where every value has chance to appear in any location of image.

2.3.1 The $L^1(P)$ Convergence

Remember that $h_\psi(\omega)$ is the ideal entropy rate for every $\omega \in \Omega$ defined as in Lemma 2.2.30:

$$h_\psi(\omega) \triangleq \lim_{n\to\infty} \frac{1}{|\Lambda_n|} H_{\psi(\omega)}(X_{\Lambda_n})$$

And from Lemma 2.2.24, we defined the entropy rate for the Markov approximation as

$$h^m(\omega) \triangleq \begin{cases} E_{\psi_\omega}(-\log P(X_0|X_{B_m})) & , \text{ if } \lim_{n\to\infty} \frac{-1}{|\Lambda_n|} \log P^m(X_{\Lambda_n})(\omega) \text{ converges.} \\ \log|\mathcal{X}| & , \text{ o.w.} \end{cases}$$

such that $\lim_{n\to\infty} \frac{-1}{|\Lambda_n|} \log P^m(X_{\Lambda_n})(\omega) = h^m(\omega)$ P-a.e. Since $\lim_{n\to\infty} \frac{-1}{|\Lambda_n|} \log P^m(X_{\Lambda_n})(\omega)$ is convergent P-a.e. (Lemma 2.2.24), our objective is equivalent to show

$$\lim_{m\to\infty} h^m(\omega) \;=\; h_{\psi(\omega)}$$

in the sense of $L^1(P)$.

Theorem 2.3.1. *Under the stationary condition, we have*

$$h^m \to h_\psi \text{ in } L^1(P)$$

as $m \to \infty$.

Proof. Since $h^m(\omega) \geq h_\psi(\omega)$ for any $\omega \in \Omega$, we can calculate

$$\begin{aligned}
E_P|h^m - h_\psi| &= E_P(h^m) - E_P(h_\psi) \\
&\overset{\text{Thm 2.2.20}}{=} E_P(E_\psi(-\log P(X_0|X_{B_m^0})) - h_P \\
&\overset{\text{Thm 2.2.18(b)}}{=} E_P(E_P(-\log P(X_0|X_{B_m^0})|\psi) - h_P \\
&= E_P(-\log P(X_0|X_{B_m^0}) - h_P \\
&\to 0
\end{aligned}$$

as $m \to \infty$, which is from Lemma 2.2.3. □

One the other hand, the P-a.e. convergence of h^m is not as easy to see, we will discuss that in the next subsection.

2.3.2 The P-a.e. Convergence

Define

$$f_m(\omega) \triangleq \begin{cases} -\log P(X_0|X_{B_m})(\omega) & , \text{ if } P(X_{B_m})(\omega) > 0 \\ 0 & , \text{ o.w.} \end{cases}$$

for $\omega \in \Omega$, $m \in \mathbb{N}$. Note that $\{\omega \in \Omega | P(X_{B_m})(\omega) = 0\}$ is a measure 0 set with respect of P. We will see that, in Theorem 2.3.4, the P-a.e. convergence of h^m to h_ψ can be achieved when

$$\lim_{m\to\infty} E_{\psi(\omega)}(f_m) = E_{\psi(\omega)}(\lim_{m\to\infty} f_m) \ P\text{-a.e.} \tag{2.3.1}$$

The next two lemmas prepares Theorem 2.3.4. And we will discuss two conditions that achieve Equation 2.3.1.

Lemma 2.3.2. *We have*

$$\lim_{m\to\infty} f_m = -\log P(X_0|X_{(<0)})$$

P-a.e. and P_δ a.e. for $\delta \in \psi(\Omega)$.

Proof. We will show this with martingale theory. Define $\mathcal{F}_n \triangleq \sigma(\cup_{m=1}^n X_{B_m})$, thus $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ for all $n \in \mathbb{N}$, i.e. $\{\mathcal{F}_n\}$ is a filtration on$(\Omega, \mathcal{F}, P)$. Combine this with the fact that $1_{X_0=x_0}$ is an integrable random variable for any $x \in \mathcal{X}$ on Ω, we have

$$\lim_{n\to\infty} E(1_{X_0=x_0}|\mathcal{F}_n) = E(1_{X_0=x_0}|\mathcal{F}_\infty^0) \text{ a.e.}$$

by martingale theorem, where $\mathcal{F}_\infty \triangleq \sigma(\cup_{n=1}^{\infty}\mathcal{F}_n) = \sigma(X_{(<0)})$. That is,

$$\lim_{m\to\infty} P(X_0|X_{B_m}) \;=\; P(X_0|X_{(<0)}) \text{ a.e.}$$

So for any $x_0 \in \mathcal{X}$ s.t. $P(X_0 = x_0|X_{(<\ell)}) > 0$, we have

$$\lim_{m\to\infty} \log\frac{1}{P(X_0|X_{B_m})} \;=\; \log\frac{1}{P(X_0|X_{(<0)})} \text{ a.e.}$$

Thus

$$\begin{aligned} P(\lim_{n\to\infty} \log\frac{1}{P(X_0|X_{B_m})} = \log\frac{1}{P(X_0|X_{(<0)})}) &\geq P(\omega : P(X_0|X_{(<0)})(\omega) > 0) \\ &= 1 - P(\omega : P(X_0|X_{(<0)})(\omega) = 0) \\ &= 1, \end{aligned}$$

since

$$P(\omega : P(X_0|X_{(<0)})(\omega) = 0) \;=\; P(\omega : P(X_{(\leq 0)})(\omega) = 0) = 0.$$

So we have proved the convergence in P-a.e. sense. Then we can use the fact $P_\delta \ll P$ to imply the convergence is also P_δ-a.e. □

The following lemma is modified to 2D from lemma 8.6.2 of Gray's book [16]. The basic idea is that: When you conditioning on infinite past, you are actually conditioning on the ergodic component.

Lemma 2.3.3. *The mapping ψ is measurable with respect to $\sigma(X_{(<0)})$. In particular*

$$P(X_0|X_{(<0)}) \quad = \quad P(X_0|X_{(<0)},\psi) = \psi(X_0|X_{(<0)}).$$

Proof. Note that the second equality is just from the definition of ergodic decomposition. And the first equality is a direct result of $\psi \in \sigma(X_{(<0)})$.

To see this, first note that from the Birkhoff ergodic theorem on images (Theorem 2.2.19), for any $F \in \mathcal{F}$ we have for P-a.e. $\omega \in \Omega$,

$$\lim_{m\to\infty} \frac{1}{|B_m|} \sum_{\ell\in B_m} 1_F(T^{-\ell}\omega) \quad = \quad E_P(1_F|\mathcal{I}),$$

since $\{B_m\}_{m\in\mathbb{N}}$ is a special averaging sequence. Thus we have

$$\lim_{m\to\infty} \frac{1}{|B_m|} \sum_{\ell\in B_m} 1_F(T^{-\ell}\omega) \quad = \quad \psi_\omega(F)$$

P-a.e. Next, using the fact that the probability ψ can be determined by generating set $\mathcal{C}$, we can show ψ is $\sigma(X_{(<0)})$-measurable by choosing an event $G \in \mathcal{G}$, say G is $\sigma(X_{\Lambda_n^k})$-measurable, where $k \in \mathbb{Z}^2$ and $n \in \mathbb{N}$. Without loss of generality, we set $k = (0,0)$. Set $a_n \triangleq (n+1, n+1) \in \mathbb{Z}^2$, we have for P-a.e. ω:

$$\begin{aligned} \psi_\omega(G) &= \psi_\omega(T^{-a_n}G) \\ &= \lim_{m\to\infty} \frac{1}{|B_m|} \sum_{\ell\in B_m} 1_{T^{-a_n}G}(T^{-\ell}\omega), \end{aligned}$$

which is a limit of $\sigma(X_{(<0)})$ measurable functions since $T^{-a_n}F_0$ is $\sigma(X_{(<0)})$-measurable and $1_A(T^{-\ell}\omega)$ is $\sigma(X_{(<0)})$-measurable if A is $\sigma(X_{(<0)})$-measurable and $\ell < 0$. Thus we have proved that for any event in generating set $\psi_\omega(G)$ is $\sigma(X_{(<0)})$-measurable.

This fact implies that ψ is $\sigma(X_{(<0)})$-measurable. $\square$

Theorem 2.3.4. *If* $\lim_{m\to\infty} E_{\psi(\omega)}(f_m) = E_{\psi(\omega)}(\lim_{m\to\infty} f_m)$ *P-a.e., we have*

$$\lim_{m\to\infty}\lim_{n\to\infty}\frac{-1}{|\Lambda_n|}\log P^m(X_{\Lambda_n})(\omega) \;=\; h_{\psi(\omega)}$$

$P-a.e.$

Proof. Note that $\lim_{n\to\infty}\frac{-1}{|\Lambda_n|}\log P^m(X_{\Lambda_n})(\omega) = E_{\psi(\omega)}(f_m)$ for P-a.e. $\omega\in\Omega$, so we only need to prove $\lim_{m\to\infty} E_{\psi(\omega)}(f_m) = h_{\psi(\omega)}$ P-a.e. Given an $\omega\in\Omega$ with $\psi(\omega)=\delta$ and $P(\psi^{-1}(\delta))>0$, we have

$$\begin{aligned}
\lim_{m\to\infty} E_\delta(f_m) &= E_\delta(\lim_{m\to\infty} f_m)\\
&\underset{\text{Lem. 2.3.2}}{=} E_\delta(-\log P(X_0|X_{(<0)}))\\
&\underset{\text{Thm. 2.2.18(b)}}{=} E(-\log P(X_0|X_{(<0)})|\psi=\delta)\\
&\underset{\text{Lem. 2.3.3}}{=} E(-\log \psi(X_0|X_{(<0)})|\psi=\delta)\\
&= E_\delta(-\log P_\delta(X_0|X_{(<0)}))\\
&= h_\delta.
\end{aligned}$$

Next, for the rest of the ω, say $A=\{\omega\in\Omega|P(\psi^{-1}(\psi(\omega)))=0\}$, we will prove it is a P measure 0 set. Since $\Omega=\mathcal{X}^{\mathbb{Z}^2}$ is countable,

$$P(A)=\sum_{\omega\in A}P(\omega)$$

and because of $\omega \in \psi^{-1}\psi(\omega)$, we have

$$\begin{aligned} P(\omega) &\leq P(\psi^{-1}\psi(\omega)) \\ &= 0 \end{aligned}$$

for $\omega \in A$. Hence $P(A) = 0$. □

The next two lemmas give two cases that are sufficient for Theorem 2.3.4.

Lemma 2.3.5. *If $\psi(\Omega)$ has only countably many elements, then*

$$\lim_{m\to\infty} E_{\psi(\omega)}(f_m) = E_{\psi(\omega)}(\lim_{m\to\infty} f_m)$$

for P-a.e. $\omega \in \Omega$.

Proof. Since $\psi(\Omega)$ has only countably many elements, we can write

$$P(F) = \sum_{\eta\in\psi(\Omega)} q_\eta P_\eta(F)$$

for any $F \in \mathcal{F}$, where $q_\eta \triangleq P(\psi^{-1}(\eta))$.

For any ω s.t. $P(\omega) > 0$, define $\delta = \psi(\omega)$. Note that $q_\delta > P(\omega) > 0$. We will prove this by stating that $\{f_m\}_{m\in\mathbb{N}}$ is uniformly integrable with respect to probability P_δ. This implies the desired property for any ω s.t. $P(\omega) > 0$, which means the property

holds for P-a.e. $\omega \in \Omega$. Given $k \in \mathbb{N}$, calculate

$$\begin{aligned}
\int_{f_m \in [k,k+1)} f_m dP_\delta &\leq (k+1) P_\delta(f_m \in [k, k+1)) \\
&\leq (k+1) P_\delta(P(X_0|X_{B_m}) \in (2^{-k-1}, 2^{-k}]) \\
&\leq (k+1) P_\delta(P(X_0|X_{B_m}) \leq 2^{-k}) \\
&= (k+1) \sum_{x_{B_m} \in \chi^{B_m}} \{P_\delta(P(X_0|X_{B_m}) \leq 2^{-k}|X_{B_m} = x_{B_m}) P_\delta(X_{B_m} = x_{B_m})\} \\
&= (k+1) \sum_{x_{B_m} : P_\delta(x_{B_m}) > 0} \{P_\delta(P(X_0|x_{B_m}) \leq 2^{-k}|x_{B_m}) P_\delta(x_{B_m})\}
\end{aligned}$$

Note that we can ignore those x_{B_m} with $P(x_{B_m}) = 0$, since $P(F|x_{B_m}) = 0$ for any $F \in \mathcal{F}$ and those term will vanish. Also observe that for those x_{B_m} s.t. $P_\delta(x_{B_m}) > 0$, we have $P(x_{B_m}) > 0$. Let's calculate the the first term inside the summation:

$$\begin{aligned}
P_\delta(P(X_0|x_{B_m}) \leq 2^{-k}|x_{B_m}) &= P_\delta(\frac{P(X_{B_m \cup 0})}{P(x_{B_m})} \leq 2^{-k}|x_{B_m}) \\
&= P_\delta(P(X_{B_m \cup 0}) \leq P(x_{B_m}) \cdot 2^{-k}|x_{B_m}) \\
&= P_\delta(\sum_{\eta \in \psi(\Omega)} q_\eta P_\eta(X_{B_m \cup 0}) \leq P(x_{B_m}) \cdot 2^{-k}|x_{B_m}) \\
&\leq P_\delta(q_\delta \cdot P_\delta(X_{B_m \cup 0}) \leq P(x_{B_m}) \cdot 2^{-k}|x_{B_m}) \\
P_\delta(x_{B_m}) > 0 \quad &= P_\delta(q_\delta \cdot P_\delta(X_0|x_{B_m}) \leq \frac{P(x_{B_m})}{P_\delta(x_{B_m})} \cdot 2^{-k}|x_{B_m}) \\
&= P_\delta(P_\delta(X_0|x_{B_m}) \leq q_\delta^{-1} \frac{P(x_{B_m})}{P_\delta(x_{B_m})} \cdot 2^{-k}|x_{B_m}) \\
&= \sum_{x_0 \in \mathcal{X} : P_\delta(x_0|x_{B_m}) \leq q_\delta^{-1} \frac{P(x_{B_m})}{P_\delta(x_{B_m})} \cdot 2^{-k}} P_\delta(x_0|x_{B_m}) \\
&\leq q_\delta^{-1} \frac{P(x_{B_m})}{P_\delta(x_{B_m})} \cdot 2^{-k} |\mathcal{X}|.
\end{aligned}$$

So,

$$\begin{aligned}\int_{f_m\in[k,k+1)} f_m dP_\delta &\leq (k+1)\sum_{x_{B_m}:P_\delta(x_{B_m})>0}\{q_\delta^{-1}\frac{P(x_{B_m})}{P_\delta(x_{B_m})}\cdot 2^{-k}|\mathcal{X}|P_\delta(x_{B_m})\}\\ &= (k+1)\sum_{x_{B_m}:P_\delta(x_{B_m})>0}\{q_\delta^{-1}\cdot P(x_{B_m})\cdot 2^{-k}|\mathcal{X}|\}\\ &\leq (k+1)q_\delta^{-1}2^{-k}\cdot|\mathcal{X}|.\end{aligned}$$

Thus we can check the uniformly integrability of f_m by

$$\begin{aligned}\int_{f_m>k} f_m dP_\delta &= \sum_{i=1}^{\infty}\int_{f_m\in[k+i,k+i+1)} f_n dP_\delta\\ &\leq \sum_{i=1}^{\infty}(k+i+1)q_\delta^{-1}2^{-k+i+\log|\mathcal{X}|},\end{aligned}$$

which converges to 0 as $k\to\infty$ uniformly in m. □

Lemma 2.3.6. *If there is a $\xi\in(0,1)$ s.t. for any $m\in\mathbb{N}$, $P(X_0|X_{B_m})(\omega)>\xi$ for P-a.e. $\omega\in\Omega$, then*

$$\lim_{m\to\infty} E_{\psi(\omega)}(f_m) = E_{\psi(\omega)}(\lim_{m\to\infty} f_m)$$

for P-a.e. $\omega\in\Omega$.

Proof. Let Ω_0 be the collection of $\omega\in\Omega$ s.t. $\lim_{m\to\infty} f_m(\omega)$ is convergent and $P(X_0|X_{B_m})(\omega)>\xi$. Note that $P(\Omega_0)=1$ by Lemma 2.3.2 and the assumption here. For $\omega\in\Omega_0$, since $P(X_0|X_{B_m})(\omega)>\xi>0$, we have $f_m\leq-\log\xi$ (note that the condition also implies $P(X_{B_m}(\omega))>0$). Thus, we have $f_m\leq-\log\xi$ P-a.e.

Thus, let $\delta=\psi(\omega)$, by the Lebesgue dominate convergence theorem, we have

$$\lim_{m\to\infty} E_\delta(f_m) = E_\delta(\lim_{m\to\infty} f_m),$$

which proves the lemma. □

2.3.3 How to Build an Ideal Compression Algorithm

Let's summarize the previous results in this section by the following corollary.

Corollary 2.3.7. *(The Ideal Algorithm) Under the stationary condition, we have*

$$\lim_{m\to\infty}\lim_{n\to\infty}\frac{-1}{|\Lambda_n|}\sum_{\ell\in\Lambda_n}\log P(X_\ell|X_{B_m^\ell\cap\Lambda_n})(\omega) \;=\; h_{\psi(\omega)}$$

in the sense of

1. $L^1(P)$, *and*
2. P*-a.e. when* $\psi(\Omega)$ *is countable, or there is a* $\xi\in(0,1)$ *s.t. for any* $m\in\mathbb{N}$, $P(X_0|X_{B_m})(\omega)>\xi$ *for* P*-a.e.* $\omega\in\Omega$

Roughly speaking, this corollary tells us that, under light conditions, with large context, the compression algorithm that build upon $P(X_0|X_{B_m\cap\Lambda_n})$ will approach the optimal entropy rate. Thus, one can focus on building a good estimator of $P(X_\ell|X_{B_m^\ell\cap\Lambda_n})$. One idea is to estimate this empirically as follows: For any $A\subset\mathbb{Z}^2$, assume $\{(y^u,y_A^u)\}_{u=1}^r$ to be an i.i.d. sample of random variable (X,X_A), and we define $\hat{P}^r$, the conditional empirical distribution of P with the library of size r , through

indicator functions as follows:

$$\hat{P}^r(X = x|X_A = x_A) = \begin{cases} \frac{\sum_{u=1}^{r} \mathbb{I}_{\{x=y^u, x_A=y_A^u\}}}{\sum_{u=1}^{r} \mathbb{I}_{\{x_A=y_A^u\}}} & , \text{ if } \sum_{u=1}^{r} 1_{\{x_A=y_A^u\}} \neq 0 \\ 0 & , \text{ o.w.,} \end{cases}$$

where $\mathbb{I}$ is an indicator function. Then for any m, n, we define $P^r(X_\ell = x|X_{B_m^\ell \cap \Lambda_n})$, the estimator of $P(X_\ell = x|X_{B_m^\ell \cap \Lambda_n})$ to be:

$$P^r(X_\ell = x|X_{B_m^\ell \cap \Lambda_n}) = \begin{cases} \hat{P}^r(X = x|X_{B_m}) & , \text{ if } B_m \cap \Lambda_n = B_m \\ \frac{1}{|\mathcal{X}|} & , \text{ o.w.} \end{cases} \tag{2.3.2}$$

By estimating with empirical distribution, we can expect that the larger amount of samples, r, we have, the better our result is. That is, "the more you know the better you get." The obvious con for doing this is that we will need enormous amount of r to get a reasonable result. We will device a feasible way of utilizing the idea described here in the next section.

2.4 Application

In this section we will introduce a compression scheme, called the comparison-based image compression algorithm (CBIC), which utilize the ideal algorithm idea described in subsection 2.3 to compress a real world image. The algorithm is introduced in Subsection 2.4.1, and we examine its performance in Subsection 2.4.2. Note that it is a valid image compression algorithm; however it is very computationally expensive in view of nowadays computer. The goal here is to build an algorithm that is able to scale up the context size, so as to estimate the minimum bits per pixel for an image. And compare the result with the best image compression algorithm today, CALIC [41].

A (black and white) image is set to have 256 depth of illumination, that is, $\mathcal{X} = \{0, 1, \cdots, 255\}$. The image we want to code is called the original image, and the training image set are called the library images.

2.4.1 Comparison-Based Image Compression Algorithm

Instead of estimating $P(X_\ell | X_{B_m})$ with indicators, like we mentioned in Subsection 2.3.3, we use a kernel estimator to build the probability. The kernel estimator we use is a kind of a Laplace distribution

$$Ker_p(z) \triangleq \frac{1}{\sum_{\tilde{z}=-|\mathcal{X}|}^{|\mathcal{X}|} p^{|\tilde{z}|}} p^{|z|}$$

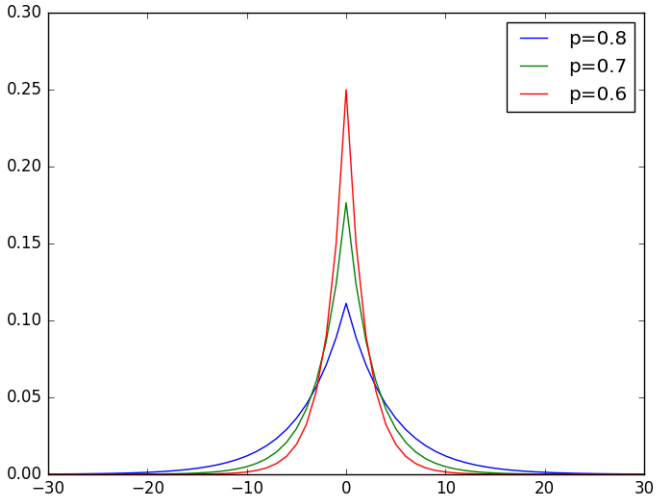

Figure 2.4.1: The kernel we use to approximate empirical distribution of $P(X_0|X_{B_m})$.

on $z \in \mathbb{Z}$ s.t. $|z| \leq 255$. for some $p \in (0,1)$. Figure 2.4.1 gives the plot of this kernel. Note that $Ker_p(z) \approx \frac{1-p}{1+p}p^{|z|}$. When $p \to 0$, the kernel becomes the indicator function like $1_{\{z=0\}}$, and when $p \to 1$, it becomes the uniform distribution.

For easier demonstration, set the target pixel of original patch locates at $(0,0)$ and its context, or say, neighbors, locates at $\mathcal{T} = B_m^0 \subset \mathbb{Z}^2$ (as shown in Figure 2.2.2). For $m = 2$, the corresponding patch is 3×5; for $m = 3$ is 4×7; $m = 4$ is 5×9. Note that $m = 4$ is the biggest context we will set in this experiments. To emphasize the importance of the context that near the target pixel, we will partition the context into several tiers. Set the tier of context $\mathcal{T}_b \subset \mathcal{T}$, $b = 1, \cdots, B$, such that $\{\mathcal{T}_b\}_{b=1}^B$ partitions $\mathcal{T}$. For a 3×5 patch, we set $B = 3$; for a 4×7 patch, $B = 4$; for a 5×9 patch, $B = 6$. For $\mathcal{T}_b$ with lower value of $b \in \{1, 2, \cdots, B\}$ means it is the tier that closer to the target pixel location. For each tier we apply different α-weight

(a) A 3×5 patch

1	1	2	2	1
1	2	3	3	2
2	3	x		

(b) A 4×7 patch

1	1	2	2	2	1	1
1	2	3	3	3	2	1
2	3	4	4	4	3	2
3	4	4	x			

(c) A 5×9 patch

1	4	4	4	9	4	4	4	1
4	4	9	9	16	9	9	4	4
4	9	16	16	25	16	16	9	4
4	9	16	25	36	25	16	9	4
9	16	25	36	x				

Figure 2.4.2: A example of α-weight corresponding to different tier and sizes of patch. Note that the weight is not yet been normalized and is shown proportionally. For example: (a) For a 3×5 patch, $\mathcal{T}_1$ has a weight proportion to 3, $\mathcal{T}_2$ has a weight proportion to 2; $\mathcal{T}_3$ has a weight proportion to 1, and $B = 3$ in this case. Note that a notation $\times$ indicates the target pixel location for each patch.

$\alpha = \{\alpha_b\}_{b=1}^{B}$, where

$$\sum_{b=1}^{B} \alpha_b |\mathcal{T}_b| = 1$$

See Figure 2.4.2 for the information about the α-weight. The value of the original patch will be represented as a pair $(\cdot, \cdot) \in \mathcal{X} \times \mathcal{X}^{\mathcal{T}}$, where the first entry is called the target pixel value, and the second entry is called the context value.

Before coding the patch, we first remove the illumination of both original patch and library patches. The mean of a patch is calculated with respect to α weight, then rounded to the closest integer. Note that every patch has its own mean. In the end we get the (value) centered original patch (y, x) and centered library patches $\{(y_i, x_i)\}_{i=1}^{N}$, where N is the size of library (patches). Also we record the α-weighted

mean, μ_y, for the original target pixel.

