Instructions for authors: All submissions should include the name, institutional affiliation (if applicable), and postal and electronic mail addresses of each author. General inquiries may be addressed to the editorial board at `editor@mathreview.uwaterloo.ca`.

Articles: *The Waterloo Mathematics Review* invites both original research and quality expository articles written for an advanced undergraduate audience on topics related to mathematics at the undergraduate level including: computer science, mathematical physics, statistics, mathematical finance, and actuarial science.

Articles may be submitted electronically or in hard copy. Electronic submissions should be in Adobe PDF format, accompanied by the relevant source files (figures, TEX files, etc.), and sent as an attachment to `submissions@mathreview.uwaterloo.ca`. Hard copies may be sent to

The Waterloo Mathematics Review
Faculty of Mathematics
University of Waterloo
200 University Ave. West
Waterloo, Ontario, N2L 3G1

Submissions should include an abstract and bibliography (preferably in BibTEX format). Figures must be of a production quality, and submitted in a vector format (e.g. encapsulated PostScript). The *Review* can provide assistance in vectorizing hand-drawn figures.

Mathematical Art: The *Review* invites graphics of a mathematical nature. Submissions of graphics should include both black-and-white and colour versions in production quality, along with a description of the mathematics being depicted and a few words on how the image was produced. Source code submission is not required when images are produced using a custom program, however, should an artist request it the *Review* will publish the source code on our website, to accompany the electronic version of the issue.

Join The Review: Are you passionate about mathematics? The *Review* is currently looking for new members for the review board. If you are an interested undergraduate at the University of Waterloo please contact the editorial board at `editor@mathreview.uwaterloo.ca`. If reviewing papers does not interest you but you would like to be involved (or you are a student at another institution and would like to join the effort) the *Review* is also looking for a new General Manager, a webmaster, and local distribution partners at other institutions. Contact the editorial board at `editor@mathreview.uwaterloo.ca` for more information.

The Cover: The cover is the initial ocean surge being sent into towards the Maeslantkering, a storm surge gate on the Nieuwe Waterweg in the Netherlands. Unimpeded, the waterway risks being flooded by ocean surges and the Maeslantkering prevents this by reflecting ocean waves off its curved metal structure. Pictured is when the Maeslantkering is partially closed and an initial ocean surge is being sent down the waterway. This numerical simulation demonstrates how triangular grids with a finite volume scheme can model realistic physical systems with complicated geometries.

Archive: Past, present, and future issues of the *Review* are available in whole or in part on our website:

`mathreview.uwaterloo.ca`

Machine-readable citations and bibliography data from articles are also available.

Subscriptions: The *Review* currently does not offer subscriptions. Instead it is distributed free of charge at institutions across Canada. If you would like to see the *Review* at your school, please contact the editors.

ISSN 1927-1417 (Print) ISSN 1927-1425 (Online)

The Waterloo Mathematics Review
Volume 2, Issue 1

Winter 2012

Contents

Remarks

Articles

Remarks

From the Editors

Dear Reader,

This issue marks the first issue of the *Waterloo Mathematics Review* without major involvement from, our now editor emeritus, Edgar Bering. This issue represents a transition period in the editorship and, as such, the production of Volume II Issue II has been a little delayed. We are also excited to have Saifuddin Syed join as an editor.

In this issue, we have authors from all over Canada, including schools such as the University of Waterloo, University of Ottawa and University of British Columbia covering a breadth of topics such as Statistical Learning Theory, Number Theory and Applied Mathematics.

Overall, we have been very excited with the quality of the submitted articles from authors all over Canada as well as the hard work of the reviewers from the University of Waterloo. We have grown considerably since first launching the *Waterloo Mathematics Review* and we hope to keep expanding the reach and distribution of the journal even further. Our goal is to continue to produce the *Waterloo Mathematics Review* in the foreseeable future and to continue to uphold the strong precedent set by Volume I.

Regards,
Frank Ban
Saifuddin Syed
Eeshan Wagh
Editors
`editor@mathreview.uwaterloo.ca`

From the CUMC

The Canadian Undergraduate Mathematics Conference (CUMC) is Canada's premier conference showcasing undergraduate research in mathematics related fields. From its beginning at McGill University in 1994, the CUMC has grown over the last 19 years to be one of Canada's largest undergraduate conferences. The conference aims to give students a valuable experience in mathematics, beyond what is available in the usual academic setting. During the conference, students can spend up to five days with other students who share a passion for mathematics. They have the valuable opportunity to give mathematical talks to an audience of their peers. In turn, students will also be exposed to ideas from areas of mathematics outside of their expertise; the conference features students with interests in, but not limited to, computer science, economics, physics, statistics, pure mathematics, and applied mathematics.

Students are invited to give a talk, but are welcome to simply attend and learn from others. As well, students who are not involved in research are invited to talk about anything they are passionate about in mathematics. Students may choose to talk about a particular aspect of mathematics that they find interesting such as: a proof they enjoy working through, a historical aspect of the evolution of mathematics, or any other mathematics topic they wish to share. Students will encounter many new topics from these presentations, and will have the opportunity to both give and receive feedback from their peers throughout the conference. The CUMC is a non-competitive forum for building connections between students of all levels and backgrounds and aims for diversity in its presenters. In fact, the three core principles of the conference are bilingualism, non-competitiveness, and regional diversity.

The conference also features renowned keynote speakers from a variety of disciplines. The keynote speakers who are invited to the CUMC are either prominent research figures or rising stars in their fields, and most importantly are people who care about undergraduate students and their exposure to mathematics.

This year we are proud to welcome Dr. Heinz Bauschke (UBC Okanagan), Dr. Catherine Beauchemin (Ryerson University), Dr. Gerda de Vries (University of Alberta), Dr. Donovan Hare (UBC Okanagan), Dr. Jennifer Hyndman (UNBC), Dr. Dominikus Noll (Université Paul Sabatier), and Dr. Tim Swartz (SFU).

In addition to exposing students to other researchers in mathematics, the CUMC gives undergraduate students a chance to visit a different Canadian mathematics department at the hosting university. The CUMC could not continue without its host schools, and for this reason it is an important tradition that the students attending the conference decide which Canadian university will hold the next conference. We strongly encourage students to consider making a bid to host CUMC 2013 at their school. Student groups interested in organizing the conference must first obtain their University's permission and then make a presentation to students at the CUMC meeting.

This is the largest conference of its kind in North America, and we hope that you will get involved and join others from across the country. All attending students will leave with new experiences, new understanding, new ideas, and a new passion for mathematics!

For further information on the conference, please visit our website at cumc.math.ca or contact us with any questions at cumc.2012@ubc.ca.

Bounding the Fat Shattering Dimension of a Composition Function Class Built Using a Continuous Logic Connective

Hubert Haoyang Duan
University of Ottawa
hduan065@uottawa.ca

Abstract: The paper deals with an important combinatorial parameter of a function class, the Fat Shattering dimension. An important known result in statistical learning theory is that a function class is distribution-free Probably Approximately Correct learnable if it has finite Fat Shattering dimension on every scale.

As the main new result, we explore the construction of a new function class from a collection of existing ones, obtained by forming compositions with a continuous logic connective (a uniformly continuous function from the unit hypercube to the unit interval). Vidyasagar had proved that such a composition function class has finite Fat Shattering dimension of all scales if the classes in the original collection do; however, no estimates of the dimension were known. Using results by Mendelson-Vershynin and Talagrand, we bound the Fat Shattering dimension of scale ϵ of this new function class in terms of a sum of the Fat Shattering dimensions of the collection's classes.

1 Introduction

In the area of statistical learning theory, the Probably Approximately Correct (PAC) learning model formalizes the notion of learning by using sample data points to produce valid hypotheses through algorithms.

Our main new result provides an upper bound on the Fat Shattering dimension of a function class, which consists of functions from a domain X to the unit interval $[0, 1]$, built using a continuous logic connective. An introduction to PAC learning is included in the paper to provide all the necessary prerequisites for stating our result. Hence, we first introduce the PAC learning model applied to learning a concept class $\mathcal{C}$, a collection of subsets of X, and more generally, a function class $\mathcal{F}$. We also explain the Vapnik-Chervonenkis and the Fat Shattering dimensions and cover some known results relating learning under this model to these dimensions.

This paper involves mostly concepts from analysis and some concepts from probability theory; the reader is recommended to have a good understanding of basic notions in measure theory.

Outline of Paper

Section 2 provides a brief overview of measure theory and analysis. In Section 3, we give two definitions of PAC learning, one for a concept class $\mathcal{C}$ and the other for a function class $\mathcal{F}$. Then, in Sections 4 and 5, we explore two combinatorial parameters, the Vapnik-Chervonenkis (VC) dimension and the Fat Shattering dimension of scale ϵ, for $\mathcal{C}$ and $\mathcal{F}$, respectively. We also discuss how these dimensions relate to the PAC learnability of concept and function classes.

In Section 6, as the main original result of our research, given function classes $\mathcal{F}_1, \ldots, \mathcal{F}_k$ and a "continuous logic connective" (that is, a continuous function $u : [0, 1]^k \to [0, 1]$), we consider the construction of a new composition function class $u(\mathcal{F}_1, \ldots, \mathcal{F}_k)$, consisting of functions $u(f_1, \ldots, f_k)$ defined by

$$u(f_1, \ldots, f_k)(x) = u(f_1(x), \ldots, f_k(x))$$

for $f_i \in \mathcal{F}_i$. We then bound the Fat Shattering dimension of scale ϵ of this class in terms of a sum of the Fat Shattering dimensions of scale $\delta(\epsilon, k)$ of $\mathcal{F}_1, \ldots, \mathcal{F}_k$, where $\delta(\epsilon, k)$ only depends on ϵ and k. There is a previously known analogous estimate for a composition of concept classes built using a usual connective of classical logic [Vid97]. We deduce our new bound using results from Mendelson-Vershynin and Talagrand.

In this paper, any propositions or examples given with proofs, unless mentioned otherwise, are done by us and are independent of any sources.

2 Brief Overview of Measure Theory and Analysis

This section lists some definitions and results in measure theory and analysis, found in standard textbooks, such as [Doo94], [Vid97], and [AC05], which are used in this paper.

Probability Spaces

A *measurable space* $(X, \mathcal{S})$ is a set X equipped with a *σ-algebra* $\mathcal{S}$, a non-empty collection of subsets of X closed under complements and countable unions. If $(X, \mathcal{S})$ and $(Y, \mathcal{T})$ are two measurable spaces, a function $f : X \to Y$ is called *measurable* if $f^{-1}(T) \in \mathcal{S}$ for all $T \in \mathcal{T}$.

Suppose $(X, \mathcal{S})$ is a measurable space; a *measure* is a function $\mu : \mathcal{S} \to \mathbb{R}^+ = \{r \in \mathbb{R} : r \geq 0\}$ satisfying $\mu(\emptyset) = 0$ and

$$\mu\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} \mu(A_i),$$

for every collection $\{A_i \in S : i \in \mathbb{N}\}$ of pairwise disjoint sets. The triple $(X, \mathcal{S}, \mu)$ is called a *measure space*. If in addition, μ satisfies $\mu(X) = 1$, then μ is a *probability measure* and $(X, \mathcal{S}, \mu)$ is called a *probability space*.

Given a probability space $(X, \mathcal{S}, \mu)$, one can measure the difference between two subsets $A, B \in \mathcal{S}$ of X by looking at their symmetric difference $A \triangle B = (A \cup B) \setminus (A \cap B)$. More generally, given two measurable functions $f, g : X \to [0, 1]$, one can look at the expected value of their absolute difference by integrating with respect to μ:

$$\int_X |f(x) - g(x)| \, d\mu(x).$$

This paper does not go into any details involving the Lebesgue integral nor does it discuss any integrability or measurability issues; we assume that integration of measurable functions to the real numbers, which is a measure space, makes sense and is linear and order-preserving.

Validating hypotheses in the PAC learning model uses the idea of measuring the symmetric difference of two subsets of a probability space $(X, \mathcal{S}, \mu)$ and calculating the expected value of the difference of $f, g : X \to [0, 1]$. The structure of metric spaces arises naturally from these two notions.

Metric Spaces

A *metric space* (M, d) is a set M equipped with a *metric* $d : M \times M \to \mathbb{R}^+$, which is symmetric and satisfies the triangle inequality and the condition that $d(m_1, m_2) = 0$ if and only if $m_1 = m_2$. Given a metric space (M, d), a *metric sub-space* of M (which is a metric space in its own right) is a nonempty subset $M' \subseteq M$ equipped with the distance $d_{|_{M'}}$, the restriction of d to M'.

A *normed vector space* (V, ρ) is a vector space V over $\mathbb{R}$ equipped with a *norm* $\rho : V \to \mathbb{R}^+$ satisfying

1. $\rho(v_1) = 0$ if and only if $v_1 = 0$
2. $\rho(rv_1) = |r|\rho(v_1)$
3. $\rho(v_1 + v_2) \leq \rho(v_1) + \rho(v_2)$

for all $v_1, v_2 \in V$ and $r \in \mathbb{R}$. The structure of a metric space exists in every normed vector space since the function $d : V \times V \to \mathbb{R}^+$ defined by $d(u, v) = \rho(u - v)$ is always a metric on V. In this case, d is called the *metric induced by the norm* ρ on V.

The following subsection provides a few examples of metric spaces which will be encountered in this paper.

Examples of Metric Spaces

The real numbers $(\mathbb{R}, \rho)$, with the absolute value norm $\rho(r) = |r|$ for $r \in \mathbb{R}$, is a normed vector space so $\mathbb{R}$ can be equipped with the metric $d(r, r') = \rho(r - r') = |r - r'|$. The unit interval $[0, 1]$ is a subset of $\mathbb{R}$, so it is a metric sub-space of $(\mathbb{R}, d)$.

In addition, given a probability space $(X, \mathcal{S}, \mu)$, the set V of all bounded measurable functions from X to $\mathbb{R}$ is a vector space, with point-wise addition and scalar multiplication. The function $\rho : V \to \mathbb{R}^+$ defined by

$$\rho(f) = \sqrt{\left(\int_X (f(x))^2 d\mu(x)\right)}$$

is a norm on V if any two functions $f, g : X \to \mathbb{R}$ which agree on a subset of X with full measure, $\mu(\{x \in X : f(x) = g(x)\}) = 1$, are identified via an equivalence relation. The norm ρ is called the $L_2(\mu)$ *norm* on V and we normally write $||f||_2 = \rho(f)$ for $f \in V$. As a result, V can be turned into a metric space.