Set the following parameters:

1. $r \in [0, \infty)$: A parameter for calculating distance between two context values x and $\hat{x}$, where a distance is defined as

$$d_r(x, \hat{x}) \triangleq \sum_{b=1}^{B} \frac{b^{-r}||x_{\mathcal{T}_b} - \hat{x}_{\mathcal{T}_b}||_{L^1}}{\sum_{b=1}^{B} b^{-r}|\mathcal{T}_b|}.$$

2. $T \in (0, \infty)$: This parameter gives us more control of the distance between two context.

3. $K \in \mathbb{N}$: *Essential sample size* determines the amount of library patches which are similar to the original one.

Given parameters (p, r, T, K), we estimate $P(y|x)$ out of library $\{(y_i, x_i)\}_{i=1}^{N}$ as the following (simplified) procedures:

1. Find the suitable library set:

   (a) Compare the α-weighted L^1-norm between x and x_i for $i = 1, \cdots, N$. Say $d_r^i = d_r(x, x_i)$ for $i = 1, \cdots, N$.

   (b) Choose K smallest values among$\{d_r(x, x_i)\}_{i=1}^{N}$, say $\{d_r^{(k)}\}_{k=1}^{K}$, and find the corresponding target pixel value of the library patches, say $\{y^{(k)}\}_{k=1}^{K}$. We will use these samples corresponds to $\{d^{(k)}\}_{k=1}^{K}$ to build the empirical distribution in the next step.

2. Calculate the the following density for $T > 0$:

$$\tilde{P}^N(y|x;\alpha,p,r,T,K) = \sum_{k=1}^{K} \frac{e^{-d_r^{(k)}/T} Ker_p(y-\tilde{y}^{(k)})}{\sum_{j=1}^{K} e^{-d_r^{(j)}/T}}$$

where

$$\tilde{y}^{(k)} \triangleq \max(0, \min(255, y^{(k)} + \mu_y)).$$

Then $\tilde{P}^N(y|x)$ is the approximated conditional probability to $P(y|x)$. Remember that μ_y is the α-weighted mean we recorded while standardizing the original path (y, x).

Note that the density is consisted of kernel estimators weighted by context similarities. One would expect that when $N \to \infty$, $d^{(k)}$ become 0 for all $k = 1, \cdots, K$. And when the essential sample number K grows, the optimal p would goes to 0, which indicates that $\tilde{P}^N$ is an approximate of the empirical conditional distribution. The whole procedure is illustrated in Figure 2.4.3.

Assume there are L interior pixels (that is, pixels that have complete B_m neighbor), the total bits that need to be transmitted for interior pixels is

$$f(\alpha,p,r,T,K) = -\sum_{\ell=1}^{L} \log_2 \tilde{P}_\ell^N(y_\ell|x_\ell;\alpha,p,r,T,K),$$

where $\tilde{P}_\ell^N$ is $\tilde{P}^N$ specified by pixel ℓ's neighbors x_ℓ.

Ideally, one would like to find the parameters α, p, r, T, and K, that minimize the total bits. Since sending these parameter won't add many bits. Realistically,

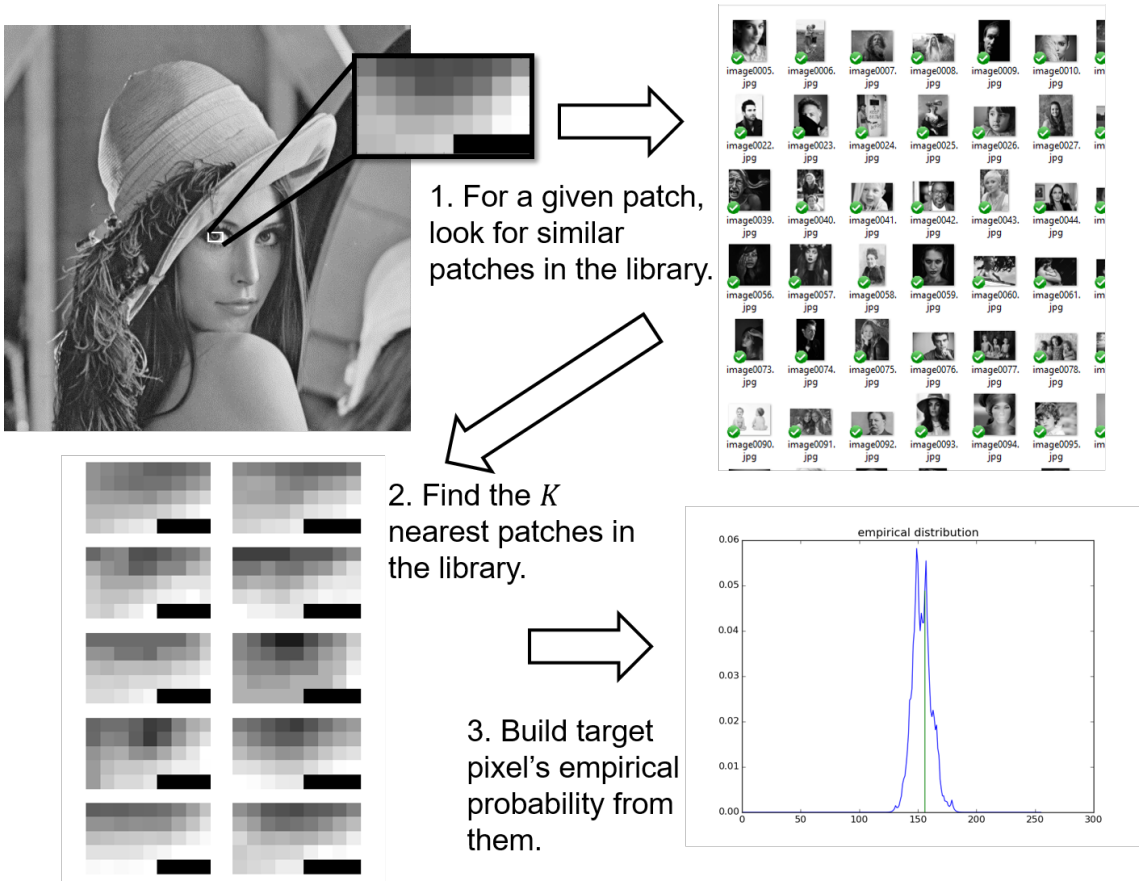

Figure 2.4.3: Demonstration of CBIC algorithm.

we can reduce the problem to just optimize the total bits with p, r, and T. First, the parameter α is used to centering the data, so that we can use the library more efficiently. That is, without the illumination, we leave only the texture of the patches, which makes the matching procedure easier. (We have tried about standardize the patch with its standard deviation but this adds more burden computationally and didn't gain more success.) If α is chosen badly, we can still overcome it with enormous library. So the role of α should have less impact on the result. In the experiment, we will simply choose α-weight as in Figure 2.4.2. Second, as for essential sample size K, the bigger K means more samples on estimating the empirical distribution, but it also means the quality of sample will drop since total library size N is limited. However, the quality of the sample can also be controlled by r and T, so higher number of K won't affect much if we can choose r and T freely.

The task left is how we optimize the rest of the parameters: p, r, and T. Note that, for different locations of target pixels, the algorithm use different set of K samples from library to determine the empirical probability. The size of total, N, is often enormous. For example, the highest number we use during the experiment is $N = 80$ million of 5×9 patches. It's impractical when we try to optimize the total bit since every time we choose a triple (p, r, T), we have to pull K nearest patch out of N library for each (interior) pixel, which requires expensive computation resources. The solution is that we choose κ nearest library patches for each pixel, in the sense that the patches have smaller α-weighted L^1 norms between neighbors, where $\kappa > K$; often we choose $\kappa = 1000$. Note that, again, every different locations of target pixels have its different κ library patch set. To choose κ library for a pixel, we simply use the same α-weight as we did when calculating the mean. Then it is possible to optimize the parameter r, T, and p. In practical, we iteratively optimize between r, T, and

p until they converge. Theoretically, function f is not convex with respect to any of the three parameters. However, we find no trouble as we just ask the computer to minimize with respect of either one with brute force. In fact, f shows convexity with respect to p in every our experiments. Of course, one can always use larger κ to downgrade the effect of different α. To have an idea of the parameter, for "Lena" image that we will use in the next subsection, we get the estimates of $p \approx 0.544$, $r \approx 1.7$, and $T \approx 1.2$.

A heap based partial sorting algorithm is used in our implementation, as we extract K nearest contexts out of N library. The procedure follows by sorting first K distances out of N distances, which forms a sorted partition of size K, then insert the unsorted $N - K$ distances into the sorted partition and remove the largest element in the partition. It takes $O(N \log K)$ times to do so.

For users to be able to decode the image, we need to transfer the information about the edge pixels of the original image. The decoding procedure for 2×3 patch is demonstrated in Figure 2.4.4. It is similar for a context model that has maximum patchsize 3×5, 4×7, 5×9, and so on. For example, as we code with context model with maximum 3×5 patches, which needs 2 columns on each of the left and right, and 2 rows on top of the original image (these are called edge pixels). We use the same method on 1×3, 3×1,2×3, and 3×2, patchsizes of context model to code it. To code edge pixels, with the same κ size of library for each edge pixel, we optimize T and p corresponds to the edge pixel while using parameter r acquired during the optimization for the interior pixel. One should understand that when the size of image is large, the bits that generate from edge pixels are marginal and will not affect overall bpp much.

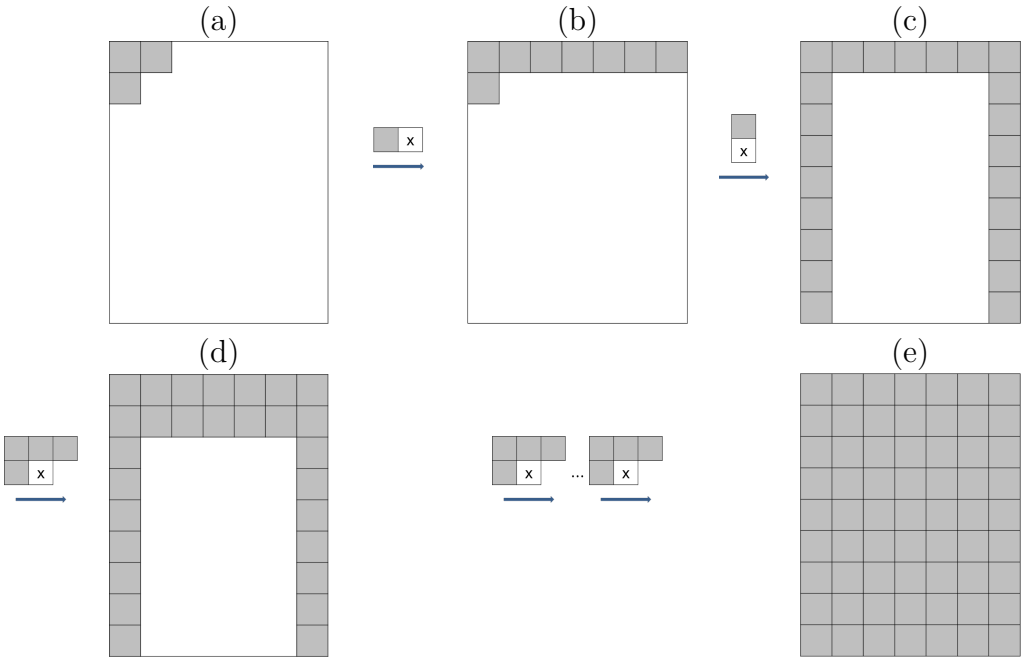

Figure 2.4.4: Raster scan as decoding the image from an algorithm built on context model with maximum patchsize 2×3. The procedure start from (a): Given the first 3 top-left corner pixel values, using a context model on 1×2 patch to get the first row of image (b). Next, apply a context model on 1×2 patch to get left and right column of image. Then use that of 2×3 to get the second row, and continue to obtain the whole image (e).

Finally, when $\tilde{P}_\ell^N$ fail to give a good estimate, say $\tilde{P}_\ell^N < c$ for some original pixel ℓ, we replace it with location information and record the pixel value directly. For example, when the illumination depth is 256, $c = \log_2 \bar{L} + \log_2 256$, where $\bar{L}$ is the total number of pixel of an image. That is, when recording $\tilde{P}_\ell^N$ costs more than the c, we simply replace it with c. So the total bits should be

$$-\sum_{\ell=1}^{L} \log_2 \tilde{P}_\ell^N \mathbb{I}_{\tilde{P}_\ell^N \leq c} + c\mathbb{I}_{\tilde{P}_\ell^N > c}.$$

When sending the code, we need to add the following bits:

- 3 pixel values for upper left corner, 8 bits each.
- 3 parameters (T, p, r) for interior bpp, 32 bits for T and r, 8 bits for p.
- 2 parameters (T, p) for edge bpp, 32 bits for T, 8 bit for p.
- 2 parameters for image dimension, 16 bits each

2.4.2 The Performance

The performance of an image compression algorithm can be measured by calculate the bits per pixel (bpp, the total file size divided by number of pixels). We prepared 4 test images as shown in Figure 2.4.5. These images are 8-bit image (2^8 possible pixel values), so the worst possible bpp result are 8 bits. Modern lossless compression algorithm often has ability to compress them to 3-5 bpp. The "big library" are 647 portraits images from internet (converted to 8 bit images), which comprise about 300

Figure 2.4.5: 4 test images: Lena, Jetplane, Goldhill, and Barbara. Goldhill has size 720×576, and others are 512×512.

million 4×7 image patches, and we will randomly choose several amount of patches in it to build the library for our experiments.

Due to the limitation of the modern computer, for CBIC algorithm, we choose the essential sample number $K = 200$ in general (we tried use higher amount of K, however, the result only improve a little). Figure 2.4.6 compares different patchsize and library sizes while compressing "Lena" image. The library for each experiment is randomly selected from the big library. Roughly speaking, when library size is small, 5×9 patchsize has slightly worse result than that from 4×7. But when the library size is high, like 40 millions, the compression rate improves from 5×9. This is because it is harder for 5×9 patches to find quality samples from smaller library. When library size is up to 40 million, a larger patch size gives a better result. In this case, the gain between 5×9 and 4×7 are marginal, compare to that between 4×7 and 3×5, which indicates the importance of neighbors diminishes along patchsize.

Figure 2.4.7 compares the result of CBIC, the proposed method, with CALIC (see [41]) and Jpeg-LS (see [25]). Note that even though we use the portrait library, the image with different subjects (e.g. Jetplane or Goldhill) still gives compatible result with CALIC. This suggests that a successful compression scheme can rely only

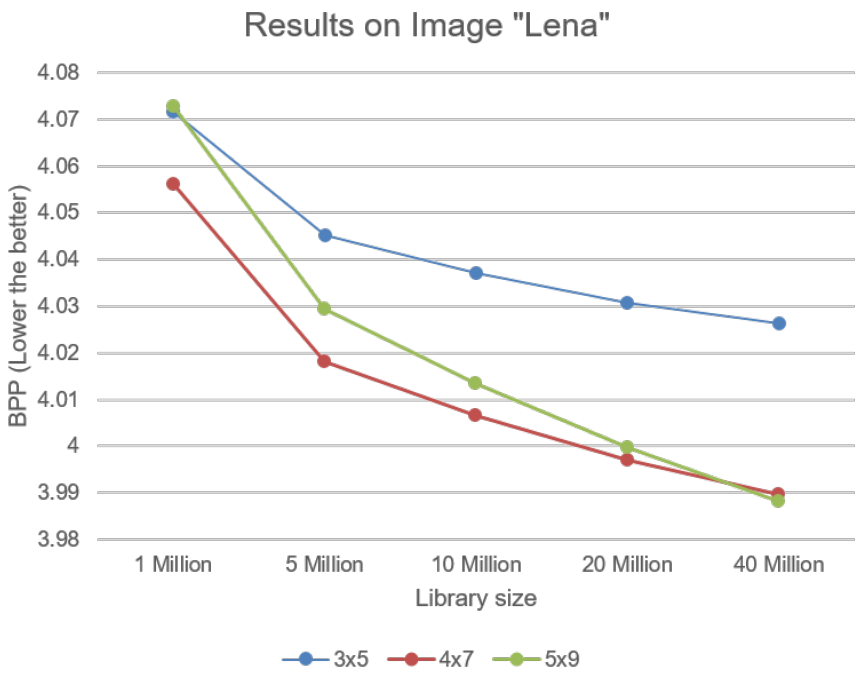

Figure 2.4.6: The bpp result of CBIC algorithm for "Lena" image with different size of patches and different size of library. In these experiments, we choose essential sample amount $K = 200$ to build the empirical distribution.

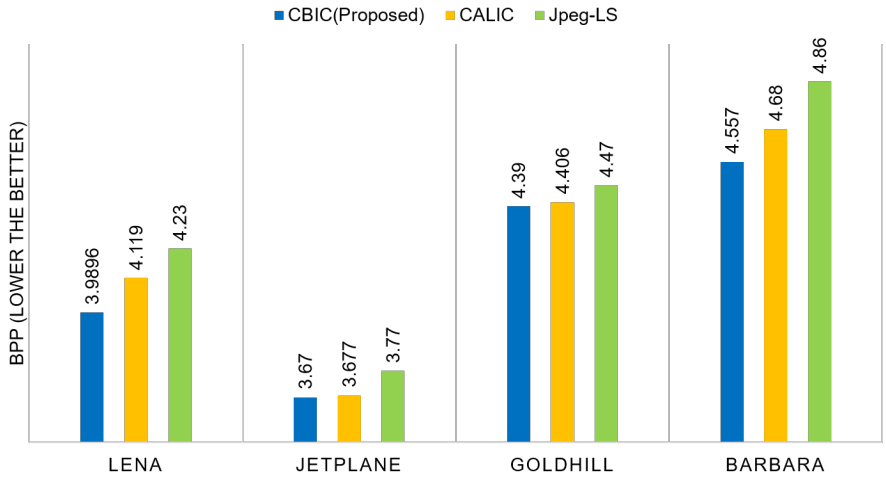

Figure 2.4.7: Comparison between CBIC and other famous lossless compression algorithms. The following information in the parenthesis shows the number of library and the patchsize used to generate the result here: Lena: (80M, 5×9); Jetplane: (80M, 4×7); Goldhill: (40M, 4×7); Barbara: (20M, 4×7).

on the local information of an image, rather than the global one. Also notice that our algorithm shows good result on compressing portrait images "Barbara" with only 20 million library (also, from Table 2.4.6, 20 million library patches also gives good result on "Lena") , and this might indicate it will be efficient to find patches within the same type of photos. For reader who followed through the article, this might be an evidence that these portrait images have similar ergodic components.

In Figure 2.4.8, we reconstruct the Lena image with pixel value that has highest probability the estimator gives (i.e. the MAP, maximum a posterior probability, estimators). That is, each pixel intensity is the one that has highest value of $\tilde{P}_\ell^N(y_\ell|x_\ell)$, note that x_ℓ is the actual pixel values of the neighborhood of ℓth pixel[8]. The estimator is obtained from our method with 4×7 patchsize, with 5 million library patches. Note that 5 million library patches is considered as the smallest amount of library in

[8]Do not confused this with lossy compression. This used the original image context, hence it is not a lossy compression result.

our experiment. One can see that a lot of detail is captured by our models, includes some of the textures of the hat and the hair.

Despite all the great performance of CBIC, the main drawback of the algorithm is the maximum memory usage and expensive computation resource requirements. While we implement the algorithm with Python combine with C that handles the most intensive sorting algorithm, it would take several hours to encode, and even more time to decode. For example, we used Amazon EC2 to enhance the computation, which gives 16 or more cores for parallel computing, yet it still take 8-10 hours to encode a small 512×512 image with 40 million library.

Why do we built such an algorithm? The main purpose is to have an image compression algorithm that is able to scale up the context size, and to achieve the optimal coding rate, so that we can see how a state of the art algorithms performed. As one can see from Table 2.4.7, we are only a little better (about 0.05 to 0.15 bpp) than one of the best scheme. Also, when we scale up the context size (see Table 2.4.6), the gain in performance is diminished. This suggests that the state of the art lossless image compression algorithm is already near optimal and it might be hopeless to find another scheme that is overwhelmingly better than the one we have today.

Figure 2.4.8: Top: Reconstruct the image with MAP estimators. Bottom: The original picture.

Chapter Three

Robust Generalized Clustering

3.1 Introduction

A traditional way to summarize a data is through regression. Over years of research, many generalizations have been developed, some of which search for multiple structures associated with different parts of the data. For example, Breiman proposed [8] the regression trees algorithm (CART, classifications and regression trees) that finds histogram-liked structures to explain the data in each of many cells that are derived by a recursive splitting of the variable associated with each dimension. Freidman introduced MARS, the multivariate regression splines [14], using hinge function as a basis to learn piece-wise linear functions. The first stage of MARS is to overfit the data with hinge functions, where hinge functions with the smallest sum of squared residuals are added iteratively. Then a backward deletion method is applied to obtain the final result. Later, Breiman combined the regression trees idea with hinge functions, and proposed an algorithm called the hinge hyperplane method, [7], where a relatively computationally inexpensive hinge-finding algorithm is introduced for learning optimal regression trees. Another example is SVR, or "support vector regression," proposed by Vapnik, et. al. [40]. SVR seeks the flattest nonlinear structure that fits most of the data, by solving a quadratic programming problem. A clear development of SVR can be found in [39].

The above-mentioned methods accommodate what we call "multiple instance data," in which there are multiple models, each designed to fit a different subset of the data. In these cited examples, there is a dependent variable and the ultimate goal is regression. We will also assume that multiple models are needed to explain the data, but in something more akin to the clustering problem than the regression

problem—e.g. see Figure 3.1.1. The best-known approach is probably the K-means algorithm: Given data $\{z_i\}_{i=1}^n \in \mathbb{R}^d$, K-means (see [27] for an early reference) seeks m points $\{c_j\}_{j=1}^m \subset \mathbb{R}^d$ that minimize the loss function, f:

$$f(c_1, \cdots, c_k) = \sum_{i=1}^{n} \min_{j=1:m} \|z_i - c_j\|^2$$

A heuristic algorithm is applied: Firstly, arbitrarily choose m points in $\mathbb{R}^d$ and associate (cluster) the data that is closest to each point. Second, calculate the mean of each cluster and repeat the first step with these means. Continue recursively. Apparently, this would not be a fruitful approach to the data in Figure 3.1.1. A generalization of this idea, replacing points with hyperplanes ("K-planes", [28]), was proposed by Manwani at 2012. However, the direct generalization quickly gets stuck stuck in local minimums and is highly sensitive to noise, as the author himself indicates in his paper. This might be the reason for why this simple generalization has found little favor in literature. Manwani's remedy is to add a local penalty function, $\|z_i - \mu_j\|^2$, to the loss function to ensure that the points in each cluster are close together. This approach can also be viewed as an instance of the EM algorithm [10], in which residuals are modeled as a Gaussian mixture with prior. We will compare our method with an orthogonal version of EM in Section 3.4. Not surprisingly, we will find that EM with Gaussian mixtures is sensitive to noise and outliers. Our approach, which will be introduced shortly, offers a way to learn multiple structures using a very simple loss function, similar to that of the K-means algorithm, but with substantial robustness.

The robustness issue already appears in the context of ordinary linear regression. Some noise or outliers can quickly undermine the result. This is the inevitable nature

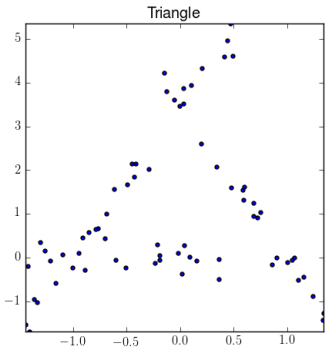

Figure 3.1.1: "Triangle data"

of a quadratic loss function, which nevertheless enjoys many excellent theoretical (and even practical properties) under a Gauss-Markov type condition. But most real-world data is not well described by these assumptions. Many methods and resources have been dedicated to remedying this. A good treatment of many of these can be found int the book by Sen and Srivastava [38]. One fruitful line of attack is called, quite generally, "robust regression." Whereas the objective is still to minimize the sum of residuals, the focus is on using different functions of the residuals. The first approaches in this direction date back to the 18th century, with the idea to replace squared residuals by the absolute value of residuals, "*least absolute value* (LAV) regression," proposed by Boscovich [5] and later studied by Laplace [22]. (A comprehensive review of LAV regression can be found in [11].) To demonstrate the difference between LAV and traditional regression, consider the data set displayed in Figure 3.1.2. We generated 100 points from a distribution concentrating near a straight line, and then added 50 outliers uniformly distributed in the rectangle. How does β affect the minimizer of

$$f(\ell) = \sum_{i=1}^{n} d(z_i, \ell)^{\beta},$$

where $d(z, \ell)$ is the Euclidean distance from a point z to a line ℓ? When $\beta = 2$, the minimizer is the line that passes through the mean of the data and has the largest variance of the projected data (i.e. the eigenvector of the sample covariance matrix corresponding to the largest eigenvalue). Here, we are using "orthogonal distance regression" [4], as defined in the formulation of principle component analysis (PCA). Notice that the optimal line when $\beta = 1$ is less affected by outliers than the corresponding line when $\beta = 2$. (Traditional LAV applies to the *vertical* residuals, as in traditional regression, rather than the orthogonal distance used in PCA.)

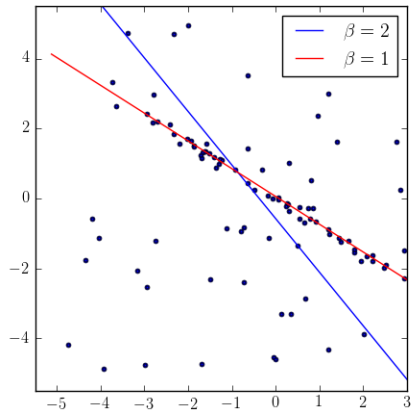

Figure 3.1.2: Minimizers of $f(\ell) = \sum_{i=1}^{n} d(z_i, \ell)^\beta$ when $\beta = 1$ and $\beta = 2$, where ℓ is a straight line in $\mathbb{R}^2$.

A unified approach is Huber's M-estimator [20], which determines regression parameters by minimizing the sum of error function

$$\sum_{i=1}^{n} \rho(\epsilon_i),$$

where ϵ_i, $i = 1, 2, \cdots, n$, are the residuals, andρ is a convex function. Note that when ρ is the square function the estimator is the ordinary regression estimator; LAV regression is recovered by taking ρ to be the absolute value. As for non-convex ρ,

Huber wrote

> "If ρ is not convex, the estimators... will, in general, no longer converge toward some constant. Apparently, one has to impose not only local but also some global conditions on ρ in order to have consistency."

In contrast to Huber's ideas, we propose to explore non-convex error functions. We forgo any reasonable notion of consistency, but we gain a strong sense of robustness, and in some cases computational advantages as well. Specifically, we consider residual functions of the form $\sum_{i=1}^{n} d(z_i, \ell)^\beta$, where $d(z_i, \ell)$ is Euclidean distance and $\beta \in (0, 1)$. When $\beta \in (0, 1)$ we call $d(z, \ell)^\beta$ a *sparse distance*.

For sparse distances, exact minimums can sometimes be computed, and oftentimes approximated (3.2.2, 3.4). As we shall see, sparser distances (smaller values of β) are more robust. As an illustration, in the data shown in Figure 3.1.3 there 70 outliers out of 100 data points. The 30 "good" data points were generated from a noisy line and the 70 outliers from the uniform distribution on the rectangle. As β is made smaller, the minimizing line "migrates" to a line that is nearly indistinguishable from the one that generated the 30 data points.

We are interested in extensions to "multiple instances": problems for which two or more shapes (e.g. lines, or lines and planes) might be the most appropriate summary. This calls for an extension of the loss function. Given data $\vec{z} = \{z_i\}_{i=1}^{n} \subset \mathbb{R}^{d1}$ and a class of models $\mathcal{M}$ (e.g. all one and two-dimensional linear manifolds in $\mathbb{R}^d$),

[1] We use $\vec{z}$, ambiguously, to represent the vector $(z_1, z_2, \ldots, z_n)$ and the (possibly multi) set $\{z_1, z_2, \ldots, z_n\}$.

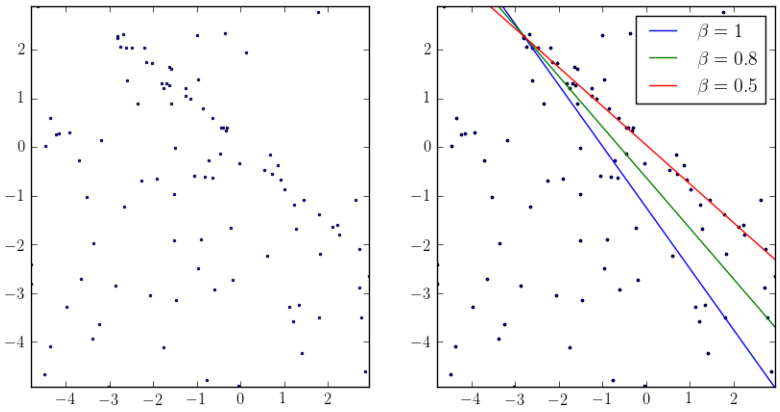

Figure 3.1.3: Left: the data includes 70 points generated from the uniform distribution. Right: The effect of β on the minimizer of $f(\ell) = \sum_{i=1}^{n} d(z_i, \ell)^\beta$, over all lines, ℓ, in $\mathbb{R}^2$.

we propose an unsupervised clustering algorithm through minimization of the loss function

$$f(L) \triangleq \gamma(L) + \frac{\lambda}{n} \sum_{i=1}^{n} \min_{\ell \in L} d(z_i, \ell)^\beta, \tag{3.1.1}$$

where d is, again, Euclidean distance and $L \in 2^{\mathcal{M}}$ is a set of instances (or "structures" or "models"). Obviously, there needs to be a penalty on the number of instances: $\gamma(L)$ is a monotone increasing function of the number of models in L.

The minimization of the loss functions with sparse distance involves the "*interpolated solutions*." By interpolating we means it is fully determined by the data points that it passes through, e.g. a two-dimensional plane passing through three linearly independent points. In LAV regression, it is known that for d dimensional data $(x_i, y_i) \in \mathbb{R}^{d-1} \times \mathbb{R}^1$, $i = 1, 2, \cdots, n$ (with full-rank matrix), the regression hyperplane will pass through at least d data points, see [1]. (Some other properties of

the estimator can be found in Chapter 1 of [23].) Formally, the minimizer, $\hat{b}$, of

$$f(b) = \sum_{i=1}^{n} |y_i - \sum_{j=0}^{d-1} x_{ij} b_j|,$$

where $x_{i0} = 1 \ \forall i$, defines a hyperplane,

$$\{(x, y) \in \mathbb{R}^{d-1} \times \mathbb{R}^1 : y = \hat{b}_0 + \sum_{j=1}^{d-1} x_j \hat{b}_j\}$$

that includes at least d of the data points. It is then said that the solution *is interpolated.* As we shall see, the phenomenon is also present in the minimization of

$$\sum_{i=1}^{n} d(z_i, \ell)^\beta,$$

over all $\ell \in \mathcal{M}$, where $\mathcal{M}$ is the set of ($d-1$ dimensional) hyperplanes and $z_i \in \mathbb{R}^d \ \forall i$. The solution passes through at least d data points. While this result for LAV regression (or for that matter the case here, when $\beta = 1$) can be proven through the dual representation, the technique does not work for $\beta < 1$. We will treat this more general case in 3.2.2 and 3.2.3, where we will describe a "interpolated solution." For more general planes, of dimension $k < d - 1$, the minimum-loss solution does not necessarily go through any data points. In this case, the k-dimensional planes that pass through $k + 1$ data points can only be considered as local minima. We discuss this in Section 3.2.4, and for special cases $k = 0$ and $k = 1$ we will give bounds on β which guarantee that the optimizer is interpolating.

It is instructive to take a close look at the simplest example, which is the clustering of points on the line. In Section 3.3 we will examine a limited type of consistency and we will show that the estimator has 50% breakdown point, which is the highest

possible for a location estimator.

In Section 3.4 we will develop an approximate minimizer of the (multiple-instance) loss function, which we will call the "Remove or Replace Algorithm" (RRA). We will perform numerous experiments, with different amounts of outliers and a variety of model classes, $\mathcal{M}$. For "mixed" model classes—those with two or more parametrized shapes (e.g. lines and $d-1$ dimensional planes), care must be taken to not unduly favor model classes with more degrees of freedom, such as a hyperplane versus a line. We will offer an extension of RRA for choosing multiple (mixed) instances from mixed model classes. In particular, we will adapt a method called *minimum sum of volumes*, which provides a way to select a final subset from a candidate list of multiple mixed instances. Candidates are chosen using the RRA algorithm. In our experiments, we obtain natural summaries of complex data, even in the presence of a high percentage of outliers. The approach, using sums of volumes, is similar, but not the same as, a method known as the *minimum volume covering ellipsoid*. The idea of clustering with minimum volume ellipsoids was first proposed by Rosen [36]; see also Barnes [2] for further developments of the algorithm.

3.2 Loss Function for Multiple Instances

The main object of study, in this section, will be a special case of Equation 3.1.1: given $\beta \in (0,1)$, let

$$f(L) \;=\; |L|^2 + \frac{\lambda}{n}\sum_{i=1}^{n} \min_{\ell \in L} d(z_i,\ell)^\beta. \tag{3.2.1}$$

where $L \subset \mathcal{M}$ and $d(z_i,\ell)$ is the (orthogonal) distance[2] from $z_i \in \mathbb{R}^d$ to the model instance $\ell \in \mathcal{M}$.

Before minimizing Function 3.2.1, we standardize the data to achieve scale and translation invariance. That is, for a raw data $\{\hat{z}_i\}_{i=1}^n$,

$$z_i = \frac{\hat{z}_i - \bar{\hat{z}}}{\sqrt{\frac{1}{n}\sum_{i=1}^n \|\hat{z}_i - \bar{\hat{z}}\|^2}},$$

where $\bar{\hat{z}} = \frac{1}{n}\sum_{i=1}^n \hat{z}_i$ and $\|\cdot\|$ is the usual L^2 norm. Thus we have $\sum_{i=1}^n z_i = 0$, and $\frac{1}{n}\sum_{i=1}^n \|z_i\|^2 = 1$.

Our main goal will be to prove that when $\mathcal{M}$ is the family of $d-1$ dimensional planes in $\mathbb{R}^d$ then, generically, the minimizer of 3.2.1 passes through at least d linearly independent points in $\vec{z}$ (i.e. "interpolates"). Subsections 3.2.2 and 3.2.3 are devoted to the proof. Section 3.2.4 gives some partial results for the more difficult case of planes of dimension less than $d-1$.

Precisely, here is what we mean by :

[2]We will only discuss *linear* model classes, but we have in mind more general shapes. In general, "orthogonal" would be replaced by minimum distance.

Definition 3.2.1. Fix $d \geq 2$ and $k \leq d-1$. Let $\mathcal{M}_k$ be the set of all k dimensional planes in $\mathbb{R}^d$, let $z_1, z_2, \ldots, z_n$ be n elements of $\mathbb{R}^d$, and write $\vec{z}$ for the *set* $\{z_1, z_2, \ldots, z_n\}$. A plane $\ell \in \mathcal{M}_k$ is *interpolating* (or, sometimes, *interpolates*) relative to $\vec{z}$ if it contains at least $k+1$ linearly independent points from $\vec{z}$.

Before turning to these topics, we will first establish, in Section 3.2.1, a straightforward result that will greatly simplify our subsequent analyses.

3.2.1 Reduction to $m = 1$

Given a minimizer of 3.2.1, and a particular model instance $\ell \in \mathcal{M}$ that is a component of the minimizer, consider the subset of data points, $A \subset \vec{z}$, which are closer to ℓ than to any other model instance in the solution set (for now, ignore ties). Does ℓ minimize $\sum_{z \in A} d(z_i, \cdot)^\beta$? The obvious answer, 'yes', is correct, but to show this we can not simply assume the contrary and then improve ℓ, since changing ℓ will, in general, also change A. Here we give a proof, and, later, we will use this result to help delineate some key properties of an optimal solution.

Definition 3.2.2. Let $L^* = \{\ell_1^*, \ell_2^*, \ldots, \ell_m^*\}$ be a minimizer of Function 3.2.1. An *optimal assignment*, given L^*, is any partition of $\{z_1, z_2, \ldots, z_n\}$,

$$A(L^*) = \{A_1(L^*), A_2(L^*), \ldots, A_m(L^*)\},$$

such that

$$z \in A_j(L^*) \Rightarrow d(z, \ell_j^*)^\beta \leq d(z, \ell_k^*)^\beta \ \forall k = 1 : m$$

To ease notation we will sometimes write $\rho_z(\ell)$ in place of $d(z,\ell)^\beta$. With this convention:

Theorem 3.2.3. *Let $L^* = \{\ell_1^*, \ell_2^*, \ldots, \ell_m^*\}$ be a minimizer of Function 3.2.1, and let $A(L^*)$ be an optimal assignment for L^*. Then for all $j = 1, \cdots, m$*

$$\ell_j^* \in \arg\min_{\ell \in \mathcal{M}} \sum_{z \in A_j(L^*)} \rho_z(\ell),$$

Proof. Suppose not. Then without loss of generality we may assume that

$$\ell_1^* \notin \arg\min_{\ell \in \mathcal{M}} \sum_{z \in A_1(L^*)} \rho_z(\ell)$$

In which case there exists $\tilde{\ell}_1 \in \mathcal{M}$ such that

$$\sum_{z \in A_1(L^*)} \rho_z(\ell_1^*) > \sum_{z \in A_1(L^*)} \rho_z(\tilde{\ell}_1)$$

Define $\tilde{L} = \{\tilde{\ell}_l, \ell_2^*, \ell_3^*, \ldots, \ell_m^*\}$, and $\tilde{\ell}_j = \ell_j^*$ for $j = 2, 3, \cdots, m$. For any $L \subset \mathcal{M}$, define $h(L) = \sum_{i=1}^n \min_{l \in L} \rho_{z_i}(\ell)$. Then

$$\begin{aligned}
h(L^*) &= \sum_{i=1}^n \min_{\ell \in L^*} \rho_{z_i}(\ell) \\
&= \sum_{j=1}^m \sum_{z \in A_j(L^*)} \rho_z(\ell_j^*) \\
&> \sum_{z \in A_1(L^*)} \rho_z(\tilde{\ell}_1) + \sum_{j=2}^m \sum_{z \in A_j(L^*)} \rho_z(\ell_j^*) \\
&= \sum_{j=1}^m \sum_{z \in A_j(L^*)} \rho_z(\tilde{\ell}_j) \\
&\geq \sum_{i=1}^n \min_{l \in \tilde{L}} \rho_{z_i}(\ell) = h(\tilde{L})
\end{aligned}$$

which contradicts the optimality of L^*. □

Basically, Theorem 3.2.3 says that the any assignment of ambiguous points (points that have the same distance to two or more instances in a minimizer) would not affect the result of the theorem, i.e. ℓ^* will still minimize its cluster. An interesting fact can be derived: When $\beta = 2$ and point model is used, there is no ambiguous point for the minimizer. This is because, if there were ambiguous points, any different assignment will result in different minimizer, which violate the property of unique solution of a (strictly) convex optimization problem. However, when $\beta < 1$, it is possible to see ambiguous points.

3.2.2 Lines in $\mathbb{R}^2$

A natural choice of model $\mathcal{M}$ is the set of straight lines in $\mathbb{R}^2$, which is called the *line model*. The main purpose here is to show that the minimizer of

$$f(\ell) = \sum_{i=1}^{n} d(z_i, \ell)^\beta, \tag{3.2.2}$$

over the line model, given data $\vec{z} = \{z_i\}_{i=1}^n \subset \mathbb{R}^2$, is interpolating—i.e. passes through at least two points (Theorem 3.2.6).

To have a better understanding of this, first consider both data point $z_i \in \mathbb{R}^1$, $\mathcal{M} = \mathbb{R}^1$. It's easy to see that the minimizer exists since f is continuous and the minimizer must be within a bounded set $[\min_{i=1:n} z_i, \max_{i=1:n} z_i]$. Also, because of $f''(x) < 0$ for $x \notin \vec{z}$, the only possible minimizer can only occur when $x \in \vec{z}$. That is,

the minimizer interpolates. Our proof of that on $\mathbb{R}^2$ is similar to this, as one will see.

Here are some notations we will use throughout the section. An element in $\mathbb{R}^d$ is considered as a d-dimensional column vector, and the superscript T on a vector means the transpose of the vector. The identity matrix is denoted as I. The superscript $\perp$ on a space is the orthogonal complement of the space, i.e. $B^\perp = \{y \in \mathbb{R}^d : x \cdot y = 0 \text{ for any } x \in B\}$ for any B is a subspace in $\mathbb{R}^d$.

Let $S \triangleq \mathbb{R} \bmod \pi$, and $v(\theta) \triangleq (\cos(\theta), \sin(\theta))^T$ for $\theta \in S_1$. Define a function $\mathcal{L}$ that maps $\mathbb{R} \times S$ to a closed set on $\mathbb{R}^2$, s.t. $\mathcal{L}(a, \theta) = \{y \in \mathbb{R}^2 : y^T v(\theta) = a\}$. Thus, we define a line $\mathcal{L}(a, \theta)$ s.t. it has normal vector $v(\theta)$ and $d(0, \mathcal{L}(a, \theta)) = |a|$. Note that this defines an unique line on $\mathbb{R}^2$, so we can define the space of lines on $\mathbb{R}^2$ as $\mathbb{L} \triangleq \mathbb{R} \times S$. Note that the space S has a natural metric

$$d_S(\theta_1, \theta_2) \quad = \quad \min\{|\theta_1 - \theta_2|, \pi - |\theta_1 - \theta_2|\}.$$

So we can define the metric of space $\mathbb{L}$ as the product metric

$$d_{\mathbb{L}}(\mathcal{L}(a_1, \theta_1), \mathcal{L}(a_2, \theta_2)) \triangleq |a_1 - a_2| + d_S(\theta_1, \theta_2)$$

for $a_i \in \mathbb{R}$, $\theta \in S$. Thus we have a metric space for lines $(\mathbb{L}, d_{\mathbb{L}})$.

To prepare Theorem 3.2.6, we need to show that the minimizer exists (Lemma 3.2.4 and 3.2.5).

Lemma 3.2.4. *Function 3.2.2 is continuous on* $(\mathbb{L}, d_{\mathbb{L}})$.

Proof. For any $z \in \mathbb{R}^2$, we only need to show that $d(z, \cdot)$ is continuous on $(\mathbb{L}, d_{\mathbb{L}})$.

Then $\sum_{i=1}^{n} d(z,\cdot)^{\beta}$ is continuous on $(\mathbb{L}, d_{\mathbb{L}})$. Set a line $\ell_{a,\theta} = \mathcal{L}(a,\theta)$ for $a \in [0,\infty)$ and $\theta \in S_1$, since $d(\cdot,\cdot)$ is the Euclidean distance. Note that $\ell_{a,\theta} = av(\theta) + \ell_{0,\theta}$. since the we have

$$\begin{aligned} d(z,\ell_{a,\theta}) &= d(z - av(\theta), \ell_{0,\theta}) \\ &= \|(z - av(\theta))^T v(\theta)\|, \end{aligned}$$

where $v(\theta) = (\cos(\theta), \sin(\theta))$. Thus, we can see that $d(z,\ell_{a,\theta})$, as a function of (a,θ), is a continuous function of a, $\cos\theta$, and $\sin\theta$. This implies $d(z,\ell_{x,\theta})$ is continuous on $(\mathbb{L}, d_{\mathbb{L}})$. □

Lemma 3.2.5. *The minimizer of Function 3.2.2 exists.*

Proof. Define a d-dimensional ball $B_v(\bar{z})$, where $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$, and $v = \max \|z_i - \bar{z}\|$. Define $B \triangleq \{\ell \in \mathbb{L} | \ell \text{ passes through } cl(B_{2v}(\bar{z}))\}$, where $cl(\cdot)$ means the closure of a set. Thus, for $\ell \notin B$ then

$$f(\ell) > nv^{\beta}.$$

Next, if $\ell_{\bar{z}}$ is a line starting at $\bar{z}$, we have

$$f(\ell_{\bar{z}}) \leq nv^{\beta}$$

for every possible direction. Thus we have

$$\min_{\ell \in B} f(\ell) \leq nv^{\beta} \leq \min_{\ell \notin B} f(\ell).$$

This implies

$$\min_{\ell \in \mathbb{L}} f(\ell) \quad = \min_{\ell \in B} \quad f(\ell),$$

which means we can restrict our search of lines on B, a compact set in $(\mathbb{L}, d_{\mathbb{L}})$. This, along with the fact that the Function 3.2.2 is continuous, the minimizer is always existed. □

Note that the minimizer of Function 3.2.2 might not be unique: Consider $d = 2$ and 4 points forms a exact square on a plane, then there are two lines admits the same minimum value of the target function.

Theorem 3.2.6. *Given data points $\{z_i\}_{i=1}^n \subset \mathbb{R}^2$, assume $\beta \in (0, 1)$, the minimizer of Function 3.2.2 interpolates.*

Proof. (Proof by contradiction) Assume the optimal line doesn't pass through any data points. Given any point $z^\star \in \mathbb{R}^2$ on the optimal line we have $z_i \neq z^\star$ for any $i = 1, \cdots, n$. We may further assume $z^\star = 0$ since we can shift the whole data to achieve it. Furthermore, set the optimal line, say $\ell(\theta^\star)$, parallel to the vector $(\cos\theta^\star, \sin\theta^\star)$ for some $\theta^\star \in [0, 2\pi)$. Define a line, $\ell(\theta)$, that passes the origin and parallel to the vector $(\cos\theta, \sin\theta)$ for $\theta \in [0, 2\pi)$. Then $\theta^\star$ should be the argument of minimum of the following function:

$$g(\theta) \quad = \quad \sum_{i=1}^{n} d(z_i, \ell(\theta))^\beta.$$

Note that, from our setting, the optimal line $\ell(\theta^\star)$ doesn't pass any data point, i.e. $\theta^\star \neq \theta_i + 2\pi k$, for any $k \in \mathbb{N}$ and $i = 1, \cdots, n$. For any data point z_i, $i = 1, \cdots, n$,

we can also write

$$z_i \;=\; ||z_i||(\cos\theta_i, \sin\theta_i)$$

for some $\theta_i \in [0, 2\pi)$. First, observe that

$$d(z_i, \ell(\theta)) \;=\; ||z_i|| \cdot |\sin(\theta - \theta_i)|.$$

For $i = 1, \cdots, n$, define $g_i(\theta) \triangleq d(z_i, \ell(\theta))^\beta$, thus we have

$$\begin{aligned} g_i(\theta) &= ||z_i||^\beta \cdot |\sin(\theta - \theta_i)|^\beta \\ \frac{\partial g_i(\theta)}{\partial \theta} &= ||z_i||^\beta \beta |\sin(\theta - \theta_i)|^{\beta-1} \cos(\theta - \theta_i) \cdot \mathrm{sgn}(\sin(\theta - \theta_i)), \end{aligned}$$

for any $\theta \neq \theta_i + 2\pi k$, where $\mathrm{sgn}(a) \triangleq 1_{a\geq 0} - 1_{a\leq 0}$. The Hessian of g_i can be calculated whenever $\theta \neq \theta_i + 2\pi k$,:

$$\begin{aligned} H_{g_i}(\theta) &= ||z_i||^\beta \beta \{(\beta-1)|\sin(\theta-\theta_i)|^{\beta-2}\cos^2(\theta-\theta_i) \cdot [\mathrm{sgn}(\sin(\theta-\theta_i))]^2 \\ &\quad + |\sin(\theta-\theta_i)|^{\beta-1} \cdot (-\sin(\theta-\theta_i)\cdot \mathrm{sgn}(\sin(\theta-\theta_i)) + 0)\} \\ &= ||z_i||^\beta \beta \{(\beta-1)|\sin(\theta-\theta_i)|^{\beta-2}\cos^2(\theta-\theta_i) - |\sin(\theta-\theta_i)|^{\beta-1}\cdot|\sin(\theta-\theta_i)|\} \\ &= ||z_i||^\beta \beta \{(\beta-1)|\sin(\theta-\theta_i)|^{\beta-2}\cos^2(\theta-\theta_i) - |\sin(\theta-\theta_i)|^{\beta}\} \\ &< 0. \end{aligned}$$

Thus for $\theta^\star \neq \theta_i + 2\pi k$,

$$H_g(\theta^\star) \;=\; \sum_{i=1}^n H_{g_i}(\theta^\star) < 0,$$

which means $\theta^\star$ is not the argument of minimum of g, we have the contradiction. Thus, the optimal line has to pass at least one data point, say z_1. Next, find the optimal direction θ by minimizing

$$g_{z_1}(\theta) = \sum_{i:z_i \neq z_1, i=1:n} (d(z_i, \ell_{z_1}(\theta))^\beta,$$

where $\ell_{z_1}(\theta)$ is a line passes z_1 with direction $(\cos\theta, \sin\theta)$. Using the same procedure as above: start from assuming the optimal line doesn't touch any data point other than those equal to z_1, and get the contradiction. One can conclude that the minimizer must admits a line that touches at least two data points. □

One might expect to have the same result for general data in $\mathbb{R}^d$—the optimal line is interpolating. However, unfortunately it is not true. The following example shows it is not necessary true.

Example 3.2.7. Consider a set of data that plotted in Figure 3.2.1

$$\begin{aligned} z_1 &= (1, 0, 0) \\ z_2 &= (-1, 0, 0) \\ z_3 &= (0, \sqrt{3}, 0) \\ z_4 &= (0, \frac{\sqrt{3}}{3}, b). \end{aligned}$$

Let $m = 1$, when b is sufficiently high, if we confine the search of the optimal line on only the line that connects two data points, the result line, say ℓ_g, will be one of

$\overline{z_1z_4}$,$\overline{z_2z_4}$, and $\overline{z_3z_4}$, which has the target function value of

$$f(\ell_g) = 2 \cdot (\frac{2\sqrt{b^2 + \frac{1}{3}}}{\sqrt{b^2 + \frac{4}{3}}})^\beta,$$

which goes to $2 \cdot 2^\beta$ when $b \to \infty$. Next, let's consider another line that go through z_4 and perpendicular to the $x - y$ plane, say ℓ_r, depicted as red line in Figure 3.2.1, will have target function value

$$f(\ell_r) = 3 \cdot (\frac{2}{\sqrt{3}})^\beta.$$

So, when $\beta \uparrow 1$, we have

$$\lim_{\beta\uparrow 1, b\to\infty} f(\ell_r) = 2\sqrt{3} \approx 3.46$$
$$\lim_{\beta\uparrow 1, b\to\infty} f(\ell_g) = 4,$$

That is, we found a line ℓ_g that has smaller value then any line that connects 2 data points. Hence, in $\mathbb{R}^d$,$d > 2$, the optimal line is not necessary interpolating.