Example 2.1. Following the notations in the paragraph above, V is a metric space with distance d defined by

$$d(f, g) = ||f - g||_2 = \sqrt{\left(\int_X (f(x) - g(x))^2 d\mu(x)\right)}.$$

Write $[0, 1]^X$ for the set of all measurable functions from a probability space $(X, \mathcal{S}, \mu)$ to $[0, 1]$. Then, it is a metric sub-space of V with distance induced by the $L_2(\mu)$ norm on V, restricted of course to $[0, 1]^X$.

Given metric spaces $(M_1, d_1), \ldots, (M_k, d_k)$, their product $M_1 \times \ldots \times M_k$ always has a natural metric structure, defined as follows.

Example 2.2. If $(M_1, d_1), \ldots, (M_k, d_k)$ are metric spaces, then their product $M_1 \times \ldots \times M_k$ is a metric space with distance d^2 defined by

$$d^2((m_1, \ldots, m_k), (m'_1, \ldots, m'_k)) = \sqrt{((d_1(m_1, m'_1))^2 + \ldots + (d_k(m_k, m'_k))^2)}.$$

The distance d^2 is normally referred to as the L_2 *product distance* on $M_1 \times \ldots \times M_k$.

Consequently, the set $[0, 1]^k$, which denotes the set-theoretic product $[0, 1] \times \ldots \times [0, 1]$, is then a metric space with the L_2 product distance. Also, following Examples 2.1 and 2.2, if $\mathcal{F}_1, \ldots, \mathcal{F}_k$ are sets of measurable functions from a probability space $(X, \mathcal{S}, \mu)$ to the unit interval, then $\mathcal{F}_i \subseteq [0, 1]^X$ for each $i = 1, \ldots, k$. Therefore, the product $\mathcal{F}_1 \times \ldots \times \mathcal{F}_k$ is a metric space with the L_2 distance as well.

3 The Probably Approximately Correct Model

Let $(X, \mathcal{S})$ be a measurable space. A *concept class* $\mathcal{C}$ on X is a subset of $\mathcal{S}$, and an element $A \in \mathcal{C}$, which is a measurable subset of X, is called a *concept*. A *function class* $\mathcal{F}$ is a collection of measurable functions from X to the unit interval $[0, 1]$. Unless stated otherwise, from this section onwards, the following notations will be used:

1. $X = (X, \mathcal{S})$: a *measurable space*
2. μ: a *probability measure* $\mathcal{S} \to \mathbb{R}^+$
3. $\mathcal{C}$: a *concept class* and $\mathcal{F}$: a *function class*

This section provides the definitions of learning $\mathcal{C}$ and $\mathcal{F}$ in the Probably Approximately Correct (PAC) learning model, introduced in 1984 by Valiant.

Concept class PAC learning involves producing a valid hypothesis for every concept $A \in \mathcal{C}$ by first drawing random points, forming a training sample, from X labeled with whether these points are contained in A. In other words, a labeled sample of m points $x_1, \dots, x_m \in X$ for A consists of these points and the evaluations $\chi_A(x_1), \dots, \chi_A(x_m)$ of the indicator function $\chi_A : X \to \{0, 1\}$, where

$$\chi_A(x) = 1 \text{ if and only if } x \in A.$$

The set of all labeled samples of m points can then be identified with $(X \times \{0, 1\})^m$, and producing a hypothesis for A with a labeled sample is exactly the process of associating the sample to a concept $H \in \mathcal{C}$ (i.e. this process is a function from the set of all labeled samples to the concept class).

Here is the precise definition of a concept class being learnable.

Definition 3.1 ([Val84]). A concept class $\mathcal{C}$ is *distribution-free Probably Approximately Correct learnable* if there exists a function (a learning algorithm) $\mathcal{L} : \cup_{m \in \mathbb{N}} (X \times \{0, 1\})^m \to \mathcal{C}$ with the following property: for every $\epsilon > 0$, for every $\delta > 0$, there exists a $M \in \mathbb{N}$ such that for every $A \in \mathcal{C}$, for every probability measure μ, for every $m \geq M$, for any $x_1, \dots, x_m \in X$, we have $\mu(H_m \triangle A) < \epsilon$ with confidence at least $1 - \delta$, where $H_m = \mathcal{L}((x_1, \chi_A(x_1)), \dots, (x_m, \chi_A(x_m)))$.

Confidence of at least $1 - \delta$ in the definition above, keeping to the same notations, simply means that the (product) measure of the set of all m-tuples $(x_1, \dots, x_m) \in X^m$, where $\mu(H_m \triangle A) < \epsilon$ for $H_m = \mathcal{L}((x_1, \chi_A(x_1)), \dots, (x_m, \chi_A(x_m)))$, is at least $1 - \delta$. An equivalent statement to $\mathcal{C}$ being distribution-free PAC learnable is that for every $\epsilon, \delta > 0$, there exists $M \in \mathbb{N}$ such that for every $A \in \mathcal{C}$, probability measure μ, and $m \geq M$,

$$\mu^m(\{(x_1, \dots, x_m) \in X^m : \mu(H_m \triangle A) \geq \epsilon\}) \leq \delta,$$

for $H_m = \mathcal{L}((x_1, \chi_A(x_1)), \dots, (x_m, \chi_A(x_m)))$. (The symbol μ^m denotes the product measure on X^m; the reader can refer to [Doo94] for the details.)

A concept class $\mathcal{C}$ is distribution-free learnable in the PAC learning model if a hypothesis H can always be constructed from an algorithm $\mathcal{L}$ for every concept $A \in \mathcal{C}$, using any labeled sample for A, such that the measure of their symmetric difference $H \triangle A$ is arbitrarily small with respect to every probability measure and with arbitrarily high confidence, as long as the sample size is large enough.

Every concept $A \in \mathcal{C}$ is a subset of X and can be associated to its indicator function $\chi_A : X \to \{0, 1\}$. Even more generally, χ_A is a function from X to $[0, 1]$; in other words, every concept class $\mathcal{C}$ can be identified as a function class $\mathcal{F}_\mathcal{C} = \{\chi_A : X \to [0, 1] : A \in \mathcal{C}\}$, so it is natural to generalize Definition 3.1 for any function class $\mathcal{F}$.

Definition 3.1 involves the symmetric difference of two concepts and its generalization to measurable functions $f, g : X \to [0, 1]$ is the expected value of their absolute difference $\mathbb{E}_\mu(f, g)$, as seen in the previous section:

$$\mathbb{E}_\mu(f, g) = \int_X |f(x) - g(x)| \, d\mu(x).$$

A simple exercise can show that if $f, g \in [0, 1]^X$ are indicator functions of two concepts $A, B \subseteq X$, then $\mathbb{E}_\mu(f, g)$ coincides with the measure of their symmetric difference: $\mathbb{E}_\mu(f, g) = \mu(A \triangle B)$, where $f = \chi_A$ and $g = \chi_B$.

With this generalization of the symmetric difference, distribution-free PAC learning for any function class can be defined. In the context of function class learning, a labeled sample of m points $x_1, \dots, x_m \in X$

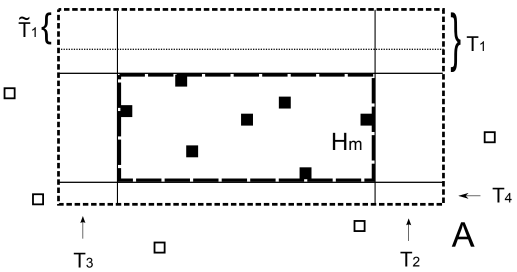

Figure 3.1: Learning an axis-aligned rectangle.

for a function $f \in \mathcal{F}$ consists of these points and the evaluations $f(x_1), \ldots, f(x_m)$. Then, the set of all labeled samples of m points can be identified with $(X \times [0,1])^m$, and producing a hypothesis is the process of associating a labeled sample to a function $H \in \mathcal{F}$ (just as in concept class learning).

Definition 3.2 ([Vid97]). A function class $\mathcal{F}$ is *distribution-free Probably Approximately Correct learnable* if there exists a function (a learning algorithm) $\mathcal{L} : \cup_{m \in \mathbb{N}}(X \times [0,1])^m \to \mathcal{F}$ with the following property: for every $\epsilon > 0$, for every $\delta > 0$, there exists a $M \in \mathbb{N}$ such that for every $f \in \mathcal{F}$, for every probability measure μ, for every $m \geq M$, for any $x_1, \ldots, x_m \in X$, we have $\mathbb{E}_\mu(H_m, f) < \epsilon$ with confidence at least $1 - \delta$, where $H_m = \mathcal{L}((x_1, f(x_1)), \ldots, (x_m, f(x_m)))$.

Both definitions of PAC learning contain the ϵ and δ parameters. The accuracy error ϵ is used because the hypothesis cannot be, in general, expected to have zero error - only an arbitrarily small error. The risk parameter δ exists because there is no guarantee that any collection of sufficiently large training points leads to a valid hypothesis; the learning algorithm is only expected to produce a valid hypothesis with the sample points with confidence at least $1 - \delta$. Hence, the name "Probably (δ) Approximately (ϵ) Correct" is used [KV94].

The following example illustrates that the set of all axis-aligned rectangles in $\mathbb{R}^2$ is distribution-free PAC learnable. Both the statement and its proof can be found in Chapter 3 of [Vid97] and Chapter 1 of [KV94].

Example 3.1. In $X = \mathbb{R}^2$, the concept class $\mathcal{C} = \{[a,b] \times [c,d] : a,b,c,d \in \mathbb{R}\}$ is distribution-free PAC learnable.

Proof. Let $\epsilon, \delta > 0$. Given a concept A and any sample of m training points $x_1, \ldots, x_m \in X$, define the hypothesis concept H_m to be the intersection of all rectangles containing only training points x_i such that $\chi_A(x_i) = 1$. In other words, H_m is the smallest rectangle that contains only the sample points *in* A.

Let μ be any probability measure, and in fact, $H_m \triangle A = A \setminus H_m$, which can be broken down into four sections $T_1, \ldots, T_4$. If we can conclude that

$$\mu\left(\bigcup_{i=1}^{4} T_i\right) < \epsilon,$$

with confidence at least $1 - \delta$, then the proof is complete.

Consider the top section T_1 and define $\tilde{T_1}$ to be the rectangle along the top parts of A whose measure is exactly $\epsilon/4$. The event $\tilde{T_1} \subseteq T_1$, which is equivalent to $\mu(T_1) \geq \epsilon/4$, holds exactly when no points in the sample $x_1, \ldots, x_m$ fall in $\tilde{T_1}$, and the probability of this event (which is the measure of all such m-tuples of $(x_1, \ldots, x_m) \in X^m$ where $x_i \notin \tilde{T_1}$ for all $i = 1, \ldots, m$) is $(1 - \epsilon/4)^m$. Similarly, the same holds for the other three sections $T_2, \ldots, T_4$. Therefore, the probability that there exists at least one T_i such that $\mu(T_i) \geq \epsilon/4$, where $i \in \{1, \ldots, 4\}$, is at most $4(1 - \epsilon/4)^m$. Hence, as long as we pick m large enough that $4(1 - \epsilon/4)^m \leq \delta$, with confidence (probability) at least $1 - \delta$, $\mu(T_i) < \epsilon/4$ for every $i = 1, \ldots, 4$ and thus,

$$\mu(H_m \triangle A) = \mu\left(\bigcup_{i=1}^{4} T_i\right) \leq \mu(T_1) + \ldots + \mu(T_4) < 4\left(\frac{\epsilon}{4}\right) = \epsilon.$$

Please note that this argument, though very intuitive, actually requires the classical Glivenko-Cantelli theorem, see e.g. [Bil95]. Figure 3.1 provides a visual illustration of the rectangles.

In summary, as long as $m \geq (4/\epsilon)\ln(4/\delta)$, with confidence at least $1-\delta$, $\mu(H_m \triangle A) < \epsilon$. We note that this estimate of the sample size only depends on ϵ and δ, so $\mathcal{C}$ is indeed distribution-free PAC learnable. □

In the next section, a fundamental theorem which characterizes concept class distribution-free PAC learning will be stated. However, in order to state this theorem, the notion of shattering, which is essential in learning theory, must be introduced.

4 The Vapnik-Chervonenkis Dimension

The Vapnik-Chervonenkis dimension is a combinatorial parameter which is defined using the notion of shattering, developed first in 1971 by Vapnik and Chervonenkis.

Definition 4.1 ([VC71]). Given any set X and a collection $\mathcal{A}$ of subsets of X, the collection $\mathcal{A}$ *shatters* a finite subset $S \subseteq X$ if for every $B \subseteq S$, there exists $A \in \mathcal{A}$ such that $A \cap S = B$.

There is an equivalent condition, which is sometimes easier to work with, to shattering, expressed in terms of characteristic functions of subsets of X.

Proposition 4.1. The collection $\mathcal{A}$ shatters a subset $S = \{x_1, \ldots, x_n\} \subseteq X$ if and only if for every $e = (e_1, \ldots, e_n) \in \{0,1\}^n$, there exists $A \in \mathcal{A}$ such that $\chi_A(x_i) = e_i$, for all $i = 1, \ldots, n$.

Definition 4.2 ([VC71]). The *Vapnik-Chervonenkis (VC) dimension* of the collection $\mathcal{A}$, denoted by $\mathrm{VC}(\mathcal{A})$, is defined to be the cardinality of the largest finite subset $S \subseteq X$ shattered by $\mathcal{A}$. If $\mathcal{A}$ shatters arbitrarily large finite subsets of X, then the VC dimension of $\mathcal{A}$ is defined to be ∞.

The VC dimension is defined for every collection $\mathcal{A}$ of subsets of any set X, so in particular, $X = (X, \mathcal{S})$ can be a measurable space and $\mathcal{A} = \mathcal{C}$ can be a concept class.