3.2.3 Hyperplanes in $\mathbb{R}^d$

In this subsection, we will study the property of the minimizer while $\mathcal{M}$ is the family of $\mathbb{R}^{d-1}$ planes in $\mathbb{R}^d$, $d > 2$, which is called the plane model. Before we define space of planes, note that a vector $r \in \mathbb{R}^d_1 \triangleq \{r \in \mathbb{R}^d | r^T r = 1\}$ can be represented as a vector of angles $\eta = (\theta_1, \theta_2, \cdots, \theta_{d-1}) \in \Theta^d \triangleq [0, \pi]^{d-2} \times [0, 2\pi)$ using spherical coordinates. And there is a bijection $v : \Theta \to \mathbb{R}^d$ to convert from spherical coordinates to Cartesian

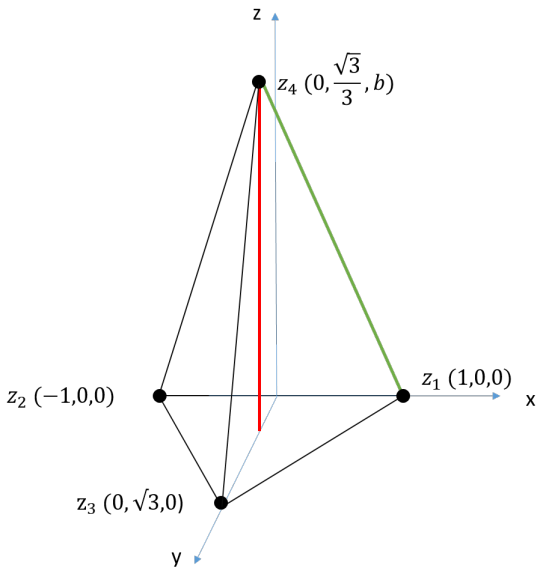

Figure 3.2.1: Counter example for the optimal line in $\mathbb{R}^3$, which explaining the optimal line might not necessary go through two data points. The data is $z_1, \cdots, z_4$, when b is high enough, the green line will be the best line (that has smaller or equal minimum target function value) among all the interpolating lines, i.e. the lines that connect two data points. However, when $\beta < 1$ and is close to 1, the red line will produce smaller target function value than that of the green one, and the red line only touches one data point.

coordinates. Note that the kth- elements of vector $v(\eta) \in \mathbb{R}^d$, $k = 1, 2, \cdots, d$, can be represented as

$$v(\eta)_k = \quad (\prod_{j=1}^{d-k} \sin\theta_j) \quad \cos\theta_{d-k+1},$$

where $\theta_d = 0$ and $\prod_{j=1}^{0} \sin\theta_j = 1$ (see [19] page 645).

Thus, an $\mathbb{R}^{d-1}$ plane in $\mathbb{R}^d$ can be represented as follows. Given $a \in \mathbb{R}$, and $\eta \in \Theta \triangleq S^{d-1}$, where $S = \mathbb{R} \bmod \pi$, a plane can be defined as a function $\mathcal{P}$ maps $\mathbb{R} \times \Theta$ to a close set on $\mathbb{R}^d$, s.t. $\mathcal{P}(a, \eta) = \{y \in \mathbb{R}^d : \ y^T v(\eta) = a\}$. That is, $\mathcal{P}(a, \eta)$ is a plane that has normal vector $v(\eta)$ and has distance $|a|$ to the origin.

As we did in Subsection 3.2.2, we define a product metric on the plane space. The distance between two planes, $D(\ell_1, \ell_2)$, $\ell_j = \mathcal{P}(a_j, \eta_j) \in \mathbb{L}$, $j = 1, 2$, is defined as

$$|a_1 - a_2| + \sum_{k=1}^{d-1} d_S(\theta_{1k}, \theta_{2k}),$$

where $\theta_{jk} = (\eta_j)_k$, $j = 1, 2$. Thus we have a metric space $(\mathbb{P}^d, D)$.

Given data $\{z_i\}_{i=1}^n \subset \mathbb{R}^d$, the target function is

$$f(\ell) = \sum_{i=1}^{n} d(z_i, \ell)^\beta \tag{3.2.3}$$

for $\ell \in \mathbb{P}$. Note that $\mathbb{P}^d$ can also be defined as the collection of planes that passing through d linearly independent points. We would like to show that the minimizer of Function 3.2.3 will pass through at least d data points (Theorem 3.2.10). As before, we first need to show the minimizer of exists (Lemma 3.2.8 and 3.2.9).

Lemma 3.2.8. *Function 3.2.3 is continuous on* $(\mathbb{P}^d, D)$.

Proof. For any $z \in \mathbb{R}^d$, we only need to show that $d(z, \cdot)$ is continuous on the product space $(\mathbb{P}, D)$. Given a plane $\ell_{a,\eta} = \mathcal{P}(a, \eta)$ for $a \in \mathbb{R}$ and $\eta = (\theta_1, \cdots, \theta_{d-1}) \in \Theta^d$, since the distance $d(\cdot, \cdot)$ is considered to be the Euclidean distance, we have

$$\begin{aligned} d(z, \ell_{a,\eta}) &= d(z - av(\theta), \ell_{0,\eta}) \\ &= \|(z - av(\theta))^T \cdot v(\eta)\|, \end{aligned}$$

which is a sum of continuous functions of a and $\theta_1, \cdots, \theta_{d-1}$. Thus, $d(z, \cdot)$ is clearly continuous on $(\mathbb{P}, D)$. □

Lemma 3.2.9. *The minimizer of Function 3.2.3 exists.*

Proof. The proof is similar to Lemma 3.2.5. Define a d-dimensional ball $B_v(\bar{z})$, where $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$, and $v = \max \|z_i - \bar{z}\|$. Define $B \triangleq \{\ell \in \mathbb{P} | \ell \text{ passes through } cl(B_{2v}(\bar{z}))\}$. Thus, if $\ell \notin B$ then

$$f(\ell) > nv^\beta.$$

Next, if $\ell_{\bar{z}}$ is a plane passing through $\bar{z}$, we have

$$f(\ell_{\bar{z}}) \leq nv^\beta$$

for every possible direction. Thus we have

$$\min_{p \in B} f(\ell) \leq nv^\beta \leq \min_{p \notin B} f(\ell).$$

This implies

$$\min_{p\in\mathbb{P}} f(\ell) \quad = \min_{p\in B} \quad f(\ell),$$

which means we can restrict our search of planes that passes in B, i.e. a compact set in $(\mathbb{P}, D)$. This, along with the fact that the Function 3.2.2 is continuous, the minimizer is always existed. □

Theorem 3.2.10. *Assume $\beta \in (0,1)$ and $span(z_{1:n}) = \mathbb{R}^d$, the minimizer of Function 3.2.3 interpolates.*

Proof. Assuming there is no data points that lies on the optimal plane ℓ. We may assume $0 \in \ell$ and have normal vector $(1, 0, \cdots, 0)$ since we can shift and rotate the whole data to achieve it. Next, rewrite our target function by restricting the domain of planes that passing through 0 and allowing only the first two coordinates of their normal vectors to move. That is, for some $\theta \in [0, 2\pi)$, let $\ell_{0,\theta}$ to be the plane that passes 0 and has normal vector $r_\theta \triangleq (\cos\theta, \sin\theta, 0, \cdots, 0)$, we search the argument of the minimum of the following function

$$g(\theta) \quad = \quad \sum_{i=1}^{n} (d(z_i, \ell_{0,\theta})^\beta.$$

And we know that $\theta = \frac{\pi}{2} + k\pi$, $k \in \mathbb{N}$, which corresponds to normal vector $(1, 0, \cdots, 0)$, minimizes the function. The following argument will show that if the optimal plane doesn't touch any data point, the Hessian of g, say H_g, will be negative when $\theta = \frac{\pi}{2} + k\pi$, and we will get the contradiction.

First, assume that when $\theta = \frac{\pi}{2} + k\pi$, the plane $\ell_{0,\theta}$ doesn't touch any data point.

For $i = 1, \cdots, n$, write z_i as $(x_i, y_i, z_{i,3} \cdots, z_{i,d}) \in \mathbb{R}^d$, and calculate

$$\begin{aligned} d(z_i, \ell_{0,\theta}) &= |z_i \cdot r_\theta| \\ &= |x_i \cos\theta + y_i \sin\theta| \\ &= \sqrt{x_i^2 + y_i^2}|\frac{x_i}{\sqrt{x_i^2 + y_i^2}} \cos\theta + \frac{y_i}{\sqrt{x_i^2 + y_i^2}} \sin\theta| \\ &= \alpha_i |\cos(\theta - \theta_i)|, \end{aligned}$$

where $\alpha_i \triangleq \sqrt{x_i^2 + y_i^2} > 0$, and θ_i satisfies $\cos\theta_i = \frac{x_i}{\sqrt{x_i^2+y_i^2}}$ and $\sin\theta_i = \frac{y_i}{\sqrt{x_i^2+y_i^2}}$.

Note that $d(z_i, \ell_{0,\theta})$ is twice differentiable only when $\theta - \theta_i \neq \frac{\pi}{2} + k\pi$. Let's establish this by the following claim:

Claim 3.2.11. For $i = 1, \cdots, n$, $d(z_i, \ell_{0,\theta})$ is twice differentiable when $\theta = \frac{\pi}{2} + k\pi$.

Proof. Given any $i = 1, \cdots, n$, since we assume the optimal plane doesn't touch any of the data point, the vector $z_i = z_i - 0$ will have the property that z_i is not perpendicular to the optimal plane's normal vector $(1, 0, \cdots, 0)$, i.e.

$$z_i \cdot (1, 0, \cdots, 0) \neq 0,$$

which implies $x_i \neq 0$, and this means $\cos\theta_i \neq 0$, i.e. $\theta_i \neq \frac{\pi}{2} + k\pi$. So we conclude that $d(z_i, \ell_{0,\theta})$ is twice differentiable when θ is in the neighborhood of $\frac{\pi}{2} + k\pi$. □

Let $g_i(\theta) \triangleq (d(z_i, \ell_{0,\theta})^\beta$ for $i = 1, \cdots, n$, we can calculate $g_i's$ derivative on the

neighborhood of $\theta = \frac{\pi}{2} + k\pi$.

$$\begin{aligned}\frac{\partial g_i(\theta)}{\partial(\theta)} &= \frac{\partial}{\partial(\theta)}\alpha_i^{\beta}|\cos(\theta-\theta_i)|^{\beta} \\ &= \alpha_i^{\beta}\beta|\cos(\theta-\theta_i)|^{\beta-1}\cdot[-\sin(\theta-\theta_i))\mathrm{sgn}(\cos(\theta-\theta_i)].\end{aligned}$$

And we can calculate the Hessian of g_i when θ is in the neighborhood of $\frac{\pi}{2} + k\pi$:

$$\begin{aligned}H_{g_i}(\theta) &= \alpha_i^{\beta}\beta\{(\beta-1)\cdot|\cos(\theta-\theta_i)|^{\beta-2}\cdot[-\sin(\theta-\theta_i)\mathrm{sgn}(\cos(\theta-\theta_i))]^2 \\ &\quad +|\cos(\theta-\theta_i)|^{\beta-1}\cdot[(-\cos(\theta-\theta_i))\mathrm{sgn}(\cos(\theta-\theta_i))+0]\} \\ &= \alpha_i^{\beta}\beta\{(\beta-1)\cdot|\cos(\theta-\theta_i)|^{\beta-2}\sin^2(\theta-\theta_i)+|\cos(\theta-\theta_i)|^{\beta-1}\cdot[-|\cos(\theta-\theta_i)|]\} \\ &= \alpha_i^{\beta}\beta\{(\beta-1)\cdot|\cos(\theta-\theta_i)|^{\beta-2}\sin^2(\theta-\theta_i)-|\cos(\theta-\theta_i)|^{\beta}]\} \\ &< 0\end{aligned}$$

when $\theta = \frac{\pi}{2} + k\pi$. So we have

$$H_g(\theta) \quad = \sum_{i=1}^{n} \quad H_{g_i}(\theta) < 0$$

when $\theta = \frac{\pi}{2} + k\pi$. This means the plane that passes 0 with the normal vector $(1, 0, \cdots, 0)$ is not the optimal plane. Thus we have the contradiction. There is at least one data point touches the optimal plane.

Next, say the data point that touches the optimal plane is z_1, we may assume that $z_1 = 0$, if not, we can shift the data to achieve that. As before, we may assume the optimal plane touches 0 and has normal vector $(1, 0, \cdots, 0)$. Then, again $\theta = \frac{\pi}{2} + k\pi$ should minimize

$$g^{(1)}(\theta) = \sum_{i\in\{1,\cdots,n|z_i\neq z_1\}} (d(z_i, \ell_{0,\theta})^{\beta}.$$

Note that we excludes the data points that is equal to 0 so that g is twice differentiable when $\theta = \frac{\pi}{2} + k\pi$. Using the same argument, we conclude that the optimal plane have to pass through another data point that is not equal to 0, say z_2.

Finally, assume the optimal plane only go through q distinct data points, $1 < q < d$, say $z_{1:q}$ and $z_1 = 0$. And the normal vector of the optimal plane in the space of

$$span(z_2, \cdots, z_q)^{\perp}.$$

For $j = 1, \cdots, d$, define $e_j \in \mathbb{R}_1^d$, such that all the components of e_j are 0 except the j^{th} component is 1. And we may assume

$$\begin{aligned} span(z_2, \cdots, z_q)^{\perp} &= span(e_{d-q+2}, e_{d-q+3}, \cdots, e_d)^{\perp} \\ &\triangleq A_q \end{aligned}$$

and the optimal plane's normal vector is e_1, since we can rotate the whole data to achieve that. Thus, if we restrict the plane that has normal vector in A_q, we should find the plane with normal vector e_1 minimize the target function. In another words, if we restrict the plane that has normal vector in $\{\ell_{0,\theta}\}_{\theta \in [0,2\pi)}$, $\theta = \frac{\pi}{2} + k\pi$ should minimize

$$g^{(q)}(\theta) \quad = \sum_{i \in \{q+1, \cdots, n | z_i \neq z_{1:q}\}} (d(z_i, p_{0,\theta})^{\beta}.$$

Note that $\ell_{0,\theta}$ has normal vector

$$(\cos\theta, \sin\theta, 0, \cdots, 0) \in A_q.$$

Then, again, the same argument will give us the contradiction. One must understand the argument is not valid when $q = d$, since

$$span(e_2, e_{d-q+3}, \cdots, e_d)^{\perp}$$

is the space that span by only one vector, e_1. So there is no freedom to move the normal vector to get smaller value on the target function.

Thus we have proved that the optimal plane must pass at least d data points. $\square$

Combine with Theorem 3.2.3 and Theorem 3.2.6, let's sum up the important result with the following corollary:

Corollary 3.2.12. *Given data $\{z_i\}_{i=1}^n \subset \mathbb{R}^d$ with $span(z_{1:n}) = \mathbb{R}^d$, $d \geq 1$ and assume the model class $\mathcal{M} = \mathbb{P}^d$, the space of $\mathbb{R}^{d-1}$ plane in $\mathbb{R}^d$ (which can be determined by d linear independent points in $\mathbb{R}^d$). If $\beta \in (0, 1)$, then there is a minimizer $L^* = \{\ell_1^*, \cdots, \ell_m^*\} \in \mathbb{P}^d$ of function*

$$f(L) = |L|^2 + \frac{\lambda}{n} \sum_{i=1}^{n} \min_{\ell \in L} d(z_i, \ell)^{\beta}, \tag{3.2.4}$$

for $L \subset \mathbb{P}^d$, such that ℓ_j^* passes through at least d linear independent data points of $\{z_i\}_{i=1}^n$ for all $j = 1, 2, \cdots, m$. i.e. the minimizer is interpolating.

3.2.4 Non-interpolating Optima: Some Partial Results

What can be said about the minimizers of 3.2.4 when $\mathcal{M} = \mathbb{P}^k$ ($k < d$, d is the dimension of the data) is the set of planes of dimension $k-1$ (or say, the planes that can be determined by k linear independent points in $\mathbb{R}^d$)?

In general, the solutions will be non-interpolating. But we can get a sense of what will happen for β small by observing that when $\beta = 0$, $d(z_i, \ell)^\beta$ becomes binary[3], zero or one, and equal to zero if and only if z_i lies on the plane ℓ. Since, generically, at most k members of $\vec{z}$ will lie on a $k-1$ dimensional plane, the minimum cost over $\mathbb{P}^k$ is $n-k$, achieved by every interpolating plane in $\mathbb{P}^k$.

With this observation in mind, we introduce the notion of an "ϵ-interpolating" plane:

Definition 3.2.13. Fix $\epsilon > 0$. A plane $\ell \in \mathbb{P}^k$ is ϵ-interpolating if $\{z_i \in \vec{z} : d(z_i, \ell) < \epsilon\}$ contains at least k linearly independent elements of $\vec{z}$.

From a computational viewpoint, knowing that the optimal solution is ϵ-interpolating suggests building a search algorithm based on the polynomial (versus exponential) problem of finding an optimal interpolating plane. Although we have no guarantees, the following theorem further supports this approach:

Theorem 3.2.14. *For any $d \geq 2$, $k \leq d$, and data $\{z_i\}_{i=1}^n \subset \mathbb{R}^d$ with span$(z_{1:n}) = \mathbb{R}^d$, let $L = \{\ell_1, \ell_2, \ldots, \ell_m\}$ be a minimizer of Function 3.2.1 over $\mathbb{P}^k$. Fix $\epsilon > 0$ and*

[3]Here we set $0^0 = 0$

choose

$$\beta < \frac{\log\left(\frac{n-k+1}{n-k}\right)}{\log\left(\frac{2\sqrt{n}}{\epsilon}\right)}$$

Then ℓ_j is ϵ-interpolating for every $j = 1, \ldots, m$.

Proof. According to the result from section 3.2.1, it is enough to consider the target function

$$f(\ell) \;=\; \sum_{i=1}^{n} d(z_i, \ell)^\beta$$

over $\ell \in \mathbb{P}^k$. Recall that $z_1, \ldots, z_n$ is standardized: $\sum_{i=1:n} z_i = 0$ and $\sum_{i=1:n} \|z_i\|^2 = n$. Consequently, $\|z_i\|^2 \leq n, \ \forall i$ and it follows that $\|z_i - z_j\|^2 \leq 4n$ for all $1 \leq i, j \leq n$. Let $\tilde{\ell}$ be any interpolating plane in $\mathbb{P}^k$. The distance from any z_i to $\tilde{\ell}$ can not be bigger than its distance to any point on $\tilde{\ell}$. But $\tilde{\ell}$ is interpolating and includes at least one data point, z_j. Since $\|z_i - z_j\|^2 \leq 4n$, it follows that $d(z_i, \tilde{\ell}) \leq 2\sqrt{n}$.

Now, since $\tilde{\ell}$ contains at least k data points,

$$\sum_{i=1}^{n} d(z_i, \tilde{\ell})^\beta \leq (n-k)(2\sqrt{n})^\beta$$

In contradiction to the claim in the theorem, assume that ℓ is not ϵ-interpolating. Then there are at least $n - k + 1$ data points that have distance from ℓ at least as great as ϵ:

$$\sum_{i=1}^{n} d(z_i, l)^\beta \geq (n-k+1)\epsilon^\beta$$

Now check that

$$\beta < \frac{\log\left(\frac{n-k+1}{n-k}\right)}{\log\left(\frac{2\sqrt{n}}{\epsilon}\right)}$$

implies that $(n-k)(2\sqrt{n})^\beta$ is smaller than $(n-k+1)\epsilon^\beta$, which means that

$$\sum_{i=1}^{n} d(z_i, \tilde{\ell})^\beta < \sum_{i=1}^{n} d(z_i, \ell)^\beta$$

and ℓ can not be optimal. □

In the following paragraphs we explore two further special cases: clustering to points in $d \geq 2$ dimensions and clustering to lines in $d \geq 3$ dimensions. In both cases we will identify a range of values for β within which optimal solutions are guaranteed to contain at least one data point.

Clustering to points in $\mathbb{R}^d$, $d \geq 2$.

When $\mathcal{M} = \mathbb{R}^d$, $d > 1$, i.e. the point model, and $m = 1$, given data $\{z_i\}_{i=1}^n \subset \mathbb{R}^d$, the target function can be written as

$$f(x) \quad = \quad \sum_{i=1}^{n} d(z_i, x)^\beta, \tag{3.2.5}$$

where $x \in \mathbb{R}^d$. Note that the optimal point is not necessary to be some of the data points. We say the optimizer is non-interpolating. One can consider three points in $\mathbb{R}^2$ that forms an equilateral triangle, the optimal point is the center of the triangle when β close to 1. The next lemma will state the fact that every data point is a local

minimum of f.

Lemma 3.2.15. *Given data* $\{z_i\}_{i=1}^n \subset \mathbb{R}^d$. *For* $i = 1, 2, \cdots, n$, z_i *is the strictly minimum point of Function 3.2.5 on the* d *dimensional ball center at* z_i *with radius*

$$\min\{\frac{1}{2}\min_{j:z_j \neq z_i} ||z_j - z_i||, \frac{1}{2}(\frac{k_i(1-\beta)}{\sum_{j:z_j \neq z_i} \frac{1}{||z_j - z_i||^{2-\beta}}})^{\frac{1}{2-\beta}}\},$$

where $k_i = \sum_{j=1}^n \mathbb{I}_{\{z_j = z_i\}}$.

Proof. It is sufficient to just consider z_1 is a local minimum. May assume $z_1 = 0$. For any radius $r > 0$, set a point $z^{(r)} = ra$, $a \in \mathbb{R}_1^d$, such that $\|z^{(r)}\| = r$. Then

$$f(z^{(r)}) \quad = \quad \sum_{i=1}^n \sqrt{(z_i - z^{(r)})^T (z_i - z^{(r)})}^{\beta}.$$

For abbreviation, fix an $a \in \mathbb{R}_1^d$ let

$$\begin{aligned} g(r) &\triangleq f(z^{(r)}) \\ g_i(r) &\triangleq \sqrt{(z_i - z^{(r)})^T (z_i - z^{(r)})}. \end{aligned}$$

We will show that when r is within the radius, we have negative Hessian of g. So for any such a point $z^{(r)}$, there is a direction (toward 0) such that it produce a negative value of the directional second derivative, i.e. $z^{(r)}$ is not a local minimum.

When $z^{(r)} = ra \neq z_i$, calculate

$$\begin{aligned} g_i'(r) &= \frac{1}{2g_i}\{-2 \cdot (z_i - z^{(r)})^T a\} \\ &= \frac{1}{g_i}\{(z^{(r)} - z_i)^T a\} \\ &= \frac{1}{g_i}\{r - {z_i}^T a\} \end{aligned}$$

since $z^{(r)T} a = ra^T a = r$. Thus

$$\frac{\partial}{\partial} g_i^\beta(r) = \beta g_i^{\beta-2}\{r - {z_i}^T a\}$$

whenever $ra \neq z$. Thus, we can calculate the Hessian of g_i^β

$$\begin{aligned} H_{g_i^\beta}(r) &= \beta\{(\beta-2) g_i^{\beta-3} \frac{1}{g_i}(r - {z_i}^T a)^2 + g_i^{\beta-2}\} \\ &= \beta\{(2-\beta)\frac{-1}{g_i^{4-\beta}}(r - {z_i}^T a)^2 + \frac{1}{g_i^{2-\beta}}\}. \end{aligned}$$

Note that $g_1(r) = r$. So for those i such that $z_i = z_1$

$$\begin{aligned} H_{g_i^\beta}(r) &= \beta\{(2-\beta)\frac{-1}{r^{4-\beta}}(r)^2 + \frac{1}{r^{2-\beta}}\} \\ &= \frac{-\beta(1-\beta)}{r^{2-\beta}}. \end{aligned}$$

And for i such that $z_i \neq z_1$, let $r_i \triangleq \|z_i\|$, we have

$$H_{g_i^\beta}(r) \leq \frac{\beta}{g_i^{2-\beta}} \leq \frac{\beta}{|r_i - r|^{2-\beta}},$$

where the second inequality is because of

$$g_i(r) = d(z_i, ra) \geq |d(z_i, 0) - d(0, ra)| = |r_i - r|.$$

Furthermore, since $r < \frac{r_i}{2}$, for any i with $z_i \neq z_1$, $r_i - r > \frac{r_i}{2}$, which means

$$H_{g_i^\beta}(r) \quad < \quad \beta(\frac{2}{r_i})^{2-\beta}.$$

Put these estimations together, we get

$$H_g(r) \quad < \quad \beta\{\sum_{i:z_i \neq z_1} (\frac{2}{r_i})^{2-\beta} - \frac{k_1(1-\beta)}{r^{2-\beta}}\}.$$

So, if $r < \frac{1}{2}(\frac{k_1(1-\beta)}{\sum_{i:z_i \neq z_1} \frac{1}{r_i}^{2-\beta}})^{\frac{1}{2-\beta}}$, we have

$$r^{2-\beta} \quad < \quad \frac{k_1(1-\beta)}{\sum_{i:z_i \neq z_1} (\frac{2}{r_i})^{2-\beta}} = (\frac{1}{2})^{2-\beta} \frac{k_1(1-\beta)}{\sum_{j:z_j \neq z_i} (\frac{1}{r_j})^{2-\beta}}$$

Thus we have $H_g(r) < 0$ when

$$r < \min\{\frac{1}{2} \min_{j:z_j \neq z_1} r_j, \frac{1}{2}(\frac{k_1(1-\beta)}{\sum_{j:z_j \neq z_i} (\frac{1}{r_j})^{2-\beta}})^{\frac{1}{2-\beta}}\}.$$

This completes the proof. □

Lemma 3.2.16. *Given data $\{z_i\}_{i=1}^n \subset \mathbb{R}^d$, if $d(z_i, z_j) > q$ for any $i \neq j$ s.t. $z_i \neq z_j$, then z_i, $i = 1, 2, \cdots, n$ is the strictly minimum point on the d dimensional ball center*

at z_i with radius

$$\alpha(n,q,\beta) \;=\; \begin{cases} \frac{q}{2}(\frac{1-\beta}{n-1})^{\frac{1}{2-\beta}} & ,\ for\ n>2 \\ \frac{q}{2} & ,\ o.w. \end{cases}.$$

Proof. Note that $n \le 2$ is a trivial case. Set $n > 2$, we can estimate the following for every $1 = 1, 2, \cdots, n$:

$$\begin{aligned} \frac{k_i(1-\beta)}{\sum_{j:z_j\neq z_i}\frac{1}{||z_j-z_1||^{2-\beta}}} &= k_i(1-\beta)(\sum_{j:z_j\neq z_i}\frac{1}{||z_j-z_i||^{2-\beta}})^{-1} \\ &\ge k_i(1-\beta)(\sum_{j:z_j\neq z_i}\frac{1}{q^{2-\beta}})^{-1} \\ &= k_i(1-\beta)(\frac{n-k_i}{q^{2-\beta}})^{-1} \\ &= \frac{k_i(1-\beta)}{n-k_i}q^{2-\beta} \\ &\ge \frac{1-\beta}{n-1}q^{2-\beta} \end{aligned}$$

for any $i = 1, 2, \cdots, n$. Thus, we have for $n \ge 2$,

$$\frac{q}{2}(\frac{1-\beta}{n-1})^{\frac{1}{2-\beta}} \;\le\; \frac{1}{2}(\frac{k_i(1-\beta)}{\sum_{j:z_j\neq z_i}\frac{1}{||z_j-z_i||^{2-\beta}}})^{\frac{1}{2-\beta}}.$$

Also, since $(\frac{1-\beta}{n-1})^{\frac{1}{2-\beta}} < 1$, we have

$$\frac{q}{2}(\frac{1-\beta}{n-1})^{\frac{1}{2-\beta}} < \frac{q}{2} \le \frac{1}{2}\min_{j:z_j\neq z_i}||z_j-z_i||$$

for any $i = 1, 2, \cdots, n$. So

$$\frac{1}{2}(\frac{1-\beta}{n-1}q)^{\frac{1}{2-\beta}} \le \min\{\frac{1}{2}\min_{j:z_j\neq z_i}||z_j-z_i||, \frac{1}{2}(\frac{k_i(1-\beta)}{\sum_{j:z_j\neq z_i}\frac{1}{||z_j-z_i||^{2-\beta}}})^{\frac{1}{2-\beta}}\}$$

for any $i = 1, 2, \cdots, n$. This completes the proof. $\square$

One quick observation of the function $\alpha(n, q, \beta)$ is that it is increasing when β gets smaller. So that the radius will get bigger when we lower the value of β, until the radius hits $\frac{1}{2}\min_{j:z_j \neq z_i} \|z_j - z_i\|$.

Even though in $\mathbb{R}^d$, $d \geq 2$, the best point that minimize Function 3.2.5 might not be one of the data, we can "force" the optimal point to be one of the data by choosing a more sparse distance, i.e. smaller β. One can say that a sparser distance tends to creat an interpolating optimizer. The next theorem states this fact.

Theorem 3.2.17. *For data $\{z_i\}_{i=1}^n \subset \mathbb{R}^d$, with $q, s \in \mathbb{R}$ s.t. $q < \|z_i - z_j\| < s$ for all $i \neq j = 1, 2, \cdots, n$. Given any $\beta_0 \in (0, 1)$, if*

$$\beta \leq \min\{\beta_0, \beta_1(\beta_0, n, s, q)\},$$

where

$$\beta_1(\beta_0, n, s, q) \triangleq \begin{cases} \log(\frac{n}{n-1})/\log(\frac{2s(n-1)}{q(1-\beta_0)} & , \text{ for } n > 2 \\ \beta_0 & , \text{ for } n \leq 2, \end{cases}$$

then the global minimum points of Function 3.2.5 are among the data point.

Furthermore, for $n \geq 2$ if $\bar{\beta} = \bar{\beta}(n, s, q)$ satisfies the equation

$$\bar{\beta} = \log(\frac{n}{n-1})/\log(\frac{2s(n-1)}{q(1-\bar{\beta})}),$$

then

$$\bar{\beta} \geq \min\{\beta_0, \log(\frac{n}{n-1})/\log(\frac{2s(n-1)}{q(1-\beta_0)})\}$$

for every $\beta_0 \in (0,1)$.

Proof. Note that $n \leq 2$ is a trivial case. Set $n > 2$, given any $x \in \mathbb{R}^d$ s.t. $x \notin \{z_i\}_{i=1}^n$ and is a local minimum point of f, the Function 3.2.5. Note that for $\alpha = \alpha(n, q, \beta)$ that defined in the previous lemma, i.e.

$$\alpha(n, q, \beta) \triangleq \frac{q}{2}(\frac{1-\beta}{n-1})^{\frac{1}{2-\beta}}.$$

So x will not be located inside any d dimensional ball $B_\alpha(z_i)$ for any $i = 1, 2, \cdots, n$ because of the previous lemma. Thus we have

$$f(x) \;\; \geq \;\; n\alpha^\beta.$$

Furthermore, given $\beta_0 \in (0,1)$, for any $\beta < \beta_0$, we have $1 - \beta > 1 - \beta_0$, which implies

$$\alpha(n, q, \beta) \;\; \geq \;\; \frac{q}{2}(\frac{1-\beta_0}{n-1})^{\frac{1}{2-\beta}}.$$

We can further estimate it by

$$\begin{aligned} \alpha^\beta \;\; &\geq \;\; (\frac{q}{2})^\beta (\frac{1-\beta_0}{n-1})^{\frac{\beta}{2-\beta}} \\ &> \;\; (\frac{q}{2})^\beta (\frac{1-\beta_0}{n-1})^{\beta} \\ &= \;\; (\frac{q(1-\beta_0)}{2(n-1)})^{\beta}, \end{aligned}$$

where the second inequality is because of $\frac{\beta}{2-\beta} < \beta$ and $\frac{1-\beta_0}{n-1} < 1$.

On the other hand, we have

$$\min_i f(z_i) \quad \leq \quad (n-1)s^\beta$$

for any $i = 1, 2, \cdots, n$ because s is the maximum length of two data points.

Thus, if

$$\begin{aligned} \beta &\leq \log(\frac{n}{n-1})/\log(\frac{2s(n-1)}{q(1-\beta_0)} \\ &= \log(\frac{n-1}{n})/\log(\frac{q(1-\beta_0)}{2s(n-1)}) \end{aligned}$$

we have

$$(\frac{q(1-\beta_0)}{2s(n-1)})^\beta \quad \geq \quad \frac{n-1}{n},$$

note that $\log(\frac{q(1-\beta_0)}{2s(n-1)}) < 0$. This means

$$f(x) \geq n\alpha \geq n(\frac{q(1-\beta_0)}{2(n-1)})^\beta \quad \geq \quad (n-1)s^\beta \geq \min_i f(z_i).$$

So we finish the proof of the first part.

The second part of the theorem is valid since

$$\log(\frac{n}{n-1})/\log(\frac{2s(n-1)}{q(1-\beta_0)})$$

is increasing when β_0 goes to 0. □

The next lemma explains a property of the bound.

Lemma 3.2.18. *Given any $\beta_0 \in (0,1)$, the function $\beta_1(\beta_0, n, s, q)$ or $\bar{\beta}(n, s, q)$ is non-increasing when n increases.*

Proof. The scale invariance is easy to see, since any scale imposed on the data will not affect the value $\frac{s}{q}$. For the first property, we only need to look at $n > 2$ case. Defining

$$\begin{aligned} \eta(x) &= \beta_1(\beta_0, x, s, q) \\ &= \frac{\log x - \log(x-1)}{\log(x-1) + c}, \end{aligned}$$

where $c = \log \frac{2s}{q(1-\beta_0)} > 0$. And calculate for $x > 2$

$$\begin{aligned} \eta'(x) &\propto (\log(x-1)+c)(\frac{1}{x} - \frac{1}{x-1}) - \frac{1}{x-1}(\log x - \log(x-1)) \\ &= (\log(x-1)+c)\frac{1}{x} - \frac{c}{x-1} - \frac{1}{x-1}\log x \\ &= \frac{\log(x-1)+c}{x} - \frac{\log x + c}{x-1} \\ &< 0 \end{aligned}$$

So, for $n > 2$, $\beta_1(\beta_0, n, s, q)$ is decreasing when n increases. □

Clustering to lines in $\mathbb{R}^d$, $d \geq 3$.

So, when β is small enough, the optimizer of the target function on the point model will be interpolating. An immediate generalization can be derived on the line model:

when β is small enough, the optimal line will pass though at least one data point. This will be stated in the next theorem.

Theorem 3.2.19. *For $\mathbb{R}^d$ data, $d \geq 3$, with the line model. Given $\beta_0 \in (0,1)$ and data $\{z_i\}_{i=1}^n$, the the global minimum line of the function*

$$f(\ell) \quad = \quad \sum_{i=1}^{n} d(z_i, \ell)^{\beta}$$

will pass at least one data point when

$$\beta \leq \begin{cases} \beta_1(\beta_0, n, s, q) & , \ n > 3 \\ \beta_0 & , \ o.w. \end{cases}$$

Proof. The $n \leq 3$ is a trivial case. For $n > 3$ and given $\beta_0 \in (0,1)$, assume the optimal line, say $\ell^\star$, doesn't pass any data point. So the target function has value

$$f(\ell^\star) = \sum_{i=1}^{n} d(z_i, \ell^\star)^{\beta}.$$

May assume $\ell^\star = \mathcal{L}(0, r)$, i.e. it passes 0 with direction $r \in \mathbb{R}_1^d$. We can project all the data point to the plane P_r that is perpendicular to r and passes 0, say $\{\hat{z}_i(r)\}_{i=1}^n$. Then

$$f(\ell^\star) \quad = \quad \sum_{i=1}^{n} d(\hat{z}_i(r), 0)^{\beta},$$

which is equivalent to the point model in $\mathbb{R}^{d-1}$. Lemma 3.2.16 says that when

$$\beta \quad \leq \quad \beta_1(\beta_0, n, s, q),$$

a data point (note that no data point is 0), say z_1, will be the minimum point of the function

$$h(z) = \sum_{i=1}^{n} d(\hat{z}_i(r), z)^{\beta}.$$

So, we can move $\ell^{\star}$ parallelly (so that it is still the normal vector of plane P_r) such that it touches z_1. We thus find a better line $\mathcal{L}(z_1, r)$ that have lower target function value. So we have the first contradiction, the optimal line must pass at least one data point. $\square$

3.3 Observations about Sparse-distance Clustering in $\mathbb{R}^1$

In this Section we examine the minimizer of the function

$$f(\ell) = \sum_{i=1}^{n} d(z_i, \ell)^\beta$$

when $\ell \in \mathbb{R}$. The intention is to understand the simplest case so that one could have some general ideas of sparse distance.

Suppose our data $\boldsymbol{z} = \{z_i\}_{i=1}^n \subset \mathbb{R}$ is generated from some distribution $F(\cdot - \xi)$, where $\xi \in \mathbb{R}$ is a location parameter. To estimate the location parameter. We can define the set $[T_n(z)]$ as

$$[T_n(\boldsymbol{z})] = \{\theta \in \mathbb{R}^1 : \sum_{i=1}^{n} |z_i - \theta|^\beta \text{ reaches minimum}\}, \tag{3.3.1}$$

and then define $T = T_n = T_n(\boldsymbol{z})$ as mean of the set $[T_n(\boldsymbol{z})]$. The definition is similar to the M-estimator from Huber [20]. The existence of the estimator for any given data set is easy to verify, since T_n must be within

$$[\min_{i=1:n} z_i, \max_{i=1:n} z_i],$$

then we are solving a minimizing problem of a continuous function on a compact set. Thus the T_n exists. Another easy but important property is that T_n is location invariant, i.e. $T_n(\boldsymbol{z} + c) = T_n(\boldsymbol{z}) + c$ for some constant c.

3.3.1 Asymptotic Properties

However, the estimator T_n does not necessary converge to a point when $\beta \leq 1$, unlike $\beta > 1$ case, which is easy to verify under some regular conditions on F. To quickly see this, note that one can obtain the limit of the function $\frac{1}{n}\sum_{i=1}^{n}|z_i - \theta|^\beta$ asymptotically under some condition of strong law of large number (SLLN), this function will converge to $\phi(\theta) = E(|Z_1 - \theta|^\beta)$ a.e. It is easy to verify that when $\beta > 1$, $\phi(\theta)$ is a strictly convex function under some regular condition on distribution F, so T_n exist and unique in this case. For the reader that when $\beta = 1$, some choice of F may give multiple solutions, and the estimator might not be convergent. For example, think about the distribution with density

$$\frac{1}{2}(\mathbb{I}_{\{x=-1\}} + \mathbb{I}_{\{x=1\}}),$$

then T_n will take values $-1, 0$, and 1 infinite often and won't converge. Thus, to have a unique asymptotic solution, one must impose some conditions on the distribution F.

When $\beta \in (0, 1)$, it's not hard to see that $[T_n]$ can only contain points in $\{z_i\}_{i=1}^n$. Since the Hessian of $|z_i - \theta|^\beta$ (as a function of θ) is

$$\beta(\beta - 1)|z_i - \theta|^{\beta-2} < 0,$$

when $\theta \neq z_i$, $i = 1, 2, \cdots, n$, which results in negative Hessian of target function for any $\theta \notin \{z_i\}_{i=1}^n$.

The following lemma shows that if density of F is symmetric and unimodal, then T_n is a consistent estimator of location parameter under some mild conditions.

Lemma 3.3.1. *When $\beta \in (0,1)$, given $\{x_i\}_{i=1}^n$ i.i.d. from a continuous distribution $F(\cdot - \xi)$. Assume the density g of distribution F satisfies*

1. *g is symmetric and unimodal at 0.*
2. *g' exist and continuous,*
3. *$E_g|X| < \infty$, $E_g|X|^{\beta-1} < \infty$*

then $T_n(\boldsymbol{x})$ converge to ξ a.e.

Proof. May assume $\xi = 0$. Since the elements of $[T_n]$ satisfies $\min_\theta \sum_{i=1}^n |z_i - \theta|^\beta$, it satisfies

$$\min_\theta \frac{1}{n} \sum_{i=1}^n |z_i - \theta|^\beta$$

also. To use SLLN, calculate

$$\begin{aligned} E(|Z_1 - \theta|^\beta) &= \int_{-\infty}^{\infty} |x-\theta|^\beta g(x)dx \\ &= \int_{\theta-1}^{\theta+1} |x-\theta|^\beta g(x)dx + \int_{\mathbb{R}\setminus(\theta-1,\theta+1)} |x-\theta|^\beta g(x)dx \\ &\leq \int_{\theta-1}^{\theta+1} g(x)dx + \int_{\mathbb{R}\setminus(\theta-1,\theta+1)} |x-\theta| g(x)dx \\ &< \infty \end{aligned}$$

because of Assumption 3. So we can use SLLN to conclude

$$\frac{1}{n} \sum_{i=1}^n |z_i - \theta|^\beta \quad \to \quad E(|Z_1 - \theta|^\beta),$$

a.e. Thus the elements in the limit of $[T_n]$ are arguments of minimum of

$$\psi(\theta) \triangleq E(|Z_1 - \theta|^\beta).$$

Observe that ψ is a even function since Z_1 and $-Z_1$ have the same distribution.

Next, we will show that ψ has only one global minimum at 0, which implies $T_n \to 0$ a.e. Note that

$$\psi(\theta) = \int_{-\infty}^{\infty} |x - \theta|^\beta g(x)dx.$$

First we have to argue that $\psi'(\theta)$ is existent. Since

$$\begin{aligned}\psi'(\theta) &= \beta \int |x - \theta|^{\beta-1} \cdot sgn(x - \theta) \cdot g(x)dx \\ &= -\beta \int_{-\infty}^{\theta} (\theta - x)^{\beta-1} g(x)dx + \beta \int_{\theta}^{\infty} (x - \theta)^{\beta-1} g(x)dx,\end{aligned}$$

and when $\theta > 0$, the second term is bounded from above by $\beta \int_0^\infty x^{\beta-1} g(x)dx$; when $\theta < 0$, the first term is bounded from below by $-\beta \int_{-\infty}^0 (-x)^{\beta-1} g(x)dx$. And the assumption 3 states that $E_g|X|^{\beta-1}$ exist and finite, which implies $\int_0^\infty (-x)^{\beta-1} g(x)dx$ and $\int_0^\infty x^{\beta-1} g(x)dx$ are both finite. This indicates that $\psi'(\theta)$ exist for all $\theta \in \mathbb{R}$.

Calculate

$$\begin{aligned}\psi(\theta) &= \int_{-\infty}^{\infty} |x - \theta|^\beta g(x)dx \\ &= \int_{-\infty}^{\infty} |y|^\beta g(y + \theta)dy.\end{aligned}$$

Thus

$$\begin{aligned}\psi'(\theta) &= \int_{-\infty}^{\infty} |y|^{\beta} g'(y+\theta)dy \\ &= \int_{-\infty}^{\infty} |x-\theta|^{\beta} g'(x)dx.\end{aligned}$$

Observe that $\psi'(0) = 0$ since g' is an odd function (because g is an even function from assumption 1). Also, we know that ψ' is an odd function. Continuing calculating ψ':

$$\psi'(\theta) = \int_{-\infty}^{0} |x-\theta|^{\beta} g'(x)dx + \int_{0}^{\infty} |x-\theta|^{\beta} g'(x)dx.$$

The first part can be estimated with the change of variable $y = -x$, then

$$\begin{aligned}\int_{-\infty}^{0} |x-\theta|^{\beta} g'(x)dx &= -\int_{\infty}^{0} |-y-\theta|^{\beta} g'(-y)dy \\ &= -\int_{0}^{\infty} |y+\theta|^{\beta} g'(y)dy.\end{aligned}$$

Thus

$$\begin{aligned}\psi'(\theta) &= -\int_{0}^{\infty} |x+\theta|^{\beta} g'(x)dx + \int_{0}^{\infty} |x-\theta|^{\beta} g'(x)dx \\ &= \int_{0}^{\infty} (|x-\theta|^{\beta} - |x+\theta|^{\beta}) g'(x)dx.\end{aligned}$$

When $\theta > 0$, because $g'(x) < 0$ when $x > 0$ (Assumption 1) and $|x-\theta| < |x+\theta|$, we

have $\psi'(\theta) > 0$. Combine the observation of ψ' being an odd unction, we have

$$\psi'(\theta) \text{ is } \begin{cases} > 0 & , \text{ when } \theta > 0 \\ < 0 & , \text{ when } \theta < 0. \end{cases}$$

Thus we have the result that ψ attend its global minimum at 0. □

3.3.2 Breakdown Point

Breakdown point is a way to measure the robustness of an estimator. First concept is proposed by Hampel [18]. Donoho and Huber [12] gives a simple version of it called the finite sample breakdown point, which we will adopt in this subsection. Given a sample $\boldsymbol{x} = \{x_i\}_{i=1}^n$, the finite sample breakdown points at sample $\boldsymbol{x}$ for T is defined as

$$BP(T, \boldsymbol{x}) \quad = \quad \min_k \{\frac{k}{n} : \sup_{x^k \in \mathcal{X}^k(\boldsymbol{x})} |T(\boldsymbol{x}) - T(\boldsymbol{x}^k)| = \infty\},$$

where

$$\mathcal{X}^k(\boldsymbol{x}) \quad \triangleq \quad \{\boldsymbol{y} \in \mathbb{R}^n : y_i = x_i \text{ except first } k \text{ elements.}\}$$

for $k < n$. The highest breakdown point for a location estimator is 50%.

We would like to show that when $\beta \in (0, 1)$, the estimator T has 50% breakdown

points. Let's first prepare some notations:

$$Q(\boldsymbol{x},\theta) \triangleq \sum_{i=1}^{n} |x_i - \theta|^{\beta}.$$

Note that all the local minimizer of $Q(\boldsymbol{x}, \cdot)$ are data points x_i, $i = 1, 2, \cdots, n$.

Let's introduce a lemma.

Lemma 3.3.2. *If $T(\boldsymbol{x}) = c \in \mathbb{R}$, and $\hat{\boldsymbol{x}} = \boldsymbol{x}$ except one of the element being changed to c, then*

$$T(\hat{\boldsymbol{x}}) = c$$

Proof. Note that since T is defined as midpoint of $[T]$, c may not necessary to be a data point. Assume the changed element is x_1, that is, $\hat{x}_1 = c$. Define

$$\begin{aligned} Q(\cdot) &= Q(\boldsymbol{x}, \cdot) \\ Q'(\cdot) &= Q(\hat{\boldsymbol{x}}, \cdot). \end{aligned}$$

Note that we have $Q(x) \geq Q(c)$ for any $x \in \hat{\boldsymbol{x}}$ since we assume $T(\boldsymbol{x}) = c$. We will split the proof into two claims: $T(\hat{\boldsymbol{x}}) = c$ when (1) $x_1 < c$. (2) $x_1 > c$.

Claim. When $x_1 < c$, $T(\hat{\boldsymbol{x}}) = c$.

Proof. We will to show that $Q'(x) > Q'(c)$ for any $x \in \hat{\boldsymbol{x}}$ s.t. $x \neq c$, so that the value of T is not changed. We will split the problem into 3 situations: (a) $x \leq x_1$, (b) $x \in (x_1, c)$ (c) $x > c$, where $x \in \hat{\boldsymbol{x}}$, which covers all the the possible minimum points in $\hat{\boldsymbol{x}}$.

(a) When $x \leq x_1$, calculate

$$\begin{aligned} Q'(x) &= Q(x) - |x_1 - x|^\beta + |c - x|^\beta \\ &> Q(x) \\ &\geq Q(x) - |c - x_1|^\beta \\ &\geq Q(c) - |c - x_1|^\beta \\ &= Q'(c). \end{aligned}$$

Note that the first inequality is because $x_1 < c$.

(b) When $x \in (x_1, c)$, calculate

$$\begin{aligned} Q'(x) &= Q(x) - |x_1 - x|^\beta + |c - x|^\beta \\ &\geq Q(x) - (x - x_1)^\beta \\ &> Q(x) - (c - x_1)^\beta \\ &\geq Q(c) - (c - x_1)^\beta \\ &= Q'(c). \end{aligned}$$

Note that the second inequality is because of $x_1 < x < c$.

(c) When $x > c$, calculate

$$\begin{aligned} Q'(x) &= Q(x) - |x_1 - x|^\beta + |c - x|^\beta \\ &= Q(x) - [(x - x_1)^\beta - (x - c)^\beta] \\ &> Q(x) - (c - x_1)^\beta \\ &\geq Q(c) - (c - x_1)^\beta \\ &= Q'(c), \end{aligned}$$

where the first inequality is because of $a^\beta - b^\beta \leq (a - b)^\beta$ if $a > b$. □

Claim. When $x_1 > c$, $T(\hat{\boldsymbol{x}}) = c$.

Proof. Again, we split the problem into 3 cases: (a) $x < c$, (b) $x \in (c, x_1)$, (c) $x > x_1$, where $x \in \hat{\boldsymbol{x}}$.