The following is an example, which we believe to be original, illustrating the calculation of the VC dimension for a concept class in the context of $X = \mathbb{R}^n$. In order to prove the VC dimension of a concept class $\mathcal{C}$ is d, we must provide a subset $S \subseteq X$ with cardinality d which is shattered by $\mathcal{C}$ and prove that no subset with cardinality $d+1$ can be shattered by $\mathcal{C}$. The reader can refer to [KV94] and [Pes10b] for more examples on calculating VC dimensions.

Example 4.1. Consider the space $X = \mathbb{R}^n$. A hyperplane $H_{\vec{a},b}$ is defined by a nonzero vector $\vec{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and a scalar $b \in \mathbb{R}$:

$$\begin{aligned} H_{\vec{a},b} &= \{\vec{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n : \vec{x} \cdot \vec{a} = b\} \\ &= \{\vec{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 a_1 + \ldots + x_n a_n = b\}. \end{aligned}$$

Write $\mathcal{C}$ as the set of all hyperplanes: $\mathcal{C} = \{H_{\vec{a},b} : \vec{a} \in \mathbb{R}^n \setminus \{\vec{0}\}, b \in \mathbb{R}\}$. Then $\mathrm{VC}(\mathcal{C}) = n$.

Proof. Consider the subset $S = \{\vec{e}_1, \ldots, \vec{e}_n\} \subseteq \mathbb{R}^n$, where $\vec{e}_i$ is the vector with 1 on the i-th component and 0 everywhere else. Suppose $B \subseteq S$ and there are two cases to consider:

1. If $B = \emptyset$, then let $\vec{a} = (1, 1, \ldots, 1) \in \mathbb{R}^n$ and the hyperplane $H_{\vec{a},-1} = \{\vec{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 + \ldots + x_n = -1\}$ is disjoint from S.

2. If $B \neq \emptyset$, then set $\vec{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n \setminus \{\vec{0}\}$, where $a_i = \chi_B(\vec{e}_i)$. Then the hyperplane $H_{\vec{a},1} = \{\vec{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 a_1 + \ldots + x_n a_n = 1\}$ satisfies

$$H_{\vec{a},1} \cap S = B.$$

Moreover, no subset $S = \{\vec{x}_1, \dots, \vec{x}_n, \vec{x}_{n+1}\} \subseteq \mathbb{R}^n$ with cardinality $n+1$ can be shattered by $\mathcal{C}$. At best, there exists a unique hyperplane $H_{\vec{a},b}$ containing n of these points, say $\{\vec{x}_1, \dots, \vec{x}_n\}$, so if $\vec{x}_{n+1} \in H_{\vec{a},b}$, then there are no hyperplanes that include $\vec{x}_1, \dots, \vec{x}_n$, but not $\vec{x}_{n+1}$. Otherwise, if $\vec{x}_{n+1} \notin H_{\vec{a},b}$, then there are no hyperplanes that include $\vec{x}_1, \dots, \vec{x}_n, \vec{x}_{n+1}$. □

The VC dimension is central to the PAC learning model for concept classes. In fact, the PAC learnability of a concept class is completely determined by its VC dimension.

4.1 Characterization of Concept Class PAC Learning

The following is one of the main theorems concerning PAC learning, whose proof results from Vapnik and Chervonenkis' paper [VC71] in 1971 and the 1989 paper [BEHW89] by Blumer et al.

Theorem 4.2 ([VC71] and [BEHW89]). Let $\mathcal{C}$ be a concept class of a measurable space $(X, \mathcal{S})$. The following are equivalent:

1. $\mathcal{C}$ is distribution-free Probably Approximately Correct learnable.

2. $\mathrm{VC}(\mathcal{C}) < \infty$.

Both directions of the proof for this result require expressing the number of sample training points required for learning in terms of the VC dimension of $\mathcal{C}$; a crucial lemma used in the proof is Sauer's Lemma, seen in [Sau72]. Given a concept class $\mathcal{C}$ with finite VC dimension, the lemma states that the growth of $|\{A \cap C : C \in \mathcal{C}\}|$ for any finite set A, with $|A| = n$, is bounded above by a polynomial function in n as n grows to infinity.

Using Theorem 4.2, one can more easily determine whether a given concept class is distribution-free PAC learnable.

Example 4.2. The set of all hyperplanes $\mathcal{C} = \{H_{\vec{a},b} : \vec{a} \in \mathbb{R}^n \setminus \{\vec{0}\}, b \in \mathbb{R}\}$, as defined in Example 4.1, is distribution-free PAC learnable.

Every concept class $\mathcal{C}$ can be viewed as a function class $\mathcal{F}_{\mathcal{C}} = \{\chi_A : X \to [0,1] : A \in \mathcal{C}\}$, as seen in Section 3, so a natural question is whether the notion of shattering can be generalized. Indeed, the next section introduces the Fat Shattering dimension of scale ϵ, which is a generalization of the VC dimension.

5 The Fat Shattering Dimension

Let $\epsilon > 0$ from this section onwards. A combinatorial parameter which generalizes the Vapnik-Chervonenkis dimension is the Fat Shattering dimension of scale ϵ, defined first by Kearns and Schapire in 1994.

This dimension, assigned to function classes, involves the notion of *ϵ-shattering*, but similar to the notion of (regular) shattering, it can be defined for any collection of functions $f : X \to [0,1]$, where X is any set. For the sake of this paper, the following sections (still) assume $X = (X, \mathcal{S})$ is a measurable space and the collection of functions is a function class $\mathcal{F}$.

Definition 5.1 ([KS94]). Let $\mathcal{F}$ be a function class. Given a subset $S = \{x_1, \dots, x_n\} \subseteq X$, the class $\mathcal{F}$ *ϵ-shatters S*, with *witness* $c = (c_1, \dots, c_n) \in [0,1]^n$, if for every $e \in \{0,1\}^n$, there exists $f \in \mathcal{F}$ such that

$$f(x_i) \geq c_i + \epsilon \text{ for } e_i = 1, \text{ and } f(x_i) \leq c_i - \epsilon \text{ for } e_i = 0.$$

Figure 5.1 illustrates the notion of ϵ-shattering for the subset $S = \{x_1, \dots, x_6\}$, with witness $c = (c_1, \dots, c_6)$. Given the binary vector $e = (101011)$, there is a function $f \in \mathcal{F}$ that passes above $c_1 + \epsilon, c_3 + \epsilon, c_5 + \epsilon, c_6 + \epsilon$ at the points x_1, x_3, x_5, x_6, respectively, but passes below $c_2 - \epsilon, c_4 - \epsilon$ at x_2, x_4.

Definition 5.2 ([KS94]). The *Fat Shattering dimension of scale $\epsilon > 0$* of $\mathcal{F}$, denoted by $\mathrm{fat}_\epsilon(\mathcal{F})$, is defined to be the cardinality of the largest finite subset of X that can be ϵ-shattered by $\mathcal{F}$. If $\mathcal{F}$ can ϵ-shatter arbitrarily large finite subsets, then the Fat Shattering dimension of scale ϵ of $\mathcal{F}$ is defined to be ∞.

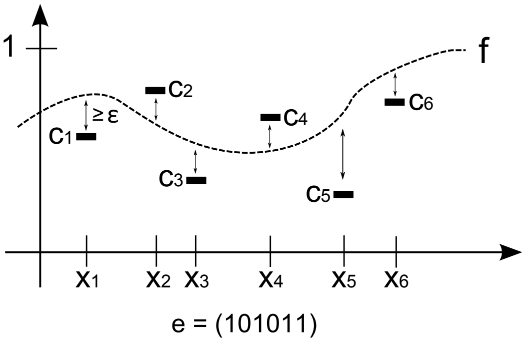

Figure 5.1: Diagram of ϵ-shattering.

When the function class $\mathcal{F}$ consists of only functions taking values in $\{0,1\}$, then the Fat Shattering dimension of any scale $\epsilon \leq 1/2$ of $\mathcal{F}$ agrees with the VC dimension of the corresponding collection of subsets of X, induced by the (indicator) functions in $\mathcal{F}$.

With the generalization from a concept class to a function class, a natural question is whether the finiteness of the Fat Shattering dimension of all scales ϵ for a function class $\mathcal{F}$ is equivalent to $\mathcal{F}$ being distribution-free PAC learnable. This question is addressed in the following subsection.

5.1 Sufficient Condition for Function Class PAC Learning

One direction of Theorem 4.2 can be generalized and stated in terms of the Fat Shattering dimension of scale ϵ of a function class.

Theorem 5.1 ([ABDCBH97] and [Vid97]). Let $\mathcal{F}$ be a function class. If $\text{fat}_\epsilon(\mathcal{F}) < \infty$ for all $\epsilon > 0$, then $\mathcal{F}$ is distribution-free PAC learnable.

However, the converse to Theorem 5.1 is false. There exists a distribution-free PAC learnable function class with infinite Fat Shattering dimension of some scale ϵ.

In fact, for every concept class $\mathcal{C}$ with cardinality $\aleph_0$ or $2^{\aleph_0}$, there is an associated function class $\mathcal{F}_\mathcal{C}$ defined as follows. Set up a bijection $b : \mathcal{C} \to [0,1/3]$ or to $[0,1/3] \cap \mathbb{Q}$, depending on the cardinality of $\mathcal{C}$, and for every $A \in \mathcal{C}$, define a function $f_A : X \to [0,1]$ by

$$f_A(x) = \chi_A(x) + (-1)^{\chi_A(x)} b(A).$$

Now, write $\mathcal{F}_\mathcal{C} = \{f_A : A \in \mathcal{C}\}$. Note that $\mathcal{F}_\mathcal{C}$ can be thought of the collection of all indicator functions of $A \in \mathcal{C}$, except that each "indicator" function f_A has two unique identifying points $b(A)$ and $1 - b(A)$, instead of simply 0 and 1. The following proposition provides many counterexamples to the converse of Theorem 5.1, which are much simpler than the one found in [Vid97].

The construction of the function class $\mathcal{F}_\mathcal{C}$ and the proposition below are developed from an idea of Example 2.10 in [Pes10a].

Proposition 5.2. Let $\mathcal{C}$ be a concept class. The associated function class $\mathcal{F}_\mathcal{C} = \{f_A : A \in \mathcal{C}\}$, defined in the previous paragraph, is always distribution-free PAC learnable; this class has infinite Fat Shattering dimension of all scales $\epsilon < 1/6$ if $\mathcal{C}$ has infinite VC dimension.

Proof. The function class $\mathcal{F}_\mathcal{C}$ is distribution-free PAC learnable because every function $f_A \in \mathcal{F}_\mathcal{C}$ can be uniquely identified with just one point $x_0 \in X$ in any labeled sample: $f_A(x_0) \in \{b(A), 1 - b(A)\}$ uniquely determines A and thus, f_A.

Furthermore, suppose $\mathcal{C}$ has infinite VC dimension. Let $n \in \mathbb{N}$ be arbitrary and because $\text{VC}(\mathcal{C}) = \infty$, there exists $S = \{x_1, \ldots, x_n\}$ such that $\mathcal{C}$ shatters S. Suppose $\epsilon < 1/6$ and we claim that $\mathcal{F}_\mathcal{C}$ ϵ-shatters S

with witness $c = (0.5, \ldots, 0.5) \in [0,1]^n$. Indeed, let $e \in \{0,1\}^n$ and there exists $A \in \mathcal{C}$ such that

$$\chi_A(x_i) = e_i,$$

for all $i = 1, \ldots, n$, by Proposition 4.1. As a result,

$$f_A(x_i) = 1 - b(A) \geq 0.5 + \epsilon \text{ for } e_i = 1$$

and

$$f_A(x_i) = b(A) \leq 0.5 - \epsilon \text{ for } e_i = 0.$$

Consequently, $\mathcal{F}_\mathcal{C}$ has infinite Fat Shattering dimension of all scales $\epsilon < 1/6$. □

One research topic we would like to consider in the future is to come up with a new combinatorial parameter for a function class, related to the notion of shattering, which would characterize PAC distribution-free learning. This new parameter would have to solve the problem of unique identifications of functions, a problem that does not occur with concept classes.

The next section explains the main result of our research: bounding the Fat Shattering dimension of scale ϵ of a composition function class which is built with a continuous logic connective.

6 The Fat Shattering Dimension of a Composition Function Class

The goals of this section are to construct a new function class from old ones by means of a continuous logic connective and to bound the Fat Shattering dimension of scale ϵ of the new function class in terms of the dimensions of the old ones. The following subsection provides this construction, which can be found in Chapter 4 of [Vid97], in the context of concept classes using a connective of classical logic.

6.1 A Review of the Construction in the Context of Concept Classes

Let $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ be concept classes, where $k \geq 2$, and let $u : \{0,1\}^k \to \{0,1\}$ be any function, commonly known as a connective of classical logic. A new collection of subsets of X arises from $\mathcal{C}_1, \ldots, \mathcal{C}_k$ as follows.

As mentioned earlier in this paper, every element $A \in \mathcal{C}_i$ can be identified as a binary function $f : X \to \{0,1\}$, namely its characteristic function $f = \chi_A$, and vice versa. Now, for any k functions $f_1, \ldots, f_k : X \to \{0,1\}$, where $f_i \in \mathcal{C}_i$ with $i = 1, \ldots, k$, consider a new function $u(f_1, \ldots, f_k) : X \to \{0,1\}$ defined by

$$u(f_1, \ldots, f_k)(x) = u(f_1(x), \ldots, f_k(x)).$$

The set of all possible $u(f_1, \ldots, f_k)$, denoted by $u(\mathcal{C}_1, \ldots, \mathcal{C}_k)$, is given by

$$u(\mathcal{C}_1, \ldots, \mathcal{C}_k) = \{u(f_1, \ldots, f_k) : f_i \in \mathcal{C}_i\}.$$

For instance, when $k = 2$, we can consider the "Exclusive Or" connective $\oplus : \{0,1\}^2 \to \{0,1\}$ defined by

$$p \oplus q = (p \wedge \neg q) \vee (\neg p \wedge q),$$

which corresponds to the symmetric difference operation. Then, our new concept class constructed from $\mathcal{C}_1$ and $\mathcal{C}_2$ is

$$\{A_1 \triangle A_2 : A_1 \in \mathcal{C}_1, A_2 \in \mathcal{C}_2\}.$$

The next known theorem states that if $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ all have finite VC dimension to start with, then regardless of u, the new collection $u(\mathcal{C}_1, \ldots, \mathcal{C}_k)$ always has finite VC dimension.

Theorem 6.1 ([Vid97]). Let $k \geq 2$. Suppose $\mathcal{C}_1, \ldots, \mathcal{C}_k$ are concept classes, each viewed as a collection of binary functions, and $u : \{0,1\}^k \to \{0,1\}$ is any function. If the VC dimension of $\mathcal{C}_i$ is finite for all $i = 1, \ldots, k$. Then there exists a constant $\alpha = \alpha_k$, which depends only on k, such that

$$\mathrm{VC}(u(\mathcal{C}_1, \ldots, \mathcal{C}_k)) < d\alpha_k,$$

where $d = \max\limits_{i=1}^{k} \mathrm{VC}(\mathcal{C}_i)$.

The proof of this theorem can be found in [Vid97] and uses Sauer's Lemma to bound the VC dimension of $u(\mathcal{C}_1, \ldots, \mathcal{C}_k)$. The main objective of our research is to generalize this theorem for function classes, in terms of the Fat Shattering dimension of scale ϵ, but the connective of classical logic u would have to be replaced by a *continuous logic connective*, which is simply a continuous function $u : [0,1]^k \to [0,1]$.

6.2 Construction of New Function Class with Continuous Logic Connective

In first-order logic, there are only two truth-values 0 or 1, so a connective is a function $\{0,1\}^k \to \{0,1\}$ in the classical sense. However, in continuous logic, truth-values can be found anywhere in the unit interval $[0,1]$. Therefore, we should consider a function $u : [0,1]^k \to [0,1]$, which will transform function classes, and require that u be a continuous logic connective. In other words, u should be continuous from the (product) metric space $[0,1]^k$ to the unit interval [YBHU08]; in fact, because u is continuous from a compact metric space to a metric space, it is automatically uniformly continuous.

The following provides the definition of a uniformly continuous function u from any metric space to another, but we must first qualify u with a modulus of uniform continuity.

Definition 6.1 (See e.g. [YBHU08]). A *modulus of uniform continuity* is any function $\delta : (0,1] \to (0,1]$.

Definition 6.2 (See e.g. [YBHU08]). Let (M_1, d_1) and (M_2, d_2) be two metric spaces. A function $u : M_1 \to M_2$ is *uniformly continuous* if there exists (a modulus of uniform continuity) $\delta : (0,1] \to (0,1]$ such that for all $\epsilon \in (0,1]$ and $m_1, m_2 \in M_1$, if $d_1(m_1, m_2) < \delta(\epsilon)$, then $d_2(u(m_1), u(m_2)) < \epsilon$.

Such a δ is called a *modulus of uniform continuity for u.*

Given function classes $\mathcal{F}_1, \ldots, \mathcal{F}_k$ and a uniformly continuous function $u : [0,1]^k \to [0,1]$, consider the new function class $u(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ defined by

$$u(\mathcal{F}_1, \ldots, \mathcal{F}_k) = \{u(f_1, \ldots, f_k) : f_i \in \mathcal{F}_i\},$$

where $u(f_1, \ldots, f_k)(x) = u(f_1(x), \ldots, f_k(x))$ for all $x \in X$, just as in Section 6.1 for concept classes, with $f_i \in \mathcal{F}_i$ and $i = 1, \ldots, k$. Our main result states that the Fat Shattering dimension of scale ϵ of $u(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ is bounded by a sum of the Fat Shattering dimensions of scale $\delta(\epsilon, k)$ of $\mathcal{F}_1, \ldots, \mathcal{F}_k$, where $\delta(\epsilon, k)$ is a function of the modulus of uniform continuity $\delta(\epsilon)$ for u and k. It is a known result, seen in Chapter 5 of [Vid97], that this new class $u(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ has finite Fat Shattering dimension of all scales $\epsilon > 0$ (and thus, it is distribution-free PAC learnable) if each of $\mathcal{F}_1, \ldots, \mathcal{F}_k$ has finite Fat Shattering dimension of all scales, but no bounds were previously known.

6.3 Main Result

Fix $k \geq 2$ and the following theorem is our main new result.

Theorem 6.2. Let $\epsilon > 0$, $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be function classes of X, and $u : [0,1]^k \to [0,1]$ be a uniformly continuous function with modulus of continuity $\delta(\epsilon)$. Then

$$\mathrm{fat}_\epsilon(u(\mathcal{F}_1, \ldots, \mathcal{F}_k)) \leq \left(\frac{K \log(4c'k\sqrt{k}/(\delta(\epsilon/(2c'))\epsilon))}{K' \log(2)} \right) \sum_{i=1}^{n} \mathrm{fat}_{c\frac{\delta(\epsilon/(2c'))\epsilon}{k\sqrt{k}}}(\mathcal{F}_i),$$

where c, c', K, K' are some absolute constants.

Extracting the actual values of these absolute constants is not easy, and we hope to find them in future research. For this reason, comparing the bound in Theorem 6.2 with the existing estimate for the VC dimension of a composition concept class is difficult; however, in statistical learning theory, estimates for function class learning are generally much worse than estimates for concept class learning.

In order to prove Theorem 6.2, for clarity, we will introduce an auxiliary function $\phi : \mathcal{F}_1 \times \ldots \times \mathcal{F}_k \to [0,1]^X$ and prove the following.

Lemma 6.1. Let $\epsilon > 0$. If $u : [0,1]^k \to [0,1]$ is uniformly continuous with modulus of continuity $\delta(\epsilon)$, then the function $\phi : \mathcal{F}_1 \times \ldots \times \mathcal{F}_k \to [0,1]^X$ defined by

$$\phi(f_1, \ldots, f_k)(x) = u(f_1(x), \ldots, f_k(x))$$

is also uniformly continuous with modulus of continuity $\frac{\delta(\epsilon/2)\epsilon}{2k}$, from the metric space $\mathcal{F}_1 \times \ldots \times \mathcal{F}_k$ with distance $\tilde{d}^2$ to $[0,1]^X$. Also, $\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k) = u(\mathcal{F}_1, \ldots, \mathcal{F}_k)$, where the symbol $\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k)$ simply represents the image of ϕ.

Then, we will prove the next lemma, and our main result will follow directly.

Lemma 6.2. Let $\epsilon > 0$, $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be function classes of X, and $\phi : \mathcal{F}_1 \times \ldots \times \mathcal{F}_k \to [0,1]^X$ be uniformly continuous with some modulus of continuity $\delta(\epsilon, k)$, a function of ϵ and k. Then

$$\text{fat}_{c'\epsilon}(\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k)) \leq \left(\frac{K \log(2\sqrt{k}/\delta(\epsilon, k))}{K' \log(2)} \right) \sum_{i=1}^{k} \text{fat}_{c\frac{\delta(\epsilon,k)}{\sqrt{k}}}(\mathcal{F}_i),$$

where c, c', K, K' are some absolute constants.

6.4 Proofs

This subsection provides all the proofs for our main theorem.

Proof of Lemma 6.1. Suppose $u : [0,1]^k \to [0,1]$ is uniformly continuous with a modulus of continuity $\delta(\epsilon)$, where $[0,1]^k$ is a metric space with the L_2 product distance d^2. We claim that the function $\phi : \mathcal{F}_1 \times \ldots \times \mathcal{F}_k \to [0,1]^X$ defined by

$$\phi(f_1, \ldots, f_k)(x) = u(f_1(x), \ldots, f_k(x))$$

is uniformly continuous with modulus of continuity $\frac{\delta(\epsilon/2)\epsilon}{2k}$. Let $\epsilon > 0$ and

$$(f_1, \ldots, f_k), (f'_1, \ldots, f'_k) \in \mathcal{F}_1 \times \ldots \times \mathcal{F}_k.$$

Suppose

$$\begin{aligned} \tilde{d}^2((f_1, \ldots, f_k), (f'_1, \ldots, f'_k)) &= \sqrt{((||f_1 - f'_1||_2)^2 + \ldots + (||f_k - f'_k||_2)^2)} \\ &< \frac{\delta(\epsilon/2)\epsilon}{2k} = \sqrt{\frac{\delta(\epsilon/2)^2(\epsilon/2)^2}{k^2}}. \end{aligned}$$

Hence, for each $i = 1, \ldots, k$,

$$||f_i - f'_i||_2 = \sqrt{\left(\int_X (f_i(x) - f'_i(x))^2 \, d\mu(x) \right)} < \sqrt{\frac{\delta(\epsilon/2)^2(\epsilon/2)^2}{k^2}}.$$

Write $A_i = \{x \in X : |f_i(x) - f'_i(x)| \geq \sqrt{\frac{\delta(\epsilon/2)^2}{k}}\}$ and we must have that $\mu(A_i) < \frac{(\epsilon/2)^2}{k}$, for each $i = 1, \ldots, k$. Otherwise,

$$\begin{aligned}\int_X (f_i(x)-f_i'(x))^2\,d\mu(x) &= \int_{A_i}(f_i(x)-f_i'(x))^2\,d\mu(x)+\int_{X\setminus A_i}(f_i(x)-f_i'(x))^2\,d\mu(x)\\ &\geq \int_{A_i}\left(\sqrt{\frac{\delta(\epsilon/2)^2}{k}}\right)^2 d\mu(x)+\int_{X\setminus A_i}(f_i(x)-f_i'(x))^2\,d\mu(x)\\ &= \mu(A_i)\left(\sqrt{\frac{\delta(\epsilon/2)^2}{k}}\right)^2+\int_{X\setminus A_i}(f_i(x)-f_i'(x))^2\,d\mu(x)\\ &\geq \frac{(\epsilon/2)^2}{k}\frac{\delta(\epsilon/2)^2}{k}+\int_{X\setminus A_i}(f_i(x)-f_i'(x))^2\,d\mu(x)\\ &\geq \frac{\delta(\epsilon/2)^2(\epsilon/2)^2}{k^2},\end{aligned}$$

which is a contradiction. Now, write $A = A_1 \cup \ldots \cup A_k$ and we have that $X \setminus A = \{x \in X : |f_i(x) - f_i'(x)| < \sqrt{\frac{\delta(\epsilon/2)^2}{k}}$, for all $i = 1, \ldots, k\}$. Suppose $x \in X \setminus A$ and then

$$\begin{aligned}d^2((f_1(x),\ldots,f_k(x)),(f_1'(x),\ldots,f_k'(x))) &= \sqrt{|f_1(x)-f_1'(x)|^2+\ldots+|f_k(x)-f_k'(x)|^2}\\ &< \sqrt{\left(\frac{\delta(\epsilon/2)^2}{k}+\ldots+\frac{\delta(\epsilon/2)^2}{k}\right)} < \delta(\epsilon/2).\end{aligned}$$

Consequently, by the uniform continuity of u, for all $x \in X \setminus A$,

$$|u(f_1(x),\ldots,f_k(x)) - u(f_1'(x),\ldots,f_k'(x))| < \epsilon/2.$$

Finally,

$$\begin{aligned}||\phi(f_1,\ldots,f_k)-\phi(f_1',\ldots,f_k')||_2 &= \sqrt{\left(\int_X (u(f_1(x),\ldots,f_k(x))-u(f_1'(x),\ldots,f_k'(x)))^2\,d\mu(x)\right)}\\ &\leq \sqrt{\left(\int_{X\setminus A} (u(f_1(x),\ldots,f_k(x))-u(f_1'(x),\ldots,f_k'(x)))^2\,d\mu(x)\right)}\\ &+\sqrt{\left(\int_{A} (u(f_1(x),\ldots,f_k(x))-u(f_1'(x),\ldots,f_k'(x)))^2\,d\mu(x)\right)}\\ &< \sqrt{\left(\int_{X\setminus A}(\epsilon/2)^2\,d\mu(x)\right)}+\sqrt{\left(\int_A 1\,d\mu(x)\right)}\\ &\leq (\epsilon/2)+(\epsilon/2)=\epsilon,\end{aligned}$$

as $\mu(A) \leq \sum_{i=1}^k \mu(A_i) \leq k\left(\frac{(\epsilon/2)^2}{k}\right) = (\epsilon/2)^2$. □

Now, in order to prove Lemma 6.2, we first introduce the concept of an ϵ-covering number for any metric space, based on [MV03], and relate this number for a function class to its Fat Shattering dimension of scale ϵ by using results from Mendelson and Vershynin [MV03] and Talagrand [Tal03].

Definition 6.3. Let $\epsilon > 0$ and suppose (M, d) is a metric space. The *ϵ-covering number*, denoted by $N(M, \epsilon, d)$, of M is the minimal number N such that there exists elements $m_1, m_2, \ldots, m_N \in M$ with the property that for all $m \in M$, there exists $i \in \{1, 2, \ldots, N\}$ for which

$$d(m, m_i) < \epsilon.$$

The set $\{m_1, m_2, \ldots, m_N\}$ is called a *(minimal) ϵ-net* of M.

The following proposition relates the ϵ-covering number of a product of metric spaces, with the L_2 product distance d^2, $M_1 \times \ldots \times M_k$ to the $\frac{\epsilon}{\sqrt{k}}$-covering number of each space M_i.

Proposition 6.3. Let $\epsilon > 0$ and suppose $(M_1, d_1), \ldots, (M_k, d_k)$ are metric spaces, each with finite $\frac{\epsilon}{\sqrt{k}}$-covering numbers, $N_i = N(M_i, \frac{\epsilon}{\sqrt{k}}, d_i)$ for $i = 1, \ldots, k$. Then

$$N(M_1 \times \ldots \times M_k, \epsilon, d^2) \leq \prod_{i=1}^{k} N_i.$$

Proof. Let $C_i = \{a_1^i, \ldots, a_{N_i}^i\}$ be a minimal $\frac{\epsilon}{\sqrt{k}}$-net for M_i with respect to distance d_i, where $i = 1, \ldots, k$ and suppose $(a^1, \ldots, a^k) \in M_1 \times \ldots \times M_k$. Then, for each $i = 1, \ldots, k$, there exists $a_{j_i}^i \in C_i$, where $1 \leq j_i \leq N_i$ such that $d_i(a^i, a_{j_i}^i) < \frac{\epsilon}{\sqrt{k}}$. Hence,

$$\begin{aligned} d^2((a^1, \ldots, a^k), (a_{j_1}^1, \ldots, a_{j_k}^k)) &= \sqrt{\left((d_1(a^1, a_{j_1}^1))^2 + \ldots + (d_k(a^k, a_{j_k}^k))^2\right)} \\ &< \sqrt{\left(\left(\frac{\epsilon}{\sqrt{k}}\right)^2 + \ldots + \left(\frac{\epsilon}{\sqrt{k}}\right)^2\right)} = \epsilon, \end{aligned}$$

where each $(a_{j_1}^1, \ldots, a_{j_k}^k) \in C_1 \times \ldots \times C_k$, which has cardinality $\Pi_{i=1}^k N_i$. Therefore, $N(M_1 \times \ldots \times M_k, \epsilon, d^2) \leq \Pi_{i=1}^k N_i$. □

Also, if $u : M_1 \to M_2$ is any uniformly continuous function with a modulus of uniform continuity $\delta(\epsilon)$ from any metric space to another, then the image of a minimal $\delta(\epsilon)$-net of M_1 under u becomes an ϵ-net for $u(M_1)$.