(a) When $x < c$, calculate

$$\begin{aligned} Q'(x) &= Q(x) - |x_1 - x|^\beta + |c - x|^\beta \\ &= Q(x) - [(x_1 - x)^\beta - (c - x)^\beta] \\ &> Q(x) - (x_1 - c)^\beta \\ &\geq Q(c) - (x_k^k - c)^\beta \\ &= Q'(c). \end{aligned}$$

(b) $x \in (c, x_1)$, calculate

$$\begin{aligned} Q'(x) &= Q(x) - |x_1 - x|^\beta + |c - x|^\beta \\ &> Q(x) - (x_1 - x)^\beta \\ &> Q(x) - (x_1 - c)^\beta \\ &\geq Q(c) - (x_1 - c)^\beta \\ &= Q'(c). \end{aligned}$$

(c) $x \geq x_1$, calculate

$$\begin{aligned} Q'(x) &= Q(x) - |x_1 - x|^\beta + |c - x|^\beta \\ &> Q(x) \\ &> Q(x) - (x_1 - c)^\beta \\ &\geq Q(c) - (x_1 - c)^\beta \\ &= Q'(c). \end{aligned}$$

□

□

Theorem 3.3.3. *When $\beta \in (0, 1)$, the finite sample breakdown point at data $\boldsymbol{x}$ estimator T is 50%.*

Proof. Given $\boldsymbol{x} = \{x_i\}_{i=1}^n$ and $\boldsymbol{x}^k \in \mathcal{X}^k(\boldsymbol{x})$, there is an $\hat{\boldsymbol{x}}^k \in \mathcal{X}^k(\boldsymbol{x})$ with $\hat{x}_i^k = \hat{x}_j^k$ for

$i, j = 1, 2, \cdots, k$ s.t.

$$T(\boldsymbol{x}^k) \;=\; T(\hat{\boldsymbol{x}}^k).$$

This is because of Lemma 3.3.2, i.e. we can change x_i^k, $i = 1, 2, \cdots, k$, to $T(\boldsymbol{x}^k)$ one point at a time and doesn't affect the estimator T. Define

$$\hat{\mathcal{X}}^k(\boldsymbol{x}) \;=\; \cup_{c\in\mathbb{R}}\hat{\boldsymbol{x}}^{k,c}$$

and $\hat{\boldsymbol{x}}^{k,c} \in \mathcal{X}^k(\boldsymbol{x})$ s.t. $\hat{x}_i^k = \hat{x}_j^k = c$ for $i, j = 1, 2, \cdots, k$. So

$$\begin{aligned} |T(\boldsymbol{x}) - T(\boldsymbol{x}^k)| \;&\leq\; \sup_{\hat{\boldsymbol{x}}^k\in\hat{\mathcal{X}}^k(\boldsymbol{x})} |T(\boldsymbol{x}) - T(\hat{\boldsymbol{x}}^k)| \\ &=\; \sup_{c\in\mathbb{R}} |T(\boldsymbol{x}) - T(\hat{\boldsymbol{x}}^{k,c})|. \end{aligned}$$

Since $\hat{\boldsymbol{x}}^{k,c} \in \mathcal{X}^k(\boldsymbol{x})$, we have

$$\sup_{\boldsymbol{x}^k\in\mathcal{X}^k(\boldsymbol{x})} |T(\boldsymbol{x}) - T(\boldsymbol{x}^k)| \;=\; \sup_{c\in\mathbb{R}} |T(\boldsymbol{x}) - T(\hat{\boldsymbol{x}}^{k,c})|,$$

where $\hat{\boldsymbol{x}}^{k,c} \in \mathcal{X}^k(\boldsymbol{x})$ and $\hat{x}_i^k = c$ for $i = 1, 2, \cdots, k$. Set $k < n/2$ and define

$$Q(\cdot) \;=\; Q(\hat{\boldsymbol{x}}^{k,c}, \cdot).$$

We may assume that for all $i = 1, \cdots, n$, $x_i \in (-\epsilon, \epsilon)$ for some $\epsilon > 0$, and $T(\boldsymbol{x}) = 0$. Calculate

$$Q(c) \;=\; \sum_{i=k+1}^{n} |c - x_i|^\beta > (n-k)|c-\epsilon|^\beta.$$

And

$$\begin{aligned} Q(0) &= k \cdot |c|^{\beta} + \sum_{i=k+1}^{n} |x_i|^{\beta} \\ &< k \cdot |c|^{\beta} + (n-k)\epsilon^{\beta}. \end{aligned}$$

Since $n-k > k$, there exist a $c_0 = c_0(\boldsymbol{x}, k) > 0$ s.t. $Q(0) < Q(c)$ whenever $|c| > c_0$. So the minimum point of Q will not be c when $|c| > c_0$, which means $|T(\boldsymbol{x}) - T(\hat{\boldsymbol{x}}^{k,c})|$ can only be among $(-\epsilon,\epsilon)$. And when $|c| \leq c_0$ we have $|T(\boldsymbol{x}) - T(\hat{\boldsymbol{x}}^{k,c})| < c_0 + \epsilon < \infty$, This implies

$$\sup_{\boldsymbol{x}^k \in \mathcal{X}^k(\boldsymbol{x})} |T(\boldsymbol{x}) - T(\boldsymbol{x}^k)| \;<\; \infty$$

when $k < n/2$. So

$$BP(T, \boldsymbol{x}) = \min_k \{\frac{k}{n} : \sup_{x^k} |T(\boldsymbol{x}) - T(\boldsymbol{x}^k)| = \infty\} = 50\%.$$

□

3.4 Minimizing the Loss Function

Given data points $\vec{z} = \{z_i\}_{i=1}^n \subset \mathbb{R}^d$, a model $\mathcal{M}$, a penalty function $\gamma : 2^{\mathcal{M}} \to \mathbb{R}$, and an error function $\rho : \mathbb{R}^d \times \mathcal{M}$, the target function we try to minimize is

$$f(L) \triangleq \gamma(L) + \frac{\lambda}{n} \sum_{i=1}^{n} \min_{\ell \in L} \rho(z_i, \ell) \tag{3.4.1}$$

for $L \in 2^{\mathcal{M}}$, or say $L \subset \mathcal{M}$. When $|\mathcal{M}| < \infty$, solving $\arg\min_{L \subset \mathcal{M}} f(L)$ with brute force involves $O(2^{|\mathcal{M}|})$ calculations. Here we propose *Remove-or-Replace algorithm (RRA)* on approximating a solution of $\arg\min_{L \subset \mathcal{M}} f(L)$ when $|\mathcal{M}| < \infty$.

3.4.1 Remove-or-Replace Algorithm

The procedure of remove-or-replace algorithm, RRA, is as follows.

1. (Initialization) Choose an initial set $L \subset \mathcal{M}$, say $L = \{\ell_1, \cdots, \ell_{|L|}\}$.

2. For $i = 1, \cdots, |L|$, set $L^{(i)} = L \backslash \ell_i$, calculate ℓ_i' with

$$\ell_i' \quad = \quad \arg \min_{\ell \in \phi \cup \mathcal{M} \backslash L} f(\ell \cup L^{(i)}).$$

   Note that we also include ϕ, called the removal element, such that $\phi \cup L^{(i)} = L^{(i)}$. That is, if the minimizer is ϕ, we know that the target function is smaller without any choice of ℓ_i (among $\mathcal{M} \backslash L$). If there are multiple arguments of minimum, pick ϕ if ϕ is among them, otherwise, pick the first occurrence of $\ell \in \mathcal{M} \backslash L$.

Record the value of target function

$$c_i \quad = \quad f(\ell_i' \cup L^{(i)}),$$

and the corresponding set $L_i = \ell_i' \cup L^{(i)}$.

3. After previous step we have $\{c_i\}_{i=1}^{|L|}$ and $\{L_i\}_{i=1}^{|L|}$. Calculate

$$r \quad = \quad \arg\min_{i=1:|L|} c_i,$$

and replace L with L_r if $f(L) < f(L_r)$. Again, if there are multiple arguments of minimum, pick the first occurrence of i in $c_1, \cdots, c_{|L|}$.

4. Repeat step 2 and 3 until there is no further decreasing on target function. The algorithm found L as an approximate minimizer of f.

The complexity of the algorithm is $O(q^2|\mathcal{M}|)$, where q is the number of element on the initial set. It is clear that every step of RRA decreases the target function f, which make it a greed algorithm. The reason we don't want to specify the number of instance, m, instead of introducing parameter λ in our loss function (Function 3.4.1) is the ability to perform RRA. Note that RRA is applicable for any function $f(L)$ where $L \subset \mathcal{M}$ with $|\mathcal{M}| < \infty$.

In the rest of this subsection, we will go through the setting of experiments on RRA. The error function we choose is denoted as the modified sparse distance, which depends on the data and the model chosen:

$$\rho(\cdot, \cdot) = \frac{d(\cdot, \cdot)^\beta}{\zeta_{\vec{z},\mathcal{M}}},$$

where $\beta \in (0,1)$ and

$$\zeta_{\vec{z},\mathcal{M}} = \frac{1}{n \cdot |\mathcal{M}|} \sum_{j=1}^{|\mathcal{M}|} \sum_{i=1}^{n} d(z_i, \ell_j)^\beta$$

is the *modified parameter* $\zeta = \zeta_{\vec{z},\mathcal{M}}$. The main reason we include it is that we would like to have similar loss function value on the sum of error term between different models. For example, given the same set of data $\vec{z}$, the point model, will have bigger value of ζ than the line model. This small modification make parameter λ less sensitive to model changes.

Note that since we assume $\beta \in (0,1)$ throughout this article, Corollary 3.2.12 greatly reduces the problem on minimize Equation 3.4.1 under certain situations. When the line model $\mathbb{L}$ is considered, the corollary tells us that we only need to consider C_2^n interpolating lines, i.e. $C_2^n < \infty$, which enables us to use RRA algorithm. More generally, when using the plane model $\mathbb{P}^d$ on $\mathbb{R}^d$, we only need to consider interpolating C_d^n planes. We also found that even when the optimizer is not interpolating, if we restrict the model to those instances that interpolate, the result is still very informative.

Thus, given the data $\vec{z} = \{z_i\}_{i=1}^n \subset \mathbb{R}^d$, the models we choose will be restricted to *interpolating models*, $\mathcal{M}_p$, where $p \in \mathbb{N}$, $p \leq d$, means the model is the collection of instance that passing p linearly independent data points $z_1, z_2, \cdots z_p$. For example, $\mathcal{M}_1$ is the interpolating point model and $\mathcal{M}_2$ is the interpolating line model, etc. Also, we will choose a penalty function $\gamma(L) = |L|^2$.

Thus, for the most of the time the target function we use is

$$f(L) \triangleq |L|^2 + \frac{\lambda}{n \cdot \zeta_{\vec{z},\mathcal{M}}} \sum_{i=1}^{n} \min_{\ell \in L} d(z_i, \ell)^\beta \tag{3.4.2}$$

for $L \subset \mathcal{M}_p$, $p \leq d$. Note that generally speaking we have $\zeta_{\vec{z},\mathcal{M}_{p_2}} < \zeta_{\vec{z},\mathcal{M}_{p_1}}$ if $d \geq p_2 > p_1 \in \mathbb{N}$. The error function is modified so that the parameter λ is less susceptible to p.

To evaluate how good the approximation of RRA is, we will also calculate the exact solution of the minimizer of $f(L)$ for small amount of instances. Since the penalty function only depends on the number of instances, we can compare RRA and the exact solution by the sum of error, h. That is, for each m, the number of instances, we examine the instances acquired by RRA or by the exact solution with the *sum of error*

$$h(L) \triangleq \sum_{i=1}^{n} \min_{\ell \in L} \rho(z_i, \ell), \tag{3.4.3}$$

with $|L| = m$ and $L \subset \mathcal{M}_p$. The h value will be shown on figures that we provided in the next few subsections. An easy observation is that when $m = 1$, the exact solution will be the same as RRA gives, so we will omit displaying it. Note that when m is known, to look for exact solution will cause C_m^N, where $N = C_p^n$, computations, and we usually cannot find the exact solution for larger p and m.

About the initial set, since we can easily choose n instances of model that make data term vanished, there is no point to have more then n elements in the initial set L. (Actually, for the line model case, we don't need more than $[\frac{n}{2}]$ elements, because any possible line must connect two data points.) When point model $\mathcal{M}_1$ is considered, the initial set is the same as $\mathcal{M}_1$, i.e. the data points. For other models, there are

two strategies to choose the initial set:

1. (*Locally best initial*) Given data $\{z_i\}_{i=1}^n$, for each data point z_i, $i = 1, \cdots, n$, we can choose the best instance, say ℓ_i, of the given model that fit locally such that ℓ_i that minimizes $\sum_{z_j \in N_k(z_i)} \rho(z_j, \ell_i)$, where $N_k(z_i) \triangleq \{z_j \in z | z_j \text{ is } z_i's \text{ } k \text{ nearest neighbor.}\}$. So we have n instances. Then, remove the redundant instances, and we get the initial set $L = \{\ell_i\}_{i=1}^q$ to work on the aforementioned algorithm. The nearest neighbor parameter k is $[\frac{n}{5}]$ through the experiments if this initialization method is chosen.

2. (*Random initial*) Randomly choose q instances, where $q \leq n$, whichever make data term vanished.

3.4.2 Robust Nature for Sparse Distance

We mentioned that we deliberately choose $\beta \in (0, 1)$ due to the robustness issue. For multiple line instances that appeared in the data, our following experiment shows the robust nature is also retained on minimizing the function

$$f(L) \triangleq |L|^2 + \frac{\lambda}{n \cdot \zeta_{z,\mathcal{M}_2}} \sum_{i=1}^{n} \min_{\ell \in L} d(z_i, \ell)^\beta, \tag{3.4.4}$$

for $L \subset \mathcal{M}_2$ in $\mathbb{R}^2$. Figure 3.4.1 plots both the result of EM algorithm (with number of instances $m = 2$ is known) and RRA result (with the line model and random initial) to the corrupted X-shaped data. The X-shaped data has originally 60 data point, and is corrupted by adding 40 outliers. The EM algorithm assumes two line instances with Gaussian error, and is initialized with the exact 2 lines that generate

the 60 un-corrupted data. We can see that the result of EM algorithm gives a poor result in this example, compare to that from RRA. In this case, the exact solution of $\min_{L\in\mathcal{M}_2} f(L)$ for $m = 2$ is the same as we got from RRA, which indicates that the loss function f is indeed a great choice for this corrupted data.

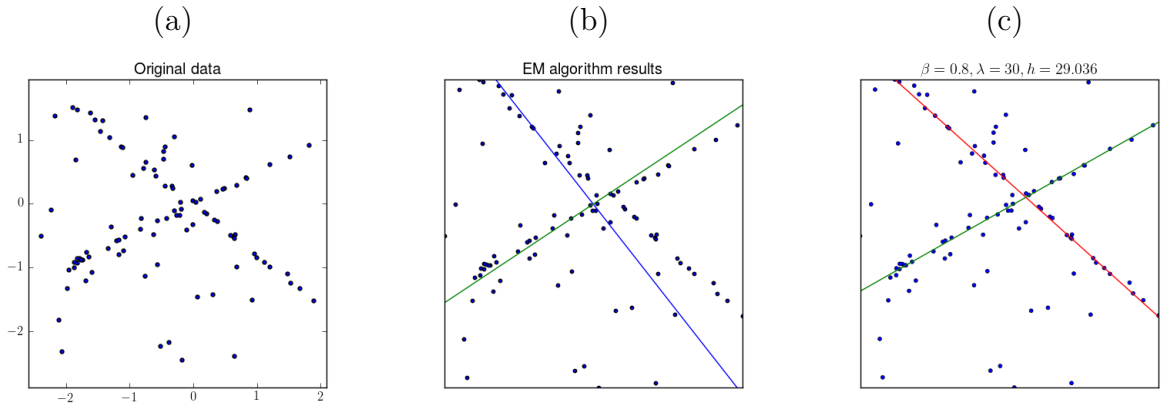

Figure 3.4.1: We examine the minimizer of Function 3.4.4 for $L \subset \mathcal{M}_2$, a set of interpolating lines in $\mathbb{R}^2$ on X-shaped data with outliers. (a) The original data consists of 100 data points that corrupted by 40 outliers. (b) EM algorithm for 2 instances. (c) RRA result for $\beta = 0.8$ and $\lambda = 30$ with random initial. In this case, the exact solution coincides RRA result.

Another things to point out is about the choice of λ. For data with this many outliers, one can decrease the value of λ to reduce the number of lines that RRA will find, so we set $\lambda = 30$ in this case. Actually, when one choose λ between 10 to 50, RRA will give two lines, see Figure 3.4.2. Note even when λ is miss-specified, the instances that RRA gives is still very informative.

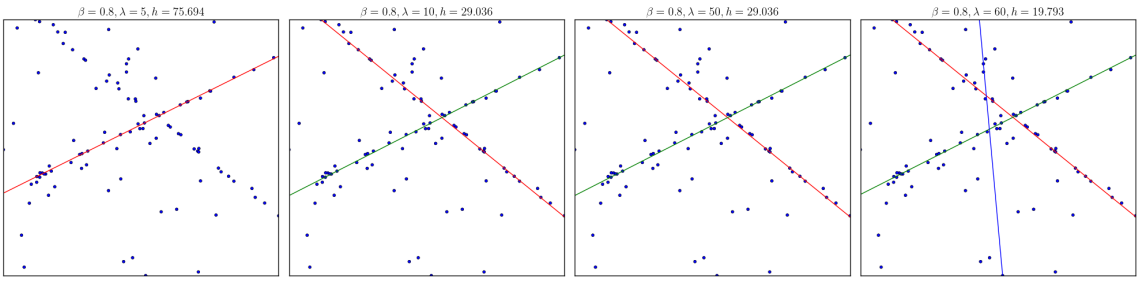

Figure 3.4.2: Different λ setting to RRA when minimizing Function 3.4.4 for $L \subset \mathcal{M}_2$, a set of interpolating lines in $\mathbb{R}^2$, on X-shaped data with outliers. One can see that for $\lambda = 10 \sim 50$, the algorithm gives two lines. Even when other λ is chosen, the instances that RRA found is still informative. Note that here we use random initial for RRA.

3.4.3 Instances from a Single Model

The intention of this subsection is to see how RRA work on various of data from a single model. For reader's reference, the target function we used here is

$$f(L) \triangleq |L|^2 + \frac{\lambda}{n \cdot \zeta_{z,\mathcal{M}_p}} \sum_{i=1}^{n} \min_{\ell \in L} d(z_i, \ell)^\beta, \tag{3.4.5}$$

where $L \subset \mathcal{M}_p$. ($\mathcal{M}_p$ is the model that collects all the interpolating $\mathbb{R}^{p-1}$ linear structure, i.e. the linear structure that determined by p linear independence data points.) We already see the loss function produces a robust result in Subsection 3.4.2. In particular, in this subsection we would like to know:

1. How general can we choose the parameter λ? Also, does initialization of RRA matters a lot?
2. How is the result from RRA compare to the exact solution?
3. How much do we sacrifice if we only look for interpolating solution? For example, for data in $\mathbb{R}^2$, if we look for instances from the interpolating point model $\mathcal{M}_1$, i.e. the data point. Another example is for data in $\mathbb{R}^3$, while we look for instances from $\mathcal{M}_2$, the collection of the lines that connect two data points.

3.4.3.1 Line Model $\mathcal{M}_2$ on $\mathbb{R}^2$ Data

For line model on $\mathbb{R}^2$ data, to minimize Function 3.4.5 on all the lines in $\mathbb{R}^2$ is equivalent to minimize on $\mathcal{M}_2$, since the minimizer is interpolating. The data that

we will examine line model on $\mathbb{R}^2$ is shown in Figure 3.4.3, which is called the "S-curve" data generated from from

$$(x_i, y_i) \quad = \quad (u_i, -0.02u_i^2 - 0.005u_i + u_i + 10 + \epsilon_i),$$

where $u_i \sim U(-10, 10)$ and $\epsilon_i \sim N(0, 1)$. The exact solution for the loss function we proposed (i.e. Function 3.4.5) is shown in Figure 3.4.4 for $m = 2$ and 3. The sum of error, h, as we mentioned in Equation 3.4.3 is shown on the top of the plots which will help examine the result from RRA.

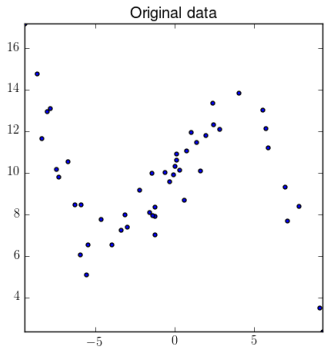

Figure 3.4.3: The "S-curve" data consists of 50 data points ($n = 50$).

The result of RRA with locally best initial (described in Subsection 3.4.1) is shown in Figure 3.4.5. We can see that the algorithm obtains the exact solution in this case, since the sum of error is the same as in the exact solution. Observe that the parameter λ controls the weight on fitting the data and the number of lines it could find. Most of the time the algorithm found a reasonable solutions.

We also tried RRA with random initial (described in Subsection 3.4.1) on the S-shaped data. The result is plot in Figure 3.4.6 where the different random seed is used. One of six initials didn't find 3 desired line instances, which gives an idea that

(a) Exact Solution for $m = 2$ (b) Exact Solution for $m = 3$

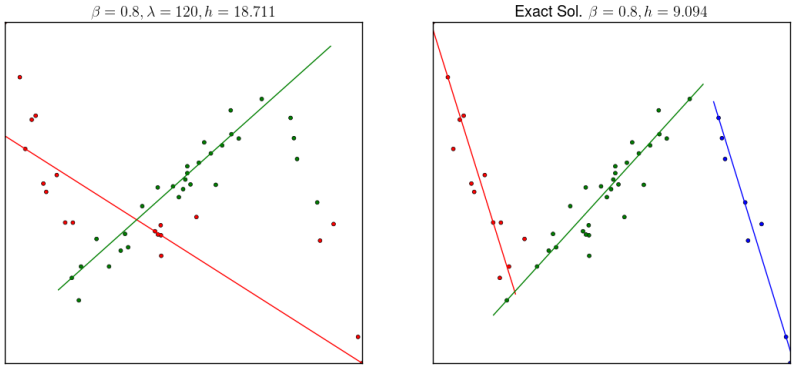

Figure 3.4.4: The exact solution (plotted as set of lines) of minimizing the loss function (Equation 3.4.5) while we use $\beta = 0.8$. (a) When the amount of instances $m = 2$ and (b) $m = 3$.

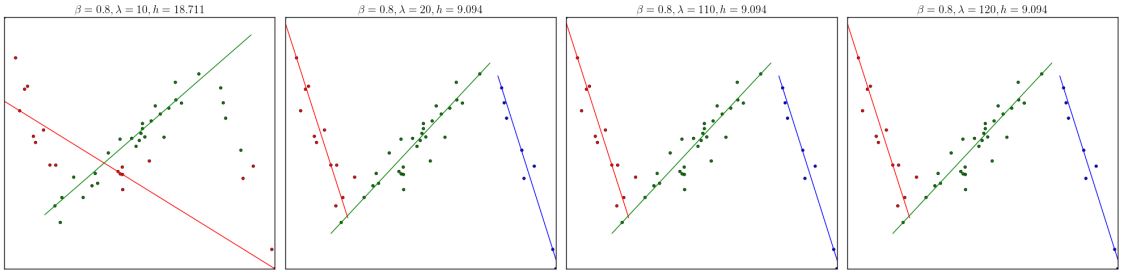

Figure 3.4.5: The result of RRA with locally best initial on S-curve data ($n = 50$) for different λ, where $\beta = 0.8$. When $\lambda = 20 \sim 110$ (the second and the third plots), RRA gives 3 line instances. Note that the sum of error, h, for $m = 2$ is 18.711 (the first plots) and for $m = 3$ is 9.094 (the second and the third plot) coincide with that of exact solution, means that RRA approximates the exact solution well in this data.

RRA is not very sensitive to initials. Generally speaking, random initial still gives a very acceptable result in this example

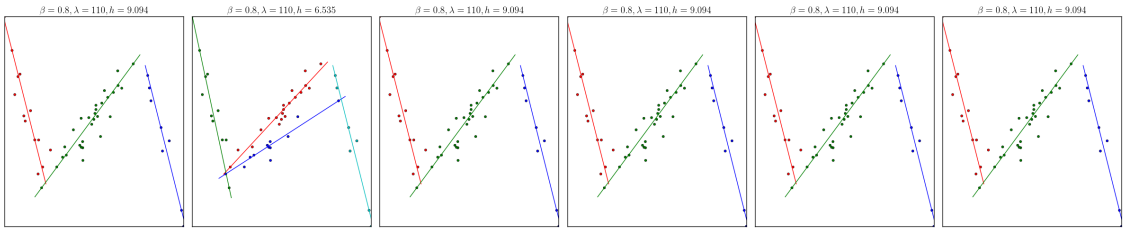

Figure 3.4.6: The result of RRA ($\beta = 0.8$, $\lambda = 110$) with random initial on S-shaped data with different random seed. From left to right, the random seed are set as 0,1,2,3,4,5, respectively. One can see that initialization does not cause huge different on RRA result in this sample.

Some other example is shown in Figure 3.4.7.