Proposition 6.4. Let $\epsilon > 0$ and suppose (M_1, d_1) and (M_2, d_2) are two metric spaces. If a function $u : M_1 \to M_2$ is uniformly continuous with a modulus of continuity $\delta(\epsilon)$, then $N(u(M_1), \epsilon, d_2) \leq N(M_1, \delta(\epsilon), d_1)$, where $u(M_1)$ denotes the image of u.

Proof. Suppose $N = N(M_1, \delta(\epsilon), d_1)$ is the $\delta(\epsilon)$-covering number for M_1 and let $\{m_1, \ldots, m_N\}$ be a $\delta(\epsilon)$-net for M_1. Hence for every $u(m) \in u(M_1)$, where $m \in M_1$, there exists $i \in \{1, \ldots, N\}$ such that

$$d_1(m, m_i) < \delta(\epsilon),$$

which implies $d_2(u(m), u(m_i)) < \epsilon$ as u is uniformly continuous. As a result, the set

$$\{u(m_1), \ldots, u(m_N)\}$$

is an ϵ-net for $u(M_1)$, so $N(u(M_1), \epsilon, d_2) \leq N(M_1, \delta(\epsilon), d_1)$. □

In particular, we can view $\mathcal{F}_1, \ldots, \mathcal{F}_k$ as metric spaces, all with distances induced by the $L_2(\mu)$ norm and suppose $\phi : \mathcal{F}_1 \times \ldots \times \mathcal{F}_k \to [0,1]^X$ is uniformly continuous with modulus of continuity $\delta(\epsilon, k)$. Then, by Proposition 6.3, if $\mathcal{F}_1, \ldots, \mathcal{F}_k$ all have finite $\frac{\delta(\epsilon,k)}{\sqrt{k}}$-covering numbers, the metric space $\mathcal{F}_1 \times \ldots \times \mathcal{F}_k$, with the L_2 product metric $\tilde{d}^2$, also has a finite $\delta(\epsilon, k)$-covering number: if we write $N(\mathcal{F}_i, \frac{\delta(\epsilon,k)}{\sqrt{k}}, L_2(\mu))$ as the $\frac{\delta(\epsilon,k)}{\sqrt{k}}$-covering number for $\mathcal{F}_i$, then,

$$N(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k, \delta(\epsilon, k), \tilde{d}^2) \leq \prod_{i=1}^{k} N(\mathcal{F}_i, \frac{\delta(\epsilon, k)}{\sqrt{k}}, L_2(\mu)).$$

Now, by Proposition 6.4,

$$\begin{aligned} N(\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k), \epsilon, L_2(\mu)) &\leq N(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k, \delta(\epsilon, k), \tilde{d}^2) \\ &\leq \prod_{i=1}^{k} N(\mathcal{F}_i, \frac{\delta(\epsilon, k)}{\sqrt{k}}, L_2(\mu)). \end{aligned}$$

In other words, the ϵ-covering number for $\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k)$ is bounded by a product of the $\frac{\delta(\epsilon,k)}{\sqrt{k}}$-covering numbers of each $\mathcal{F}_i$. To prove Lemma 6.2, we now state the main theorem of a paper written by Mendelson and Vershynin, which relates the ϵ-covering number of a function class to its Fat Shattering dimension of scale ϵ.

Theorem 6.5 ([MV03]). Let $\epsilon > 0$ and let $\mathcal{F}$ be a function class. Then for every probability measure μ,

$$N(\mathcal{F}, \epsilon, L_2(\mu)) \leq \left(\frac{2}{\epsilon}\right)^{K \mathrm{fat}_{c\epsilon}(\mathcal{F})}$$

for absolute constants c, K.

And Talagrand provides the converse.

Theorem 6.6 ([Tal03]). Following the notations of Theorem 6.5, there exists a probability measure μ such that

$$N(\mathcal{F}, \epsilon, L_2(\mu)) \geq 2^{K' \mathrm{fat}_{c'\epsilon}(\mathcal{F})},$$

for absolute constants c', K'.

Proof of Lemma 6.2. By Propositions 6.3 and 6.4,

$$N(\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k), \epsilon, L_2(\mu)) \leq \prod_{i=1}^{k} N(\mathcal{F}_i, \frac{\delta(\epsilon, k)}{\sqrt{k}}, L_2(\mu)),$$

so

$$\log(N(\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k), \epsilon, L_2(\mu))) \leq \sum_{i=1}^{k} \log(N(\mathcal{F}_i, \frac{\delta(\epsilon, k)}{\sqrt{k}}, L_2(\mu))).$$

By Theorem 6.5,

$$\log N(\mathcal{F}_i, \frac{\delta(\epsilon, k)}{\sqrt{k}}, L_2(\mu)) \leq K \mathrm{fat}_{c\frac{\delta(\epsilon,k)}{\sqrt{k}}}(\mathcal{F}_i) \log(2\sqrt{k}/\delta(\epsilon, k)),$$

for any probability measure μ where c, K are absolute constants. Moreover, by Theorem 6.6 for some probability measure μ and absolute constants c', K',

$$\log(N(\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k), \epsilon, L_2(\mu))) \geq K' \mathrm{fat}_{c'\epsilon}(\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k)) \log(2)$$

and altogether,

$$\begin{aligned} \mathrm{fat}_{c'\epsilon}(\phi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k)) &\leq \frac{\sum_{i=1}^{k} K \mathrm{fat}_{c\frac{\delta(\epsilon,k)}{\sqrt{k}}}(\mathcal{F}_i) \log(2\sqrt{k}/\delta(\epsilon, k))}{K' \log(2)} \\ &= \left(\frac{K \log(2\sqrt{k}/\delta(\epsilon, k))}{K' \log(2)}\right) \sum_{i=1}^{k} \mathrm{fat}_{c\frac{\delta(\epsilon,k)}{\sqrt{k}}}(\mathcal{F}_i). \end{aligned}$$

□

Finally, we will prove our main theorem.

Proof of Theorem 6.2. By Lemma 6.1, if $u : [0,1]^k \to [0,1]$ is uniformly continuous with modulus of continuity $\delta(\epsilon)$, then $\phi : \mathcal{F}_1 \times \ldots \times \mathcal{F}_k \to [0,1]^X$ defined by

$$\phi(f_1, \ldots, f_k)(x) = u(f_1(x), \ldots, f_k(x))$$

is also uniformly continuous with modulus of continuity $\frac{\delta(\epsilon/2)\epsilon}{2k}$. Then, apply Lemma 6.2 with $\delta(\epsilon, k) = \frac{\delta(\epsilon/2)\epsilon}{2k}$ and with a simple change of variables $c'\epsilon' \to \epsilon$, Theorem 6.2 follows directly. □

Altogether, we can summarize the maps in this section in the following two diagrams (where i is the diagonal map):

$$\begin{tikzcd} X \arrow[r, "i"] & X^k \arrow[r, "f_1 \times \ldots \times f_k"] & {[0,1]^k} \arrow[r, "u"] & {[0,1]} \end{tikzcd},$$

while

$$\begin{tikzcd} \mathcal{F}_1 \times \ldots \times \mathcal{F}_k \arrow[r, "\phi"] & {[0,1]^X} \end{tikzcd}.$$

This result is potentially useful because it allows us to construct new function classes using common continuous logic connectives and bound their Fat Shattering dimensions of scale ϵ. For instance, the function $u : [0,1]^2 \to [0,1]$ defined by $u(r_1, r_2) = r_1 \cdot r_2$ (multiplication) is uniformly continuous with a modulus of continuity $\delta(\epsilon) = \frac{\epsilon}{2}$. Indeed, let $\epsilon > 0$ and consider $(r_1, r_2), (r_1', r_2') \in [0,1]^2$. Suppose $d^2((r_1, r_2), (r_1', r_2')) < \delta(\epsilon) = \frac{\epsilon}{2}$, so

$$|r_1 - r_1'| < \sqrt{|r_1 - r_1'|^2 + |r_2 - r_2'|^2} < \frac{\epsilon}{2}$$

and similarly, $|r_2 - r_2'| < \frac{\epsilon}{2}$. Then,

$$\begin{aligned} |u(r_1, r_2) - u(r_1', r_2')| &= |r_1 r_2 - r_1' r_2'| \\ &= |r_1 r_2 - r_1 r_2' + r_1 r_2' - r_1' r_2'| \\ &\leq |r_1(r_2 - r_2')| + |r_2'(r_1 - r_1')| \\ &\leq |r_2 - r_2'| + |r_1 - r_1'| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

As a result, if $\mathcal{F}_1$ and $\mathcal{F}_2$ are two function classes with finite Fat Shattering dimensions of some scale ϵ, then the function class $u(\mathcal{F}_1, \mathcal{F}_2) = \mathcal{F}_1\mathcal{F}_2 = \{f_1 \cdot f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$, defined by point-wise multiplication, also has finite Fat Shattering dimension of scale ϵ, up to some constant factor, and Theorem 6.2 provides an upper bound.

We have made an interesting connection, which has not been explored much in the past, between continuous logic and PAC learning, and we plan to investigate this connection even further. For instance, the relationship of compositions of function classes and continuous logic may be interesting to study because compositions of uniformly continuous functions are again uniformly continuous. Furthermore, we can try to add some topological structures to concept or function classes to see how PAC learning can be affected.

7 Acknowledgements

I would like to thank Professor Vladimir Pestov from the University of Ottawa for all his help and support and for introducing me to learning theory. The work in this paper was partially funded by the NSERC Undergraduate Student Research Award in the fall of 2010. I would like to thank the reviewers of *The Waterloo Mathematics Review* for their thorough reading of the paper and their feedback.

This paper is based on the author's undergraduate honours project, supervised by Professor Vladimir Pestov and completed in April 2011 at the University of Ottawa.

References

[ABDCBH97] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, *Scale-sensitive dimensions, uniform convergence, and learnability*, Journal of the ACM **44** (1997), no. 4, 615–631.

[AC05] G. Auliac and J. Y. Caby, *Mathématiques: Topologie et analyse*, 3rd ed., EdiScience, Belgium, 2005.

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, *Learnability and the Vapnik-Chervonenkis Dimension*, Journal of the ACM **36** (1989), no. 4, 929 – 965.

[Bil95] P. Billingsley, *Probability and Measure*, 3rd ed., Wiley-Interscience, New York, 1995.

[Doo94] J. L. Doob, *Measure Theory*, Springer-Verlag, New York, 1994.

[KS94] M. J. Kearns and R. Schapire, *Efficient Distribution-free Learning of Probabilistic Concepts*, Journal of Computer System Sciences **48** (1994), no. 3, 464–497.

[KV94] M. J. Kearns and U. V. Vazirani, *An Introduction to Computational Learning Theory*, The MIT Press, Cambridge, Massachusetts, 1994.

[MV03] S. Mendelson and R. Vershynin, *Entropy and the Combinatorial Dimension*, Inventiones Mathematicae **152** (2003), 37 – 55.

[Pes10a] V. Pestov, *A Note on Sample Complexity of Learning Binary Output Neural Networks Under Fixed Input Distributions*, Proc. 2010 Eleventh Brazilian Symposium on Neural Networks, IEEE Computer Society, Los Alamitos-Washington-Tokyo (2010), 7 – 12.

[Pes10b] ———, *Indexability, Concentration, and VC Theory*, Proc. of the 3rd International Conf. on Similarity Search and Applications (SISAP 2010) (2010), 3 – 12.

[Sau72] N. Sauer, *On the Densities of Families of Sets*, J. Combinatorial Theory **13** (1972), 145 – 147.

[Tal03] M. Talagrand, *Vapnik-Chervonenkis Type Conditions and Uniform Donsker Classes of Functions*, Annals of Probability **31** (2003), no. 3, 1565 – 1582.

[Val84] L. G. Valiant, *A Theory of the Learnable*, Communications of the ACM **27** (1984), no. 11, 1134 – 1142.

[VC71] V. N. Vapnik and A. Y. Chervonenkis, *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*, Theory of Prob. and its Appl. **16** (1971), no. 2, 264 – 280.

[Vid97] M. Vidyasagar, *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*, Springer-Verlag London Limited, London, 1997.

[YBHU08] I. B. Yaacov, A. Berenstein, C. W. Henson, and A. Usvyatsov, *Model Theory for Metric Structures*, London Math Society Lecture Note Series **350** (2008), 315 – 427.

Integers with a predetermined prime factorization

Eric Naslund
University of British Columbia
naslund.eric@gmail.com

Abstract: A classic question in analytic number theory is to find asymptotics for $\sigma_k(x)$ and $\pi_k(x)$, the number of integers $n \leq x$ with exactly k prime factors, where $\pi_k(x)$ has the added constraint that all the factors are distinct. This problem was originally resolved by Landau in 1900, and much work was subsequently done where k is allowed to vary. In this paper we look at a similar question about integers with a specific prime factorization. Given $\boldsymbol{\alpha} \in \mathbb{N}^k$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ let $\sigma_{\boldsymbol{\alpha}}(x)$ denote the number of integers of the form $n = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$ where the p_i are not necessarily distinct, and let $\pi_{\boldsymbol{\alpha}}(x)$ denote the same counting function with the added condition that the factors are distinct. Our main result is asymptotics for both of these functions.

1 Introduction

One of the major problems in the 19th century was to find the growth rate of the number of primes less then x, that is the function

$$\pi(x) := \sum_{p \leq x} 1.$$

In 1797, Legendre conjectured that $\pi(x)$ is asymptotic to $\frac{x}{\log x}$, written as $\pi(x) \sim \frac{x}{\log x}$, which means that we have the limit

$$\lim_{x \to \infty} \frac{\pi(x)}{x/\log x} = 1.$$

Although a more precise conjecture was given by Gauss, little progress was made over the next 50 years. In 1848 and 1850, Chebyshev made several contributions, and managed to prove weaker upper and lower bounds. A major breakthrough occurred in 1859, when Riemann published his seminal paper, "On the Number of Primes Less Than a Given Magnitude," in which he outlined a proof of Legendre's conjecture using complex analysis and the zeta function. In 1896, 99 years after Legendre made his conjecture, Hadamard and de la Vallée Poussin rigorously completed Riemann's outline, proving what is known today as the prime number theorem [MV07]. In particular, we can write down the explicit error term :

$$\pi(x) = \frac{x}{\log x} + O\left(\frac{x}{\log^2 x}\right), \tag{1.1}$$

but to be more precise than this we would need to introduce the function from Gauss's conjecture.

A natural follow up question is whether or not we have similar asymptotics for the number of integers with exactly k prime factors. There are two reasonable ways to define the counting function; let $\sigma_k(x)$ denote the number of integers less then x with exactly k prime factors, and let $\pi_k(x)$ be the same but with the added constraint that the k prime factors must be distinct. For convenience, we also define the sets $\mathcal{P}_k^\sigma = \{n : \ n = p_1 \cdots p_k\}$ and $\mathcal{P}_k^\pi = \{n : \ n = p_1 \cdots p_k \text{ where } i \neq j \Rightarrow p_i \neq p_j\}$, so that we may write

$$\sigma_k(x) = \sum_{\substack{n \leq x \\ n \in \mathcal{P}_k^\sigma}} 1 \quad \text{and} \quad \pi_k(x) = \sum_{\substack{n \leq x \\ n \in \mathcal{P}_k^\pi}} 1.$$

In 1900 by Landau [Lan00] found the growth rate of these functions, and he proved that for fixed k we have

$$\pi_k(x) \sim \sigma_k(x) \sim \frac{x\left(\log\log x\right)^{k-1}}{(k-1)!\log x}. \tag{1.2}$$

E. M. Wright then gave a short elementary proof of this in 1954 [Wri54]. Heuristically we might expect this kind of asymptotic since $\sum_{k=1}^{\infty}\sigma_k(x) = \lfloor x\rfloor$, and if we could ignore the error term and sum over all $k \leq \log x$, we would arrive back at this equality again as

$$\sum_{k=1}^{\infty}\sigma_k(x) \approx \sum_{k=1}^{\infty}\frac{x\left(\log\log x\right)^{k-1}}{(k-1)!\log x} = \frac{x}{\log x}\sum_{k=0}^{\infty}\frac{\left(\log\log x\right)^{k}}{k!} = x.$$

Note that even though this works out, the heuristic is not entirely reliable. It seems to suggest that $\sigma_k(x) \sim \frac{x(\log\log x)^{k-1}}{(k-1)!\log x}$ even when k varies with x, which is not true when $k \approx \log\log x$ [HT88]. In his paper, Landau also gave explicit error terms, and showed that for $k \geq 2$

$$\sigma_k(x) = \frac{x\left(\log\log x\right)^{k-1}}{(k-1)!\log x} + O\left(\frac{x\left(\log\log x\right)^{k-2}}{\log x}\right) \tag{1.3}$$

and

$$\pi_k(x) = \frac{x\left(\log\log x\right)^{k-1}}{(k-1)!\log x} + O\left(\frac{x\left(\log\log x\right)^{k-2}}{\log x}\right) \tag{1.4}$$

where the notation $O(f(x))$ means that the error term is bounded in absolute value by some constant multiple of $f(x)$. (Although seperated on different lines, note that the above asymptotics are indeed the same.) In this paper we are interested in something very similar, which is counting the number of integers of a particular shape, integers of the form $p_1^{\alpha_1}\cdots p_n^{\alpha_n}$ where the α_i are fixed exponents. For example, we may ask how many integers of the form pq^3 are there less than x. To discuss this problem, we begin by introducing some notation. Given a vector $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_k) \in \mathbb{N}^k$, define $\sigma_{\boldsymbol{\alpha}}(x)$ to be the number of integers $n \leq x$ of the form $n = p_1^{\alpha_1}\cdots p_k^{\alpha_k}$, allowing prime repetitions, and $\pi_{\boldsymbol{\alpha}}(x)$ to be the number without prime repetitions. If we set $\mathcal{P}_{\boldsymbol{\alpha}}^{\sigma} = \{n : \ n = p_1^{\alpha_1}\cdots p_r^{\alpha_r}\}$, and $\mathcal{P}_{\boldsymbol{\alpha}}^{\pi} = \{n : \ n = p_1^{\alpha_1}\cdots p_r^{\alpha_r}$ where $i \neq j \Rightarrow p_i \neq p_j\}$, then as was done for $\pi_k(x)$, and $\sigma_k(x)$, we can rewrite these counting functions as

$$\sigma_{\boldsymbol{\alpha}}(x) = \sum_{\substack{n \leq x \\ n \in \mathcal{P}_{\boldsymbol{\alpha}}^{\sigma}}} 1 \quad \text{and} \quad \pi_{\boldsymbol{\alpha}}(x) = \sum_{\substack{n \leq x \\ n \in \mathcal{P}_{\boldsymbol{\alpha}}^{\pi}}} 1.$$

Our goal is to provide asymptotics for $\sigma_{\boldsymbol{\alpha}}(x)$ and $\pi_{\boldsymbol{\alpha}}(x)$, and our main theorem is:

Theorem 1.1. Let r, α be positive integers. Suppose we have a vector of the form $\boldsymbol{\alpha} = (\alpha, \cdots, \alpha, \alpha_1, \cdots, \alpha_r) \in \mathbb{N}^{k+r}$, where $k > 0$ is the multiplicity of α, and where $\alpha < \alpha_i$ for all i. Then if $\boldsymbol{\beta} = (\alpha_1, \cdots, \alpha_r) \in \mathbb{N}^r$, we have

$$\sigma_{\boldsymbol{\alpha}}(x) \sim \sigma_k\left(x^{\frac{1}{\alpha}}\right)\sum_{n\in\mathcal{P}_{\boldsymbol{\beta}}^{\sigma}} n^{-\frac{1}{\alpha}}$$

and

$$\pi_{\boldsymbol{\alpha}}(x) \sim \sigma_k\left(x^{\frac{1}{\alpha}}\right)\sum_{n\in\mathcal{P}_{\boldsymbol{\beta}}^{\pi}} n^{-\frac{1}{\alpha}}.$$

The above theorem tells us that the higher powers introduce a constant factor into the asymptotic since both of the series $\sum_{n\in\mathcal{P}_{\boldsymbol{\beta}}^{\sigma}} n^{-\frac{1}{\alpha}}$ and $\sum_{n\in\mathcal{P}_{\boldsymbol{\beta}}^{\pi}} n^{-\frac{1}{\alpha}}$ converge absolutely. The convergence of these series follows from the fact that $\frac{\alpha_i}{\alpha} > 1$ along with equation 2.1 in the next section. In particular, returning to our previous example of counting the number of integers of the form pq^3 less than x, we have that $\mathcal{P}^{\pi}_{\boldsymbol{\beta}} = \mathcal{P}^{\sigma}_{\boldsymbol{\beta}} = \{p^3 : \ p \text{ is prime}\}$, and hence

$$\pi_{(1,3)}(x) \sim \sigma_{(1,3)}(x) \sim \frac{x}{\log x} \sum_p \frac{1}{p^3} = \frac{x}{\log x} P(3)$$

where $P(s) = \sum_p p^{-s}$ is the prime zeta function. We can ask whether the constant can always be rewritten as a product of prime zeta functions, and this is answered by the following theorem:

Theorem 1.2. Suppose we are given $\alpha < \alpha_1 \leq \cdots \leq \alpha_r$, and that for any choice of $\epsilon_i \in \{-1, 0, 1\}$, we have $\sum_i \epsilon_i \alpha_i = 0$ implies $\epsilon_i = 0$ for every i. Then

$$\sum_{n \in \mathcal{P}^{\sigma}_{\boldsymbol{\beta}}} n^{-\frac{1}{\alpha}} = \prod_{i=1}^{r} P\left(\frac{\alpha_i}{\alpha}\right)$$

where $P(s) = \sum_p p^{-s}$ is the prime zeta function. This is equivalent to the condition that every $n \in \mathcal{P}^{\sigma}_{\boldsymbol{\beta}}$, where $\boldsymbol{\beta} = (\alpha_1, \ldots, \alpha_r)$, has a unique representation as $n = p_1^{\alpha_1} \cdots p_r^{\alpha_r}$.

For example, the above two theorems imply that the number of integers of the form $n = p_1 p_2 p_3^3 p_4^5 p_5^{19}$, with $n \leq x$, will be asymptotic to

$$\sigma_2\left(x\right) P(3)P(5)P(19) \sim \frac{x \log\log x}{\log x} P(3)P(5)P(19).$$

2 The Main Result

It is very important to split up the smallest power, as this is contributes the most to the sum. Throughout this section, we write our vector of exponents as $\boldsymbol{\alpha} = (\alpha, \cdots, \alpha, \alpha_1, \cdots, \alpha_r) \in \mathbb{N}^{k+r}$, with $1 \leq \alpha < \alpha_1 \leq \cdots \leq \alpha_r$, where $k > 0$ is the multiplicity of α, and let $\boldsymbol{\beta} = (\alpha_1, \cdots, \alpha_r) \in \mathbb{N}^r$. To start, we provide a simple upper bound for $\sigma_{\boldsymbol{\beta}}(x)$. Notice that

$$\pi_{\boldsymbol{\beta}}(x) \leq \sigma_{\boldsymbol{\beta}}(x) = \sum_{\substack{m \leq x \\ m \in \mathcal{P}^{\sigma}_{\boldsymbol{\beta}}}} 1 \leq \sum_{p_1^{\alpha_1} \cdots p_r^{\alpha_r} \leq x} 1,$$

where the right hand sum ranges over all vectors of primes of length r satisfying $p_1^{\alpha_1} \cdots p_r^{\alpha_r} \leq x$. Since $\alpha_1 \leq \alpha_i$ for all i, and $p_1^{\alpha_1} \cdots p_r^{\alpha_r} \leq x$ implies that $p_1^{\alpha_1} p_2^{\alpha_1} \cdots p_r^{\alpha_1} \leq x$, we see that replacing every exponent by α_1 only increases the sum. Then using 1.2 we have

$$\pi_{\boldsymbol{\beta}}(x) \leq \sigma_{\boldsymbol{\beta}}(x) \leq \sum_{p_1 \cdots p_r \leq x^{\frac{1}{\alpha_1}}} 1 = O\left(x^{\frac{1}{\alpha_1}} \frac{(\log\log x)^{r-1}}{\log x}\right). \tag{2.1}$$

The following subsection is devoted to examining $\sigma_{\boldsymbol{\alpha}}(x)$. The key will be using the hyperbola method, and most of the lemmas will apply identically to the proof for $\pi_{\boldsymbol{\alpha}}(x)$.

2.1 $\sigma_{\boldsymbol{\alpha}}(x)$

Each integer $n \in \mathcal{P}^{\sigma}_{\boldsymbol{\alpha}}$ has one part in $\mathcal{P}^{\sigma}_{k}$, and one part in $\mathcal{P}^{\sigma}_{\boldsymbol{\beta}}$, and our goal will be to split it up between these two to better understand $\sigma_{\boldsymbol{\alpha}}(x)$. With this in mind, we might expect

$$\sigma_{\boldsymbol{\alpha}}(x) \approx \sum_{\substack{mn^{\alpha} \leq x \\ n \in \mathcal{P}^{\sigma}_{k}, m \in \mathcal{P}^{\sigma}_{\boldsymbol{\beta}}}} 1.$$

However, this will not be an exact equality as an integer $k \le x$ with $k \in \mathcal{P}_{\boldsymbol{\alpha}}^{\sigma}$ may have more than one representation of the form $k = mn^{\alpha}$ with $n \in \mathcal{P}_k^{\sigma},\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}$. Since $k \in \mathcal{P}_{\boldsymbol{\alpha}}^{\sigma}$ can have at most one representation of the form $k = mn^{\alpha}$ with $n \in \mathcal{P}_k^{\pi},\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}$, we have the inequalities

$$\sum_{\substack{mn^{\alpha} \le x \\ n \in \mathcal{P}_k^{\pi}, m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} 1 \le \sigma_{\boldsymbol{\alpha}}(x) \le \sum_{\substack{mn^{\alpha} \le x \\ n \in \mathcal{P}_k^{\sigma}, m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} 1.$$

Rewriting so that we first sum over m, this is

$$\sum_{\substack{mn^{\alpha} \le x \\ n \in \mathcal{P}_k^{\sigma}, m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} 1 = \sum_{\substack{m \le x \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} \sum_{\substack{n^{\alpha} \le \frac{x}{m} \\ n \in \mathcal{P}_k^{\sigma}}} 1 = \sum_{\substack{m \le x \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} \sigma_k \left(\left(\frac{x}{m} \right)^{\frac{1}{\alpha}} \right)$$

and we have that

$$\sum_{\substack{m \le x \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} \pi_k \left(\left(\frac{x}{m} \right)^{\frac{1}{\alpha}} \right) \le \sigma_{\boldsymbol{\alpha}}(x) \le \sum_{\substack{m \le x \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} \sigma_k \left(\left(\frac{x}{m} \right)^{\frac{1}{\alpha}} \right). \tag{2.2}$$

Our first goal will be to remove all of the terms from the sum with $m \ge \frac{x}{(\log x)^C}$ for some constant $C > 2$, without introducing large error. For example, we could take $C = 3$ to prove the asymptotic. However to achieve the optimal error term we need something of the form $C = 2\alpha\alpha_1 + 1$, a choice which will become clear later on. Note that we need only bound this sum for $\sigma_k(x)$, since $\pi_k(x) \le \sigma_k(x)$, and this is covered by the following lemma.