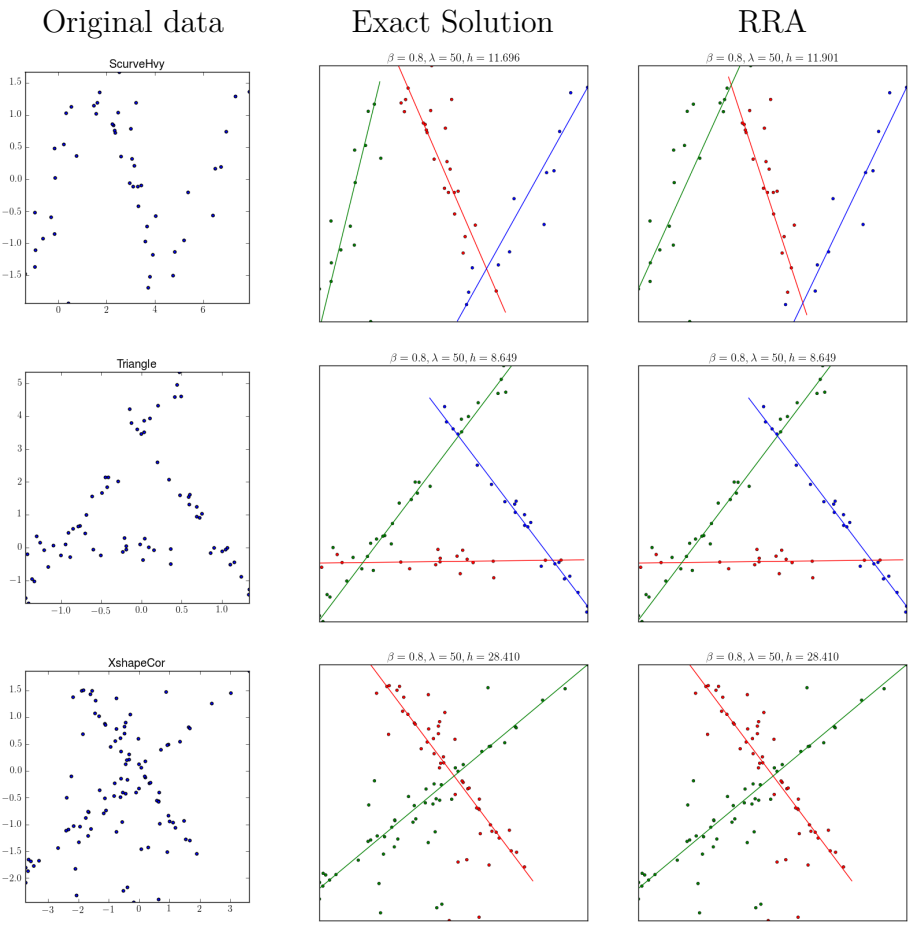

Figure 3.4.7: Some results of various data in $\mathbb{R}^2$. RRA is initialized with locally best initial.

3.4.3.2 Line Model $\mathcal{M}_2$ on $\mathbb{R}^3$ Data

Here we examine RRA on the interpolating line model on $\mathbb{R}^3$. Note that the optimizer will, in general, not interpolate. Given $r = (r_1, r_2, r_3) \in \mathbb{R}^3_1$ with $r_1 > 0$, $\sigma > 0$ and a point $z_0 = (x_1, x_2, x_3) \in \mathbb{R}^3$, a set of data $\{z_i\}_{i=1}^n = \{(x_{i1}, x_{i2}, x_{i3})\}_{i=1}^n \subset \mathbb{R}^3$ forms roughly a line in 3D where $x_m \leq x_{i1} \leq x_M$ can be generated by

$$\begin{aligned}
\theta &= (r, \sigma, z_0, x_m, x_M) \\
t_i &\sim_{iid} U(\frac{x_m - x_0}{r_1}, \frac{x_M - x_0}{r_1}) \\
\epsilon_i &\sim_{iid} N(0, \sigma^2) \\
x_{1i} &\sim_{iid} x_1 + t_i \cdot r_1 \\
x_{2_i} &\sim_{iid} x_2 + t_i \cdot r_2 + \epsilon_i \\
x_{3i} &\sim_{iid} x_3 + t_i \cdot r_3 + \epsilon_i,
\end{aligned}$$

for $i = 1, 2, \cdots, n$. The test data is generate by three different sets of θ:

$$\begin{aligned}
\theta_1 &= (0, 0, 3, 0.3, 5, -4, 1, -3, 5) \\
\theta_2 &= (1, 2, 5, 0.3, -2, 3, 4, -1, 4) \\
\theta_3 &= (1, 1, 1, 0.3, 1, 10, 3, 0, 4).
\end{aligned}$$

The original data is called "Line 3D" data, plotted in Figure 3.4.8. Figure 3.4.10 compares RRA result with exact solution. RRA is performed with locally best initial and $\beta = 0.8$, $\lambda = 50$. One can see that RRA gives a good approximation on the exact solution. We also test it with random initial with various random seeds, the result is very consistent and similar to the result from locally best initial.

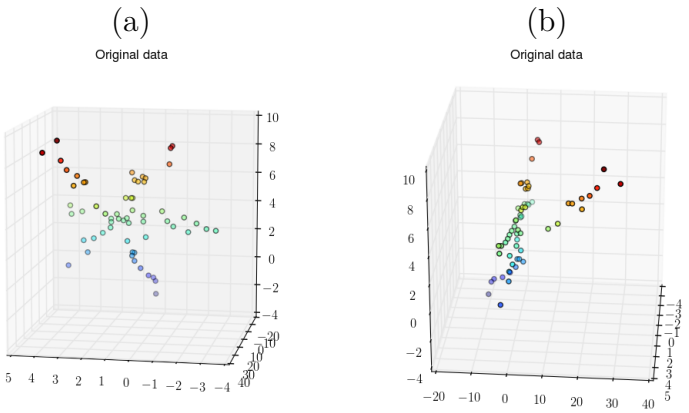

Figure 3.4.8: "Line 3D" data for the experiment on the interpolating line model on $\mathbb{R}^3$. (a) and (b) are different viewing angle.

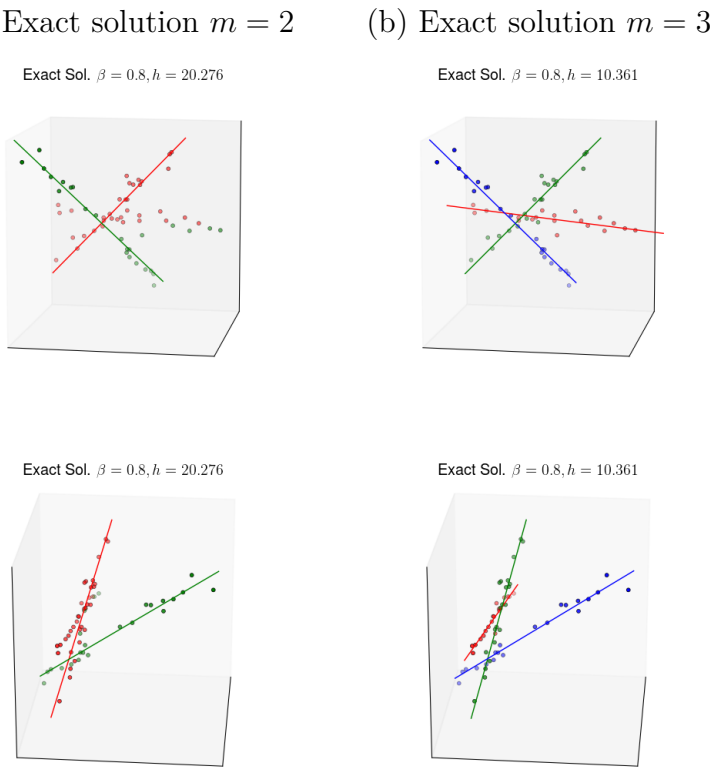

Figure 3.4.9: The exact solution for "Line 3D" data. (a) The exact solution for $m = 2$ with different viewing angles on top and bottom. (b) The exact solution for $m = 3$.

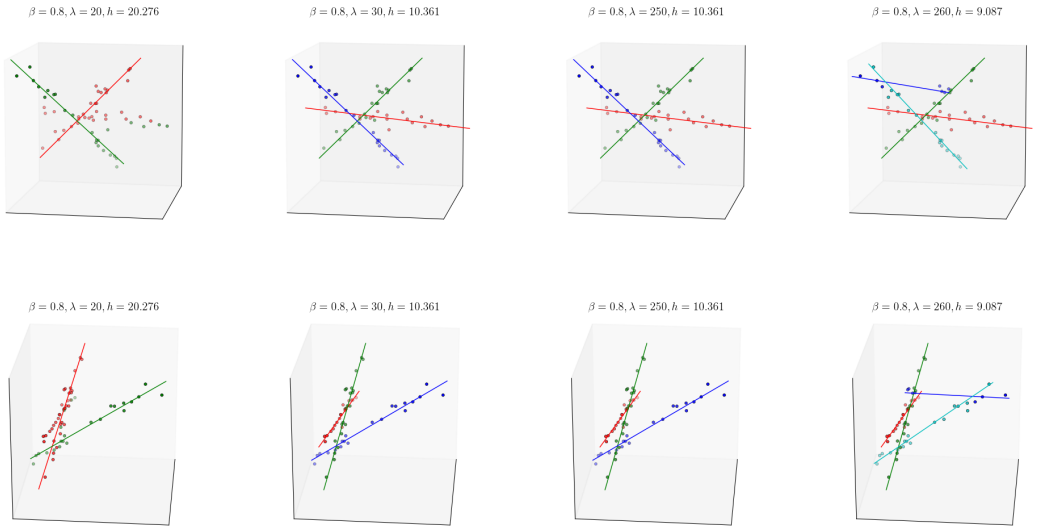

Figure 3.4.10: Result of "Line 3D" on $\mathbb{R}^3$ data. Top row and bottom row are different viewing angle of the result. From left to right is the result of different λ setting. A desired results are occurred when $\lambda \in (30, 250)$. Note that RRA coincides with the exact solution.

3.4.3.3 Plane Model $\mathcal{M}_3$ on $\mathbb{R}^3$ Data

In this subsection we examine RRA when line model is considered in $\mathbb{R}^3$. The data is generated from:

$$\begin{aligned} x &= (x_1, x_2, x_3) \in \mathbb{R}^3 \\ x &\sim \frac{1}{K}\sum_{k=1}^{K} g(x|\theta_k), \end{aligned}$$

where $g_k(\cdot)$ is a density function defined as:

$$\begin{aligned} g(x|\theta_k) &\propto \mathbb{I}_{[m_{1k},M_{1k}]}(x_1) \cdot \mathbb{I}_{[m_{2k},M_{2k}]}(x_1) \cdot \exp(-\frac{1}{2\cdot\sigma_k^2}(x_3 - \frac{(a_k x_1 + b_k x_2 + d_k)}{c_k})^2) \\ \theta_k &= (m_{1k}, M_{1k}, m_{2k}, M_{2k}, a_k, b_k, c_k, d_k, \sigma_k) \end{aligned}$$

The planes data has 40 data points generated from 2 set of parameters ($K = 2$):

$$\begin{aligned} \theta_1 &= (0, 5, -5, 5, 0, 0, 1, 0, 0.2) \\ \theta_2 &= (0, 5, -5, 5, 0, 1, 1, 0, 0.2), \end{aligned}$$

which is call "Plane 3D" data set. The original data is plotted in Figure 3.4.11 and the exact solution is shown in Figure 3.4.12 which uses $\beta = 0.8$, $\lambda = 50$. To represent the result, we plot it with two different perspective. RRA result is shown in Figure 3.4.13 (initial with the locally best initial). We can see that the result is identical to the exact solution.

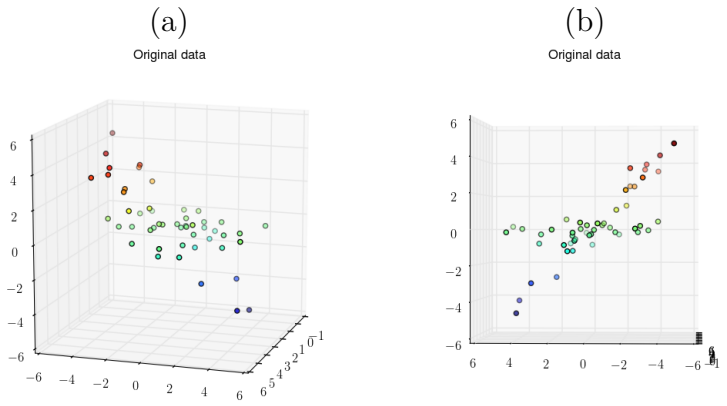

Figure 3.4.11: The"Plane 3D" data set for the experiment on the plane model on $\mathbb{R}^3$. (a) and (b) are different viewing angle.

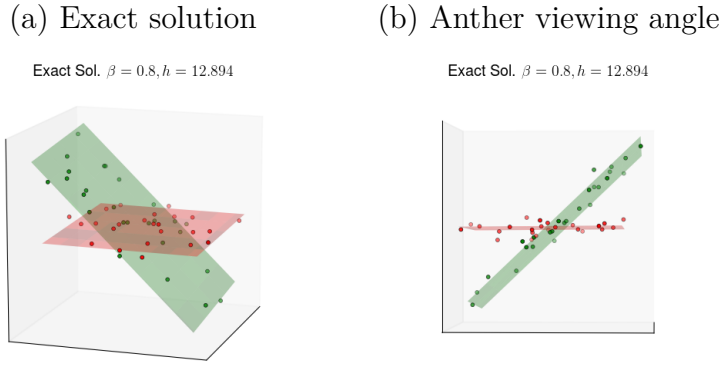

Figure 3.4.12: The exact solution for $m = 2$ for "Plane 3D" data. (a) and (b) are different viewing angle.

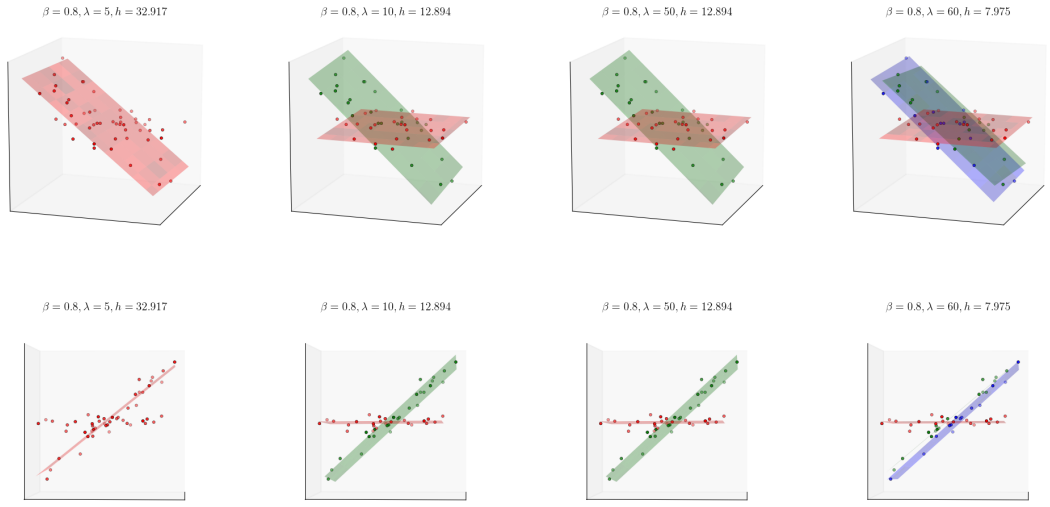

Figure 3.4.13: RRA results of "Plane 3D" on $\mathbb{R}^3$ data. Top row and bottom row are different viewing angle of the results. From left to right are the results of different λ setting. A desired results are occurred when $\lambda \in (10, 50)$. Note RRA coincides with the exact solution in this case.

3.4.4 Instances from Multiple Models

So far we have examined RRA for multiple instances from a single model. We would like to consider the case when multiple instances of different models appears in data in this subsection. In general, we want to find structures from a set of data $\vec{z} = \{z_i\}_{i=1}^n \subset \mathbb{R}^d$ that have instances from multiple (linear) interpolating models $\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_d$. Note that, for $p \leq d$, $p \in \mathbb{N}$, $\mathcal{M}_p$ is the model that contains linear structures in $\mathbb{R}^d$ that passing through p linear independent data points. One can still use RRA by setting $\mathcal{M} = \cup_{p=1}^d \mathcal{M}_p$, and look for an interpolating minimizer. However, if there is no penalty on using more sophisticated model, the sum of the error function that we preferred, $\sum_{i=1}^n \min_{\ell \in L} d(z_i, \ell)^\beta$, will favor instances from $\mathcal{M}_d$, the highest level of models. Due to this problem, we provide two strategies

1. Penalizing higher level models, and
2. Inspect with the sum of minimum volume,

as the following two subsections.

3.4.4.1 Penalizing Higher Level Models

The first strategy is to penalize the usage of the more complicated model by adding heavier weights on the error function. That is, one considers the following target function

$$f(L) \triangleq |L|^2 + \frac{\lambda}{n} \sum_{i=1}^n \min_{\ell \in L} \frac{\omega_{\psi(\ell)} \cdot d(z_i, \ell)^\beta}{\zeta_{\vec{z}, \mathcal{M}_{\psi(\ell)}}},$$

for $L \subset \mathcal{M} = \cup_{p=1}^{d} \mathcal{M}_p$, and $\psi : \mathcal{M} \to \{1, 2, \cdots, d\}$ s.t. $\psi(\ell) = p$ for $\ell \in \mathcal{M}_p$, $p \in \{1, 2, \cdots, d\}$, and $\omega : \{1, 2, \cdots d\} \to \mathbb{R}$ is the *model penalty parameter* s.t. $\omega_{p_1} \leq \omega_{p_2}$ for $p_1 < p_2 \in \{1, 2, \cdots, d\}$.

Note that the modified parameter $\zeta_{\vec{z},\mathcal{M}_p}$ also penalize the more advanced model since $\zeta_{\vec{z},\mathcal{M}_{p_1}} > \zeta_{\vec{z},\mathcal{M}_{p_2}}$ for $p_1 < p_2$. However, in practice, one often need to add model penalty term ω_p to acquire a satisfactory result. As our experiments shown, see Figure 3.4.15 and Figure 3.4.16, the results are sensitive to the choice of the model penalty term ω_p. A caveat that worth noting is that we use the error function as

$$\rho(z, \ell) = \frac{\omega(\psi(\ell)) \cdot d(z_i, \ell)^{\beta}}{\zeta_{\vec{z},\mathcal{M}_{\psi(\ell)}}},$$

which helps on finding instances, but does not possess the trait of Euclidean distance, i.e. given a data point z and two linear instances $\ell_1 \in \mathcal{M}_p$, $\ell_2 \in \mathcal{M}_q$, $p \neq q \in \{1, 2, \cdots, d\}$, $\rho(z, \ell_1) < \rho(z, \ell_2)$, does not necessary imply $d(z, \ell_1) < d(z, \ell_2)$. This might cause some problems when we try to cluster with the error function $\rho(\cdot, \cdot)$. Due to this fact, when we do clustering, we use the usual Euclidean distance $d(\cdot, \cdot)$.

We generate two sets of 2D data, "percent" and "Corrupted Percent" to examine the effect of the model penalty parameter ω, which is shown in Figure 3.4.14. One of the data set is called clean percent shape data (see Figure 3.4.14(a)) that consists of 3 instances, two point instances and a line. For another data set (see Figure 3.4.14(b)) we add 40 uniformly distributed noises onto the clean percent shape data, which help us gain the idea of how robust the method is.

The result for the clean percent shape data is displayed in Figure 3.4.15. One can

see that when $\omega_1 = 1$, and $\omega_2 = 2 \sim 7$, RRA successfully find the 3 desired structure. And when $[\omega_1, \omega_2] = [1, 1.5]$, RRA favors line structure, a higher level model, as we mentioned. On the contrary, when $[\omega_1, \omega_2] = [1, 8]$, RRA favors point instances, which indicates over compensation happened.

The result for the corrupted percenti shape data is in Figure 3.4.16. When $\omega_1 = 1$, $\omega_2 = 1.5 \sim 6$, we see that RRA find the structures, even under the presence of 30.8% noises.

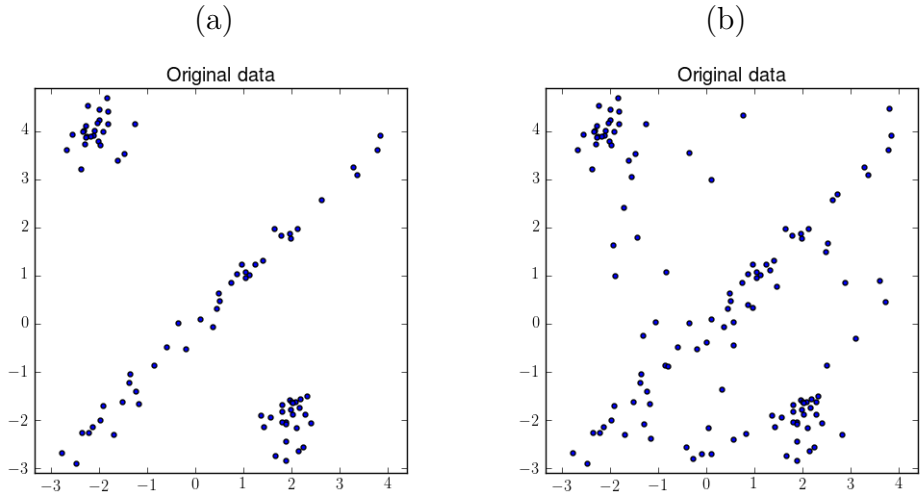

Figure 3.4.14: The 2D data "percent" and "Corrupted Percent" for experiment in this section. (a) "Percent" data consists of $n = 90$ data points. (b) "Corrupted Percent" data has 40 uniformly distributed noises as outliers in "percent" data (30.8% of the data).

3.4.4.2 Inspecting with Minimum Sum of Volumes

Another strategy is less sensitive to the choice of parameters. We called it the *minimum sum of volumes* (*MSV*) method. The idea is to utilize the result instances from RRA with different models as candidates, and pick the candidates by minimize the

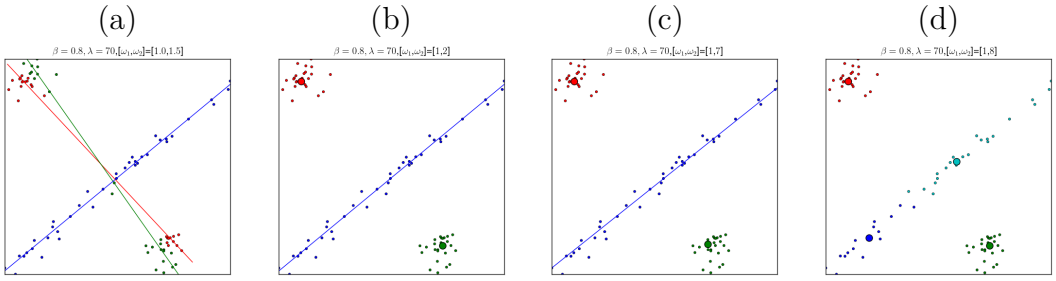

Figure 3.4.15: Results of RRA with the model penalty parameter $[\omega_1, \omega_2]$ on the "Percent" data. ($n = 90$). While setting $\omega_1 = 1$, a satisfactory result is appeared when $\omega_2 = 2 \sim 7$ (Plots (b) and (c)). The parameters are denoted on top of each plots. One can see the penalty term acts on RRA result. The big dot markers are the instances in $\mathcal{M}_1$ found by RRA; the line is the instance in $\mathcal{M}_2$ found by RRA.

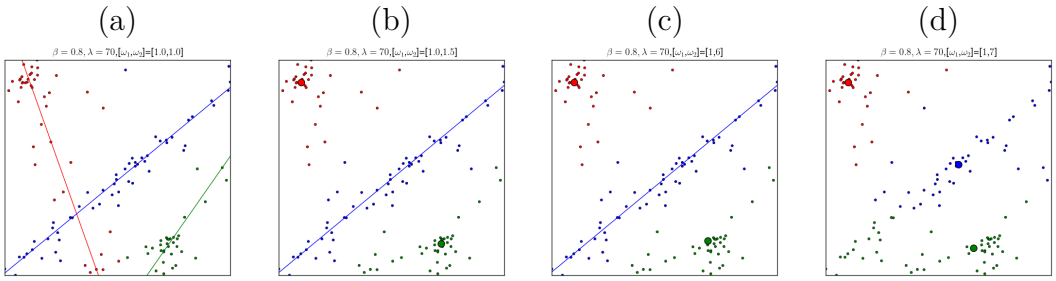

Figure 3.4.16: Results of RRA with the model penalty parameter $[\omega_1, \omega_2]$ on "Corrupted Percent" data ($n = 130$). While setting $\omega_1 = 1$, a satisfactory result is appeared when $\omega_2 = 1.5 \sim 6$ (Plots (b) and (c)). The parameters are denoted on top of each plots. One can see the penalty term acts on RRA result. The big dot markers are the instances in $\mathcal{M}_1$ found by RRA; the line is the instance in $\mathcal{M}_2$ found by RRA.

volume of the structures. That is, the candidates are picked from RRA with $\mathcal{M}_p$ for $p = 1, 2, \cdots, P$, by minimizing

$$f(L) \triangleq |L|^2 + \frac{\lambda}{n} \sum_{i=1}^{n} \min_{\ell \in L} \rho(z_i, \ell)$$

for $L \subset \mathcal{M}_p$, where

$$\rho(z_i, \ell) \quad = \quad \frac{d(z_i, \ell)^\beta}{\zeta_{\vec{z}, \mathcal{M}_p}}.$$

Given an instance $\ell^p \in \mathcal{M}_p$, a cluster of data $A \subset \vec{z}$, and $u \in (0, 100]$, the *volume* is defined as

$$v_u(\ell^p, A) \triangleq C(r^u_{\text{inclass}}, p-1) \cdot C(r_{\text{dist}}, d-(p-1)),$$

where

$$C(r, k) \triangleq \frac{\pi^{k/2}}{\Gamma(\frac{k}{2}+1)} r^k$$

is the volume of a k-dimensional ball, and

$$r^u_{\text{dist}} = r^u_{\text{dist}}(\ell^p, A) \triangleq Q_u(\{d(z, \ell^p)\}_{z \in A}),$$

where $Q_u(C)$ is the uth-percentile of the 1-dimensional data C, $u \in (0, 100]$.

$$r_{\text{inclass}} = r_{\text{inclass}}(\ell^p, A) \triangleq \max_{x \in Proj(A, \ell^p)} d(x, \bar{x}),$$

where $Proj(A, \ell^p) = \{Proj(z, \ell^p)\}_{z \in A}$ is the set of points that is projected to ℓ^p from data in A. Note that $r_{\text{inclass}}(\ell, A) = 0$ for $\ell \in \mathcal{M}_1$, a instance from the point model, and $C(0, 0) = 1$. For example, when the dimension of data $d = 3$ is considered, the volume of a point instances along with the data it clustered is a ball in $\mathbb{R}^3$ with radius

r_{dist}; the volume of a line instances along with the data it clustered is a volume of a cylinder in $\mathbb{R}^3$, which has a height r_{inclass} and a circle bottom with radius r_{dist}; the volume of a plane instances along with the data it clustered is another volume of a (flatter) cylinder in $\mathbb{R}^3$, which has a height r_{dist} and a circle bottom with radius r_{inclass}. A example of volume in $\mathbb{R}^3$ is shown in Figure 3.4.21.

Given $u \in (0, 100]$, the MSV method is a procedure for finding m instances that have minimum sum of volumes, $MSV_u(m)$, is:

1. For each $p \in \{1, 2, \cdots, P\}$, obtain candidate of instances L_p from model $\mathcal{M}_p$ by minimizing
$$f(L) \triangleq |L|^2 + \frac{\lambda}{n \cdot \zeta_{\vec{z},\mathcal{M}_p}} \sum_{i=1}^{n} \min_{\ell \in L} d(z_i, \ell)^\beta$$
for $L \subset \mathcal{M}_p$. (This is done by RRA.) In the end we have $L^{cand} = \cup_{p=1}^{P} L_p$ is the collection of candidate instances from different models. Note that one should expect $k > m$, otherwise, raise the value of λ so one can have more candidates found.

2. Set $k = |L^{cand}|$. Thus we have $N = C_m^k$ combinations of m instances out of k candidates. The set of every combination is represented as $\{M_1, M_2, \cdots, M_N\}$.

3. For any combination , say $M = \{\ell_1, \ell_2, \cdots, \ell_m\}$, obtain the cluster $A_j = \{z \in \vec{z} | d(z, \ell_j) < d(z, \ell),\ \forall \ell \in L\}$ for $j = 1, 2, \cdots, m$. If two or more instances minimize the same data, randomly choose one. Calculate the sum of volumes, $SV_u(M) = \sum_{j=1}^{m} v_u(\ell_j, A_j)$, for this combination of candidates. If there is a cluster A_j, $j = 1, 2, \cdots, m$, which has less than ϵ amount of data, i.e. $|A_j| < \epsilon$, set $SV_u(M) = \infty$.

4. Choose the combination, $\hat{M}$, that has minimum sum of volumes. i.e. $\hat{M} = \arg\min_{i=1:N} SV_u(M_i)$

Note that we use $d(\cdot,\cdot)$, the Euclidean distance, to cluster and calculate volumes, instead of using the sparse distance $d(\cdot,\cdot)^\beta$. This is because Euclidean distance is more natural to use when clustering (compare to ρ) and to calculate volumes, where both of them relies on geometrical properties of the data. For a noisy data, we use the parameter u to keep the method robust. Also, in step 3 when calculating $MSV_u(m)$, we exclude the combination that contains a cluster that less than ϵ amount of data. This is to prevent some complications that might occur, especially when some outliers are present in the data that makes the computation of volumes difficult. We set $\epsilon = [\frac{n}{15}]$.

We apply MSV method on a 2D data .The original data and candidates are shown in Figure 3.4.17, and result of MSV is shown is Figure 3.4.18. In this data set, 30.8% of the data are outliers.

The application of MSV method on a 3D data is shown with data "Corrupted 3D mixture" in Figure 3.4.19 and Figure 3.4.20, where we have 15 outliers out of 65 data points. The uncorrupted version of the data consists of a point, a line, and a plane instance (total 3 instances). Also, in Figure 3.4.21 we show the volume derived for each instance in $m = 3$ case.

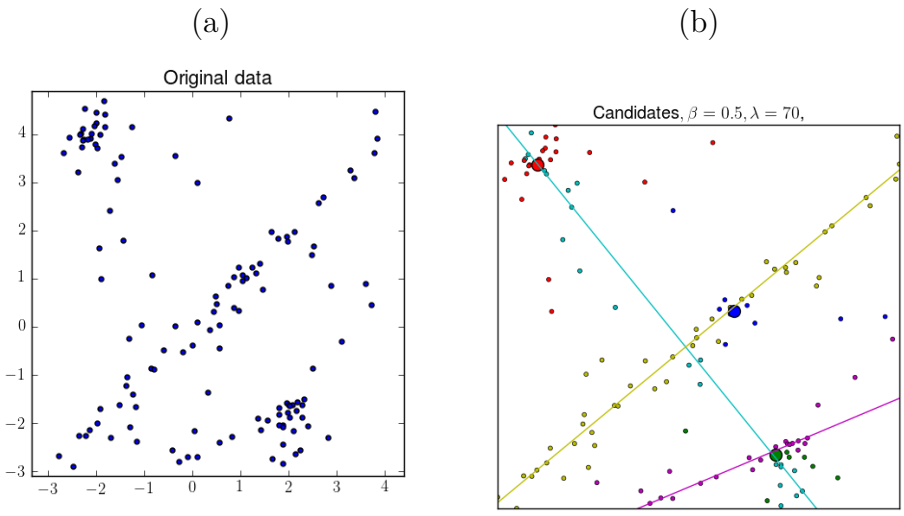

Figure 3.4.17: (a) "Corrupted percent" data, where 40 out of 130 data points are outliers (30.8% of the data). (b) Candidates for MSV method on percent data, required from RRA at step 1 of MSV method. There are total 6 candidates, represented as bigger dots and lines, appeared in this data

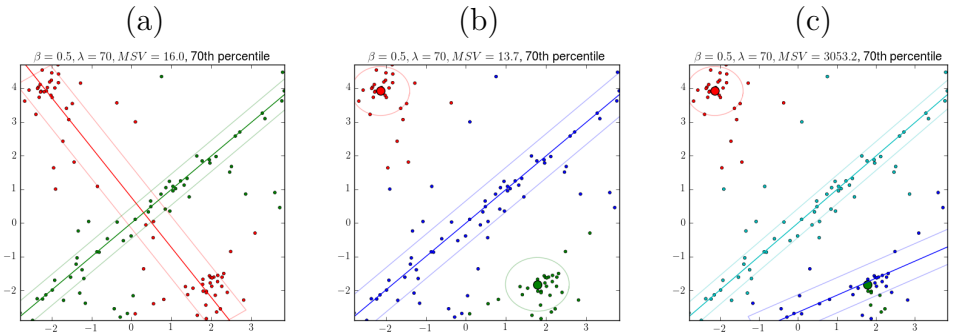

Figure 3.4.18: Results of MSV method on "Corrupted percent" data. Assuming $u = 70$, that is, roughly speaking, we use 70% of the data in a cluster to calculate the volume. (a) Assume the amount of instances $m = 2$. (b) $m = 3$. (c) $m = 4$. Note that the value of $MSV(m)$ is denoted at the top of each plot. When $m = 4$, there is no combination that provides the sufficient amount ($> \epsilon$) of data in a cluster, so $MMSV(4) = \infty$ (on the top of the plot, the MSV value is 3053.3, which is just a big value that indicate this happened). The transperant line is indicated as the "volume" of each instances.

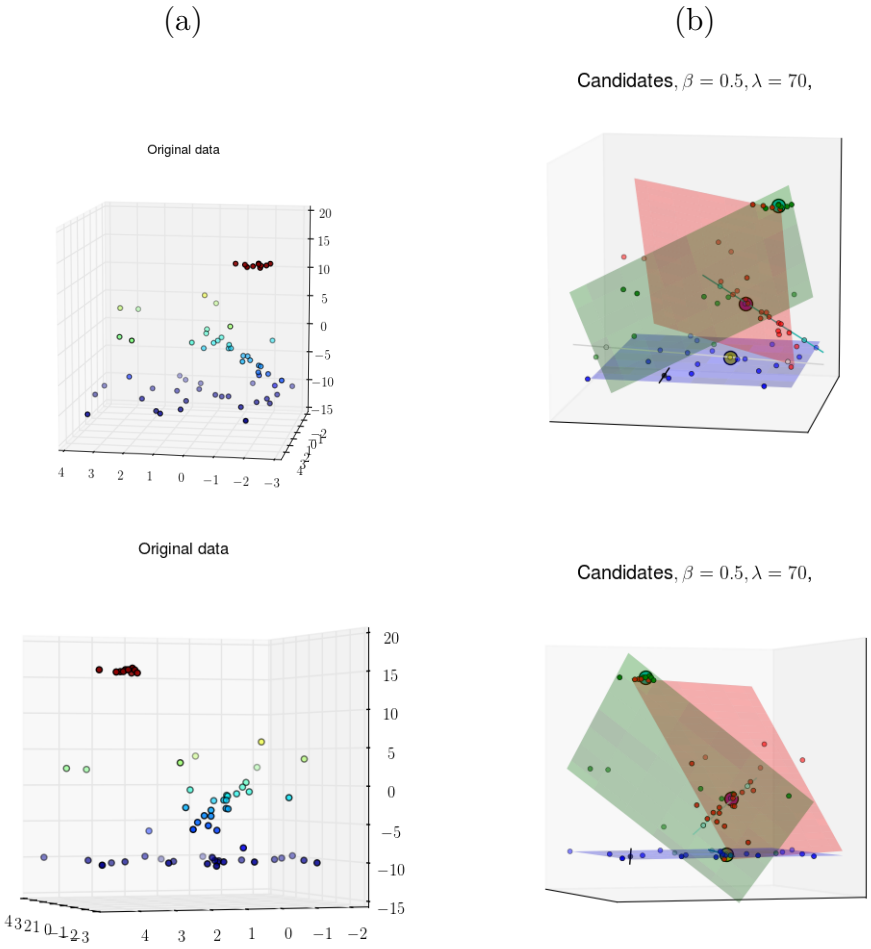

Figure 3.4.19: The "Corrupted 3D mixture" data, where 15 out of 65 data points are outliers (23.1% of the data). (a) The original data from different viewing angle (top and bottom). (b) Candidates for MSV method on percentile data, required from RRA, from different viewing angle (top and bottom). There are total 9 candidates, represented as bigger dots, lines, and planes, appeared in this data.

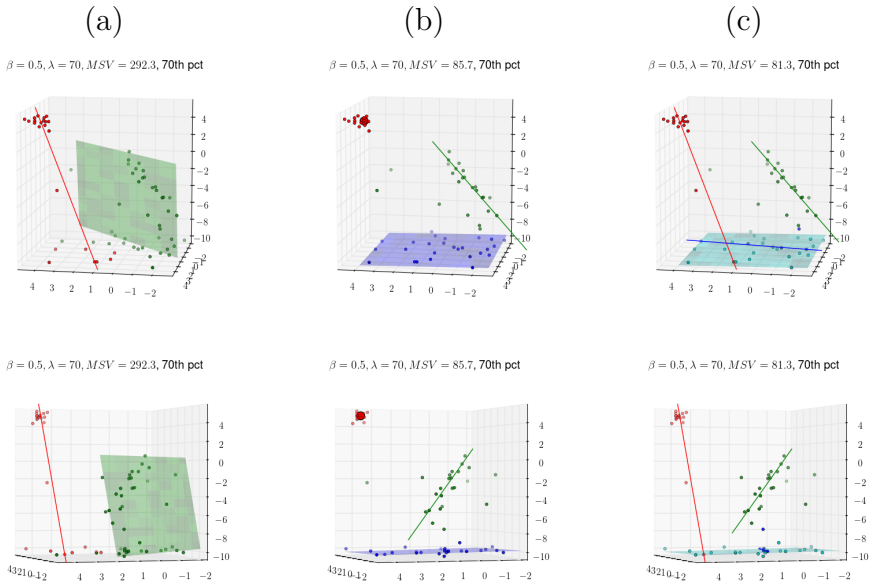

Figure 3.4.20: Results of MSV method on corrupted 3D mixture data. We provide different viewing angles by separating them at top row and bottom row. Assuming $u = 70$, i.e. 70% of data in a cluster is used to compute the volume, roughly speaking. (a) Assume the amount of instances $m = 2$. (b) $m = 3$. (c) $m = 4$. Note that the value of $MSV(m)$ is denoted at the top of each plot.

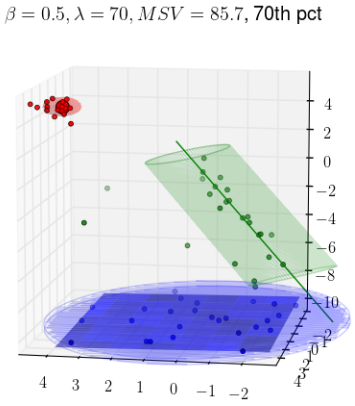

Figure 3.4.21: $m = 3$ result for "corrupted 3D mixture" data. The transparent area is the volume we calculated for each instance.

3.5 Future Work

There are many variations on the theme and as of yet few analytic guidelines. Future investigations will include:

1. **Performance Bounds.** We seek exact and tight bounds on the performance of the RRA algorithm. Our experiments, as well as the $\epsilon-$interpolation theorem of section 3.2.4 suggest that RRA, or perhaps another coarse-to-fine type search strategy, finds an excellent approximate solution. Can we bound the difference in loss between the RRA and the global optimum? The problem is reminiscent of the covering problem, where such bounds do exist.

2. **The Role of $|L|^2$.** We have suggested various ways of choosing the number of instances. The performance of any algorithm is likely to be sensitive to the complexity term (complexity "penalty") in the loss function. We have experimented exclusively with the quadratic loss. How will performance differ using other loss functions, e.g. other powers of $|L|$? More theory is needed.

3. **$\epsilon-$interpolation: can ϵ be set to 0?** Perhaps the most glaring hole in our analytic treatment concerns the theory on $\epsilon-$interpolation: In section 3.2.4 we constructed a function $\beta_o(n, k)$ such that $\beta < \beta_o$ guarantees that all instances of an optimal solution would be $\epsilon-$interpolating. Does there exist a similar result for $\epsilon = 0$? In other words, does there exist a function $\beta_o(n, k)$ (or, perhaps, $\beta_o(n, k, d)$) such that $\beta < \beta_o$ guarantees that all instances of an optimal solution are interpolating?

4. **Extending the library of shapes.** We can generalize to a far larger library

of surfaces, e.g. the quadratic surfaces in $\mathbb{R}^d$, perhaps defined by a system of implicit equations, $\mathcal{M} \leftrightarrow F(z, \theta) = 0,\ z \in \mathbb{R}^d$, one equation for each θ. If $\mathcal{M}$ is closed under Euclidean transformations, and perhaps to scale as well, then many of the same analytic and computational methods developed here for reducing search to the set of interpolating instances are still appropriate, albeit with greater theoretical and computational challenges. We plan to work towards the attractive idea of a having a rich library of manifolds that can be used to create a summary of a complex high-dimensional data set through a collection of multiple, heterogeneous instances.

Chapter Four

Conclusions

4.1 On Entropy Rate of Images

We examine how modern lossless image compression algorithms performed. Specifically, we aim at the dominated method in this field, the predictive context modeling method. To this end, the extended Shannon-McMillian-Breimen theorem is proved. First, under stationary condition, the coding rate will converge to the entropy rate of its own ergodic component for almost every image. That is, given an image, the entropy rate of the ergodic component with respect to the image is the theoretical minimum bits per pixel (bpp) of any compression algorithm. Second, with light assumptions, the predictive context model with a stationary probability will converge to the theoretical minimum when the size of context grows. In order to find out the minimum bpp for an image, we proposed CBIC, the comparison-based image compression algorithm. The compression scheme is a type of "the more you know the better you get" algorithm, which means we can put as many as resources (in our case, the library) we want to get a better result. This is done by utilizing the idea of empirical distribution from an external source. We see the bpp result converged when size of context grows, like the theory suggests. In the experiment, we get slightly better (about 0.05 to 0.15 bpp) result compare to one of the best compression scheme we have so far. This indicates that these schemes may already be close to the optimal.

4.2 On Robust Generalized Clustering

We propose a new loss function for clustering data points to shapes. Thus far we have focused on "linear shapes," meaning points, lines, and in general, high-dimensional

planes, but there are natural extensions to parametric classes of surfaces, such as quadratic and cubic surfaces. The idea is to fit multiple surfaces to the data (e.g. a heterogeneous collection of planes, including points, lines, and planes of different dimensions), assigning each point to the closest surface. Importantly, the loss function is a *concave* function of the distance from data points to their nearest surface. This induces "sparse" distances, meaning that optimal surfaces tend to be near or even on top of data points—sparse in the sense of many zero distances. Furthermore, concave loss is, by design, robust to outliers. In the case of planes, depending on the dimension of the data and the dimension of the planes, we can sometimes prove that every member of the minimizing collection is "interpolating," i.e. completely determined by a set of data points at zero distance. Related results indicate that the optimal planes will be close to or containing data points, provided that the concavity of the loss respects an explicit lower bound. Such results are useful for computation: whereas the minimization of the loss function is in general NP hard, a polynomial complexity "Remove or Replace" algorithm gives excellent approximations in cases where the actual minimum can be (laboriously) computed.

Bibliography

[1] Nabih N. Abdelmalek. On the discrete linear L_1 approximation and L_1 solutions of overdetermined linear equations. *J. Approximation Theory*, 11:38–53, 1974. ISSN 0021-9045.

[2] E. R. Barnes. An algorithm for separating patterns by ellipsoids. *IBM J. Res. Develop.*, 26(6):759–764, 1982. ISSN 0018-8646. doi: 10.1147/rd.266.0759. URL `http://dx.doi.org/10.1147/rd.266.0759`.

[3] P. Billingsley. *Ergodic Theory and Information*. Wiley, New York, 1965.

[4] Paul T. Boggs and Janet E. Rogers. Orthogonal distance regression, 1990. URL `http://dx.doi.org/10.1090/conm/112/1087109`.

[5] R. J. Boscovich. De literraria expeditione per pontificiam synopsis amplioris operis, ac habentur plura eius ex exempla ria etiam sensorum impressa. *Bononiensi Scientiarum et Artium Inst. Atque Acad. Commen.*, 4:353–396., 1757.

[6] L. Breiman. The individual ergodic theorem of information theory. *Ann. of Math. Stat*, 28:809–811, 1957.

[7] Leo Breiman. Hinging hyperplanes for regression, classification, and function

approximation. *IEEE Trans. Inform. Theory*, 39(3):999–1013, 1993. ISSN 0018-9448. doi: 10.1109/18.256506. URL http://dx.doi.org/10.1109/18.256506.

[8] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984. ISBN 0-534-98053-8; 0-534-98054-6.

[9] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006. ISBN 978-0-471-24195-9; 0-471-24195-4.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1): 1–38, 1977. ISSN 0035-9246. URL http://links.jstor.org/sici?sici=0035-9246(1977)39:1<1:MLFIDV>2.0.CO;2-Z&origin=MSN. With discussion.

[11] Terry E. Dielman. Least absolute value rregression recent contributions. *Journal of statistical computation and simulation*, 75(4):263–286, apr .

[12] David Donoho and Peter J. Huber. The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pages 157–184. Wadsworth, Belmont, CA, 1983.

[13] R. Durrett. *Probability: Theory and Examples*. Duxbury advanced series. Thomson Brooks/Cole, 2005. ISBN 9780534424411. URL https://books.google.com/books?id=NPYYAQAAIAAJ.

[14] Jerome H. Friedman. Multivariate adaptive regression splines. *Ann. Statist.*, 19 (1):1–141, 1991. ISSN 0090-5364. doi: 10.1214/aos/1176347963. URL http:

//dx.doi.org/10.1214/aos/1176347963. With discussion and a rejoinder by the author.

[15] F. Golchin and K. K. Paliwal. A lossless image coder with context classification, adaptive prediction and adaptive entropy coding. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle, Washington, USA*, pages 2545– 2548., May 1998.

[16] Robert M. Gray. *Probability, Random Processes, and Ergodic Properties*. Springer, 2010.

[17] Robert M. Gray. *Entropy and Information Theory*. Springer, 2013.

[18] Frank R. Hampel. A general qualitative definition of robustness. *Ann. Math. Statist.*, 42:1887–1896, 1971. ISSN 0003-4851. doi: 10.1214/aoms/1177693054. URL http://dx.doi.org/10.1214/aoms/1177693054.

[19] Sadri Hassani. *Mathematical physics*. Springer, Cham, second edition, 2013. ISBN 978-3-319-01194-3; 978-3-319-01195-0. doi: 10.1007/978-3-319-01195-0. URL http://dx.doi.org/10.1007/978-3-319-01195-0. A modern introduction to its foundations.

[20] Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[21] K. Jacobs. *The ergodic decomposition of the Kolmogorov-Sinai invariant*. Academic Press, New York, 1963.

[22] Pierre-Simon Laplace. *Théorie analytique des probabilités*. Mme Courcies, Paris, 1812.

[23] Kenneth D. Lawrence and Jeffrey L. Arthur. *Robust Regression: Analysis and Applications*. Marcel Dekker, 1990.

[24] E. Lindenstrauss. Pointwise theorems for amenable groups. *Invent. math*, 146, 2001.

[25] G. Seroussi M. J. Weinberger and G. Sapiro. LOCO-I: A low complexity, context-based, lossless image compression algorithm. *Data Compression Conference (Snowbird, UT)*, pages 140–149, 1996.

[26] J. J. Rissanen M. J. Weinberger and R. B. Arps. Applications of universal context modeling to lossless compression of gray-scale images. *IEEE Trans. on Image Processing*, 5(4), April 1996.

[27] J. MacQueen. Some methods for classification and analysis of multivariate observations. pages Vol. I: Statistics, pp. 281–297, 1967.

[28] Naresh Manwani and P. S. Sastry. K-plane regression. *CoRR*, abs/1211.1513, 2012. URL `http://arxiv.org/abs/1211.1513`.

[29] Bernd Meyer and Peter Tischer. TMW - a new method for lossless image compression. pages 533–538, 1997.

[30] N. M. Nasrabadi and R. A. King. Image coding using vector quantization: A review. *IEEE Trans. on communications*, 36(8), August 1988.

[31] A. Nevo and E.M. Stein. A generalization of birkhoff's pointwise ergodic theorem. *Acta. Math*, 173:135–154, 1994.

[32] D. Ornstein and B. Weiss. The shannon-mcmillan-breiman theorem for a class of amenable groups. *Isreal Journal of Mathematics*, 44(1), 1983.

[33] D. Ornstein and B. Weiss. Entropy and isomorphism theorem for actions of amenable group. *J. D'Analyse Mathematique*, 48, 1987.

[34] D. Ornstein and B. Weiss. Entropy and recurrence rate for stationary random fields. *IEEE Trans. Inf. Theory*, 48(6):1694–1697, June 2002.

[35] K. Ratakonda and N. Ahuja. Lossless image compression with multiscale segmentation. *IEEE Transactions Image Processing*, 11(11):1228–1237, 2002.

[36] J. B. Rosen. Pattern separation by convex programming. *J. Math. Anal. Appl.*, 10:123–134, 1965. ISSN 0022-247x. doi: 10.1016/0022-247X(65)90150-2. URL `http://dx.doi.org/10.1016/0022-247X(65)90150-2`.

[37] D. J. Rudolph. *Fundamentals of measurable dynamics: ergodic theory on Lebesgue spaces*. Oxford University Press, 1990.

[38] Ashish Sen and Muni Srivastava. *Regression Analysis:Theory, Methods, and Applications*. Springer, 1990.

[39] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Stat. Comput.*, 14(3):199–222, 2004. ISSN 0960-3174. doi: 10.1023/B:STCO.0000035301.49549.88. URL `http://dx.doi.org/10.1023/B:STCO.0000035301.49549.88`.

[40] Vladimir Vapnik, Steven E. Golowich, and Alex Smola. Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processing Systems 9*, pages 281–287. MIT Press, 1996.

[41] X. Wu. Context-based, adaptive, lossless image coding. *IEEE Trans. Communications*, 45(4), April 1997.