Lemma 2.1. For $C > 1$ we have that

$$\sum_{\substack{(\log x)^C < m \le x \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} \sigma_k \left(\left(\frac{x}{m} \right)^{\frac{1}{\alpha}} \right) = O \left(\frac{x^{\frac{1}{\alpha}}}{(\log x)^{(C-1)\left(1 - \frac{\alpha}{\alpha_1}\right)}} \right).$$

Proof. We may change the order of summation and write

$$\begin{aligned} \sum_{\substack{(\log x)^C \le m \le x \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} \sigma_k \left(\left(\frac{x}{m} \right)^{\frac{1}{\alpha}} \right) &= \sum_{\substack{(\log x)^C \le m \le x \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} \sum_{\substack{n^{\alpha} \le \frac{x}{m} \\ n \in \mathcal{P}_k^{\sigma}}} 1 \\ &= \sum_{\substack{n^{\alpha} \le \frac{x}{(\log x)^C} \\ n \in \mathcal{P}_k^{\sigma}}} \sum_{\substack{(\log x)^C \le m \le \frac{x}{n^{\alpha}} \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} 1. \end{aligned}$$

Using 2.1 this is bounded above by

$$\begin{aligned} \sum_{\substack{n^{\alpha} \le \frac{x}{(\log x)^C} \\ n \in \mathcal{P}_k^{\sigma}}} \sum_{\substack{m \le \frac{x}{n^{\alpha}} \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} 1 &= \sum_{\substack{n^{\alpha} \le \frac{x}{(\log x)^C} \\ n \in \mathcal{P}_k^{\sigma}}} O \left(\frac{x^{\frac{1}{\alpha_1}}}{n^{\frac{\alpha}{\alpha_1}}} \frac{(\log\log (x/n^{\alpha}))^{r-1}}{\log (x/n^{\alpha})} \right) \\ &= O \left(x^{\frac{1}{\alpha_1}} (\log\log x)^{r-1} \sum_{\substack{n^{\alpha} \le \frac{x}{(\log x)^C} \\ n \in \mathcal{P}_k^{\sigma}}} \frac{1}{n^{\frac{\alpha}{\alpha_1}}} \right). \end{aligned}$$

Taking the trivial bound, the inner sum becomes

$$\begin{aligned}\sum_{\substack{n^{\alpha} \leq \frac{x}{(\log x)^C} \\ n \in \mathcal{P}_k^{\sigma}}} \frac{1}{n^{\frac{\alpha}{\alpha_1}}} &\leq \sum_{n \leq \frac{x^{\frac{1}{\alpha}}}{\log^{\frac{C}{\alpha}} x}} \frac{1}{n^{\frac{\alpha}{\alpha_1}}} = O\left(\left(\frac{x^{\frac{1}{\alpha}}}{\log^C x}\right)^{-\frac{\alpha}{\alpha_1}-1}\right) \\ &= O\left(\frac{1}{(\log x)^{C\left(1-\frac{\alpha}{\alpha_1}\right)}}\right),\end{aligned}$$

so that we have the upper bound

$$O\left(\frac{x^{\frac{1}{\alpha}}}{(\log x)^{(C-1)\left(1-\frac{\alpha}{\alpha_1}\right)}} \frac{(\log\log x)^{r-1}}{(\log x)^{1-\frac{\alpha}{\alpha_1}}}\right) = O\left(\frac{x^{\frac{1}{\alpha}}}{(\log x)^{(C-1)\left(1-\frac{\alpha}{\alpha_1}\right)}}\right).$$

□

Combining 2.2 along with Lemma 2.1 and Landau's estimates 1.3, 1.4 for $k > 1$ yields

$$\begin{aligned}\sigma_{\boldsymbol{\alpha}}(x) = &\frac{1}{(k-1)!} \sum_{\substack{m \leq (\log x)^C \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} \alpha \frac{x^{\frac{1}{\alpha}} \left(\log\left(\frac{1}{\alpha}\log\left(\frac{x}{m}\right)\right)\right)^{k-1}}{m^{\frac{1}{\alpha}} \log\left(\frac{x}{m}\right)} \\ &+ O\left(\frac{x^{\frac{1}{\alpha}}}{(\log x)^{(C-1)\left(1-\frac{\alpha}{\alpha_1}\right)}} + \sum_{\substack{m \leq (\log x)^C \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} \frac{x^{\frac{1}{\alpha}} \left(\log\left(\frac{1}{\alpha}\log\left(\frac{x}{m}\right)\right)\right)^{k-2}}{m^{\frac{1}{\alpha}} \log\left(\frac{x}{m}\right)}\right),\end{aligned} \tag{2.3}$$

and for $k = 1$ by 1.1, the prime number theorem, we have

$$\begin{aligned}\sigma_{\boldsymbol{\alpha}}(x) = &\sum_{\substack{m \leq (\log x)^C \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} \alpha \frac{x^{\frac{1}{\alpha}}}{m^{\frac{1}{\alpha}} \log\left(\frac{x}{m}\right)} \\ &+ O\left(\frac{x^{\frac{1}{\alpha}}}{(\log x)^{(C-1)\left(1-\frac{\alpha}{\alpha_1}\right)}} + \sum_{\substack{m \leq (\log x)^C \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} \frac{x^{\frac{1}{\alpha}}}{m^{\frac{1}{\alpha}} \log^2\left(\frac{x}{m}\right)}\right).\end{aligned} \tag{2.4}$$

If we write $\left(\log\left(\frac{1}{\alpha}\log\left(\frac{x}{m}\right)\right)\right)^{k-1} = \left(\log\log\left(\frac{x}{m}\right) - \log\alpha\right)^{k-1}$ and then expand using the binomial theorem, all of the terms will be consumed by the error term except for the one with $\left(\log\log\left(\frac{x}{m}\right)\right)^{k-1}$, which allows us to change the main term in the above to

$$\frac{1}{(k-1)!} \sum_{\substack{m \leq (\log x)^C \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} \alpha \frac{x^{\frac{1}{\alpha}} \left(\log\log\left(\frac{x}{m}\right)\right)^{k-1}}{m^{\frac{1}{\alpha}} \log\left(\frac{x}{m}\right)}. \tag{2.5}$$

We may clean up the error terms by bounding each part of the sum from above. Since $m \leq (\log x)^C$, $\frac{1}{\log\left(\frac{x}{m}\right)}$ is bounded above by

$$\frac{1}{\log\left(\frac{x}{(\log x)^C}\right)} = \frac{1}{\log(x) - C\log\log x} = \frac{1}{\log x} + O\left(\frac{\log\log x}{\log^2 x}\right).$$

We also have the trivial bounds

$$\log\left(\frac{1}{\alpha}\log\left(\frac{x}{m}\right)\right) \leq \left(\log\left(\log(x)\right)\right),$$

and

$$\sum_{\substack{m \leq (\log x)^C \\ m \in \mathcal{P}^\sigma_{\boldsymbol{\beta}}}} \frac{1}{m^{\frac{1}{\alpha}}} \leq \sum_{m \in \mathcal{P}^\sigma_{\boldsymbol{\beta}}} m^{-\frac{1}{\alpha}}$$

since the right hand side is a convergent series. Combining these, for integers $A \geq 0$, $B > 1$ we have that

$$\sum_{\substack{m \leq (\log x)^C \\ m \in \mathcal{P}^\sigma_{\boldsymbol{\beta}}}} \frac{x^{\frac{1}{\alpha}}\left(\log\left(\frac{1}{\alpha}\log\left(\frac{x}{m}\right)\right)\right)^A}{m^{\frac{1}{\alpha}}\log^B\left(\frac{x}{M}\right)} = O\left(\frac{x^{\frac{1}{\alpha}}\left(\log\log x\right)^A}{\log^B(x)}\right), \tag{2.6}$$

which gives an upper bound on the error term in both cases, $k = 1$ and $k > 1$. The following lemma allows us to deal with the main term:

Lemma 2.2. For $C > 1$, we have that

$$\sum_{\substack{m \leq (\log x)^C \\ m \in \mathcal{P}^\sigma_{\boldsymbol{\beta}}}} \frac{\left(\log\log\left(\frac{x}{m}\right)\right)^{k-1}}{m^{\frac{1}{\alpha}}\log\left(\frac{x}{m}\right)} = \frac{\left(\log\log(x)\right)^{k-1}}{\log x} \sum_{\substack{m \leq (\log x)^C \\ m \in \mathcal{P}^\sigma_{\boldsymbol{\beta}}}} m^{-\frac{1}{\alpha}} + O\left(\frac{\left(\log\log x\right)^{k-1}}{\log^2 x}\right).$$

Proof. First, note that we have the bounds

$$\frac{1}{\log(x)} \leq \frac{1}{\log\left(\frac{x}{m}\right)} \leq \frac{1}{\log\left(\frac{x}{\log x}\right)}$$

and

$$\left(\log\log\left(\frac{x}{\log x}\right)\right)^{k-1} \leq \left(\log\log\left(\frac{x}{m}\right)\right)^{k-1} \leq \left(\log\log(x)\right)^{k-1}.$$

Using power series expansions we may write

$$\frac{1}{\log\left(\frac{x}{\log x}\right)} = \frac{1}{\log(x)\left(1 - \frac{\log\log x}{\log x}\right)} = \frac{1}{\log x} + O\left(\frac{\log\log x}{\log^2 x}\right)$$

and

$$\left(\log\log\left(\frac{x}{\log x}\right)\right)^{k-1} = \left(\log\log x + \log\left(1 - \frac{\log\log x}{\log x}\right)\right)^{k-1} = \left(\log\log x\right)^{k-1} + O\left(\frac{\left(\log\log x\right)^{k-1}}{\log x}\right).$$

Then 2.6 implies that

$$\sum_{\substack{m \leq (\log x)^C \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} \frac{\left(\log\log\left(\frac{x}{m}\right)\right)^{k-1}}{m^{\frac{1}{\alpha}}\log\left(\frac{x}{m}\right)} = \frac{(\log\log x)^{k-1}}{\log x} \sum_{\substack{m \leq (\log x)^C \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} m^{-\frac{1}{\alpha}} + O\left(\frac{(\log\log x)^{k-1}}{\log^2 x}\right).$$

□

Let $C = 2\alpha\alpha_1 + 1$ so that $(C-1)\left(1 - \frac{\alpha}{\alpha_1}\right) = 2\alpha\left(\alpha_1 - \alpha\right) \geq 2$. Upon combining 2.3, 2.5, 2.6, and lemma 2.2 for $k > 1$ we obtain

$$\sigma_{\boldsymbol{\alpha}}(x) = \alpha \frac{x^{\frac{1}{\alpha}}(\log\log x)^{k-1}}{(k-1)!\log x} \sum_{\substack{m \leq (\log x)^C \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} m^{-\frac{1}{\alpha}} + O\left(x^{\frac{1}{\alpha}}\frac{(\log\log x)^{k-2}}{\log x}\right). \tag{2.7}$$

Similarly, 2.4, 2.6, and lemma 2.2 together yield

$$\sigma_{\boldsymbol{\alpha}}(x) = \alpha \frac{x^{\frac{1}{\alpha}}}{\log x} \sum_{\substack{m \leq (\log x)^C \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} m^{-\frac{1}{\alpha}} + O\left(\frac{x^{\frac{1}{\alpha}}}{\log^2 x}\right)$$

for $k = 1$. To deal with the last sum, write

$$\sum_{\substack{m \leq (\log x)^C \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} m^{-\frac{1}{\alpha}} = \sum_{m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}} m^{-\frac{1}{\alpha}} - \sum_{\substack{m > (\log x)^C \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} m^{-\frac{1}{\alpha}}.$$

Applying summation by parts, we have that

$$\begin{aligned} \sum_{\substack{m > (\log x)^C \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}}} m^{-\frac{1}{\alpha}} &= \int_{(\log x)^C}^{\infty} t^{-\frac{1}{\alpha}} d\left(\sigma_{\boldsymbol{\beta}}(t)\right) \\ &= t^{-\frac{1}{\alpha}}\sigma_{\boldsymbol{\beta}}(t)\Big|_{(\log x)^C}^{\infty} + \frac{1}{\alpha}\int_{(\log x)^C}^{\infty} t^{-\frac{1}{\alpha}-1}\sigma_{\boldsymbol{\beta}}(t)dt. \end{aligned}$$

Then by 2.1 this becomes

$$O\left((\log x)^{C\left(\frac{1}{\alpha_1}-\frac{1}{\alpha}\right)}(\log\log\log x)^{r-1}\right) = O\left(\frac{1}{(\log x)^2}\right)$$

since $C\left(\frac{1}{\alpha_1} - \frac{1}{\alpha}\right) = -2 + \left(\frac{1}{\alpha_1} - \frac{1}{\alpha}\right)$. Thus for $k > 1$ we have

$$\sigma_{\boldsymbol{\alpha}}(x) = \alpha \frac{x^{\frac{1}{\alpha}}(\log\log x)^{k-1}}{(k-1)!\log x} \sum_{m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}} m^{-\frac{1}{\alpha}} + O\left(x^{\frac{1}{\alpha}}\frac{(\log\log x)^{k-2}}{\log x}\right), \tag{2.8}$$

and for $k = 1$,

$$\sigma_{\boldsymbol{\alpha}}(x) = \alpha \frac{x^{\frac{1}{\alpha}}}{\log x} \sum_{m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}} m^{-\frac{1}{\alpha}} + O\left(\frac{x^{\frac{1}{\alpha}}}{(\log x)^2}\right). \tag{2.9}$$

This yields the desired asymptotic

$$\sigma_{\boldsymbol{\alpha}}(x) \sim \alpha \frac{x^{\frac{1}{\alpha}} (\log\log x)^{k-1}}{(k-1)! \log x} \sum_{m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}} m^{-\frac{1}{\alpha}}, \tag{2.10}$$

and since

$$\sigma_k\left(x^{\frac{1}{\alpha}}\right) \sim \alpha \frac{x^{\frac{1}{\alpha}} (\log\log x)^{k-1}}{(k-1)! \log x}$$

by Landau's estimates 1.2, we conclude that

$$\sigma_{\boldsymbol{\alpha}}(x) \sim \sigma_k\left(x^{\frac{1}{\alpha}}\right) \sum_{m \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}} m^{-\frac{1}{\alpha}}, \tag{2.11}$$

proving the first part of Theorem 1.1.

2.2 $\pi_{\boldsymbol{\alpha}}(x)$

To prove the same result for $\pi_{\boldsymbol{\alpha}}(x)$, we start again by splitting integers $n \in \mathcal{P}_{\boldsymbol{\alpha}}^{\sigma}$ into two parts, one in $\mathcal{P}_k^{\pi}$, and one in $\mathcal{P}_{\boldsymbol{\beta}}^{\pi}$. With this in mind we consider

$$\sum_{\substack{n^{\alpha} m \leq x \\ n \in \mathcal{P}_k^{\pi}, m \in \mathcal{P}_{\boldsymbol{\beta}}^{\pi}}} 1.$$

This will be strictly larger then $\pi_{\boldsymbol{\alpha}}(x)$ since n and m may have prime factors in common. (Note that since all factors are distinct, we cannot have multiple representations $k = mn$.) However, we can throw out all of the terms for which $\gcd(m,n) > 1$ without affecting the asymptotic. Write $n = q_1 \cdots q_k$, and $m = p_1^{\alpha_1} \cdots p_r^{\alpha_r}$. If $\gcd(m,n) > 1$, then we must have $q_i = p_j$ for some i, j. The set of all tuples with $q_i = p_j$ is bounded above by

$$\sigma_{\boldsymbol{\alpha}_{i,j}}(x)$$

where $\boldsymbol{\alpha}_{i,j} = (\alpha, \ldots, \alpha, \alpha_1, \cdots, (\alpha_j + \alpha), \cdots, \alpha_r) \in \mathbb{N}^{k-1+r}$ and we have $k-1$ copies of α. In particular, by 2.10, we see that

$$\sigma_{\boldsymbol{\alpha}_{i,j}}(x) = O\left(x^{\frac{1}{\alpha}} \frac{(\log\log x)^{k-2}}{\log x}\right)$$

for $k > 1$, and

$$\sigma_{\boldsymbol{\alpha}_{i,j}}(x) = O_{\epsilon}\left(x^{\frac{1}{\alpha_1}+\epsilon}\right)$$

for any $\epsilon > 0$ when $k = 1$. Since there are at most $k \cdot r$ possible pairs (i,j), it follows that for $k > 1$

$$\pi_{\boldsymbol{\alpha}}(x) = \sum_{\substack{n^{\alpha} m \leq x \\ n \in \mathcal{P}_k^{\pi}, m \in \mathcal{P}_{\boldsymbol{\beta}}^{\pi}}} 1 + O\left(x^{\frac{1}{\alpha}} \frac{(\log\log x)^{k-2}}{\log x}\right),$$

and a similar error term as before when $k = 1$. The main term may be rewritten as

$$\sum_{\substack{m \leq x \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\pi}}} \sum_{\substack{n^{\alpha} \leq \frac{x}{m} \\ n \in \mathcal{P}_k^{\pi}}} 1 = \sum_{\substack{m \leq x \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\pi}}} \pi_k\left(\left(\frac{x}{m}\right)^{\frac{1}{\alpha}}\right),$$

and from here, following through the exact same sequence of steps and lemmas from the previous section will yield

$$\sum_{\substack{m \leq x \\ m \in \mathcal{P}_{\boldsymbol{\beta}}^{\pi}}} \pi_k \left(\left(\frac{x}{m} \right)^{\frac{1}{\alpha}} \right) \sim \alpha \frac{x^{\frac{1}{\alpha}} \left(\log \log x \right)^{k-1}}{(k-1)! \log x} \sum_{m \in \mathcal{P}_{\boldsymbol{\beta}}^{\pi}} m^{-\frac{1}{\alpha}}.$$

All of the upper bounds for $\sigma_{\boldsymbol{\alpha}}(x)$ still apply to $\pi_{\boldsymbol{\alpha}}(x)$, and the only change is that we are summing over $\mathcal{P}_{\boldsymbol{\beta}}^{\pi}$ rather then $\mathcal{P}_{\boldsymbol{\beta}}^{\sigma}$, which is why the final sum is different. Using 1.2, we get that

$$\pi_{\boldsymbol{\alpha}}(x) \sim \pi_k \left(x^{\frac{1}{\alpha}} \right) \sum_{m \in \mathcal{P}_{\boldsymbol{\beta}}^{\pi}} m^{-\frac{1}{\alpha}}, \tag{2.12}$$

proving the second part of Theorem 1.1. If the error term is kept throughout the above computations, we get the more precise

$$\pi_{\boldsymbol{\alpha}}(x) = \alpha \frac{x^{\frac{1}{\alpha}} \left(\log \log x \right)^{k-1}}{(k-1)! \log x} \sum_{m \in \mathcal{P}_{\boldsymbol{\beta}}^{\pi}} m^{-\frac{1}{\alpha}} + O \left(x^{\frac{1}{\alpha}} \frac{\left(\log \log x \right)^{k-2}}{\log x} \right) \tag{2.13}$$

when $k > 1$, and

$$\pi_{\boldsymbol{\alpha}}(x) = \alpha \frac{x^{\frac{1}{\alpha}} \left(\log \log x \right)^{k-1}}{(k-1)! \log x} \sum_{m \in \mathcal{P}_{\boldsymbol{\beta}}^{\pi}} m^{-\frac{1}{\alpha}} + O \left(\frac{x^{\frac{1}{\alpha}}}{\log^2 x} \right), \tag{2.14}$$

for $k = 1$.

3 The Constant Factor

Let $\alpha > 0$ be given, let $A = \{\alpha_1, \cdots, \alpha_r\}$ where $\alpha < \alpha_i \leq \alpha_j$ for all i, j, and set set $\boldsymbol{\beta} = (\alpha_1, \ldots, \alpha_r)$. If every $n \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}$ has one and only one representation of the form $n = p_1^{\alpha_1} \cdots p_r^{\alpha_r}$, then we may decompose the sum as

$$\sum_{n \in \mathcal{P}_{\boldsymbol{\beta}}^{\sigma}} n^{-\frac{1}{\alpha}} = \sum_{p_1} \sum_{p_2} \cdots \sum_{p_r} \left(p_1^{\alpha_1} \cdots p_r^{\alpha_r} \right)^{-\frac{1}{\alpha}}.$$

This equals

$$\left(\sum_{p_1} p_1^{-\frac{\alpha_1}{\alpha}} \right) \cdots \left(\sum_{p_r} p_r^{-\frac{\alpha_r}{\alpha}} \right)$$

which by definition of the prime zeta function, $P(s) = \sum_p p^{-s}$, is

$$\prod_{i=1}^{r} P \left(\frac{\alpha_i}{\alpha} \right).$$

We now show that each integer can be uniquely represented if and only if $\sum_i \epsilon_i \alpha_i = 0$ with $\epsilon_i \in \{-1, 0, 1\}$ implies that every $\epsilon_i = 0$. Suppose we are given ϵ_i, not all zero, with $\sum_i \epsilon_i \alpha_i = 0$. Then we have then we have $\alpha_{i_1} + \alpha_{i_2} + \cdots + \alpha_{i_k} = \alpha_{j_1} + \alpha_{j_2} + \cdots + \alpha_{j_l} = M$ for some M where each all of the i_n and j_m are distinct. Setting $p_{i_1} = \cdots = p_{i_k} = p$, and $p_{j_1} = \cdots = p_{j_l} = q$, we will have a factor of $q^M p^M$, and this allows us to permute q and p giving two representations of the same integer. Conversely, if we have two representations of the same integer, then it must be because of a factor of the form $q^M p^M$, which implies that we must have $\sum_i \epsilon_i \alpha_i = 0$ for some non zero choices ϵ_i. This then completes the proof of Theorem 1.2.

References

[HT88] Adolf Hildebrand and Gérald Tenenbaum, *On the number of prime factors of an integer*, Duke Math. J. **56** (1988), no. 3, 471–501. MR 948530 (89k:11084)

[Lan00] E. Landau, *Sur quelques problèmes relatifs à la distribution des nombres premiers*, Bull. Soc. Math. France **28** (1900), 25–38. MR 1504359

[MV07] Hugh L. Montgomery and Robert C. Vaughan, *Multiplicative number theory. I. Classical theory*, Cambridge Studies in Advanced Mathematics, vol. 97, Cambridge University Press, Cambridge, 2007. MR 2378655 (2009b:11001)

[Wri54] E. M. Wright, *A simple proof of a theorem of Landau*, Proc. Edinburgh Math. Soc. (2) **9** (1954), 87–90. MR 0065579 (16,448e)

Rotterdam Must Die: Triangular Finite Volume Methods Applied to the Shallow Water Equations

Luke Bovard and Katharine Hyatt
University of Waterloo

Introduction

In this paper we apply the method of finite volumes using a triangular mesh with a Roe solver to solve the shallow water wave equations. In order to demonstrate the advantages of using a triangular mesh, we solve two problems that are not easily solved using rectangular finite volume methods. We first solve the classic problem of a broken circular dam and then apply the scheme to the Maeslantkering, a movable barrier along the Nieuwe Waterweg in Holland used to regulate water flow from storms into the shipping canal, to demonstrate the complicated geometry that triangular meshes are able to model.

Background

Failure of Finite Difference Schemes

The simplest scheme avaiable to solve PDEs numerically is to use a finite difference formula. The ideas behind finite difference formula is to use simple approximations to the derivatives and numerically solve the resulting system of equations. However, finite difference schemes are fairly limited in their scope. For example, consider the well known one dimensional PDE Burger's equation [Ach90]

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0$$

the exact solution of this equation is well known and a prototypical example in the method of characteristics. However, suppose we try and apply a finite difference scheme to the above. Discretising both derivatives we have that

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{u_j^n}{\Delta x}(u_j^n - u_{j-1}^n) = 0$$

where u_j^n is the approximation at position j and time n. Suppose that we subject the above system to the initial condition

$$u_0(x) = \begin{cases} 1 & x \leq 0 \\ 0 & x > 0 \end{cases}$$

Let us re-write the above scheme explicitly

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}u_j^n(u_j^n - u_{j-1}^n)$$

Consider the first time-step. For $x_j > 0$ we would have that

$$\begin{aligned} u_j^1 &= u_j^0 - \frac{\Delta t}{\Delta x}u_j^0(u_j^0 - u_{j-1}^0) \\ &= 0 - \frac{\Delta t}{\Delta x}0(0 - u_{j-1}^0) \\ &= 0 \end{aligned}$$

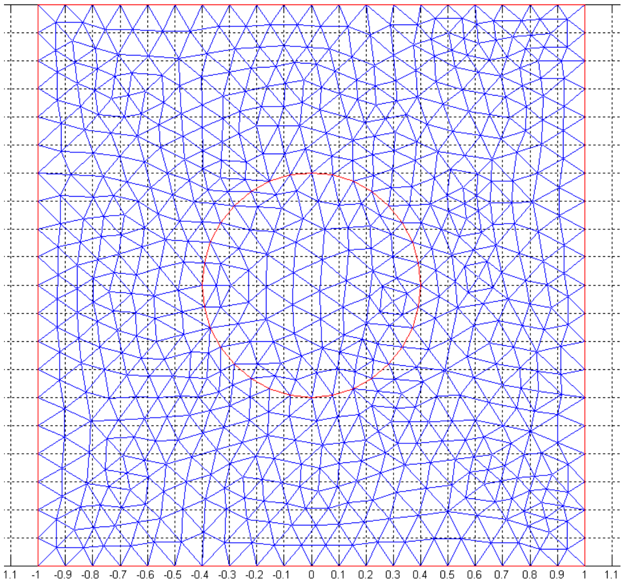

Figure 0.1: The triangular mesh used for the circular dam problem

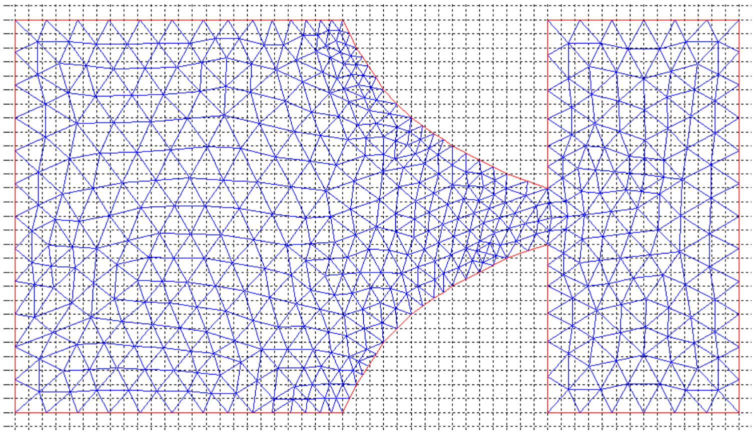

Figure 0.2: The triangular mesh used for the Maeslantkering

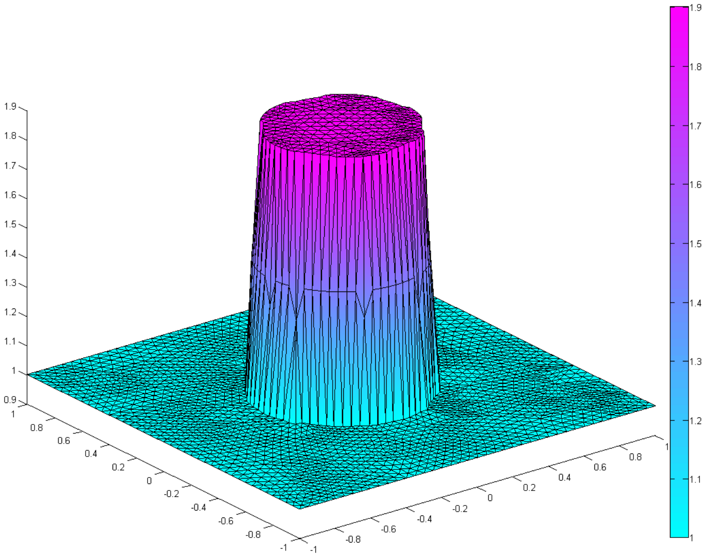

Figure 0.3: Initial condition of the circular dam problem

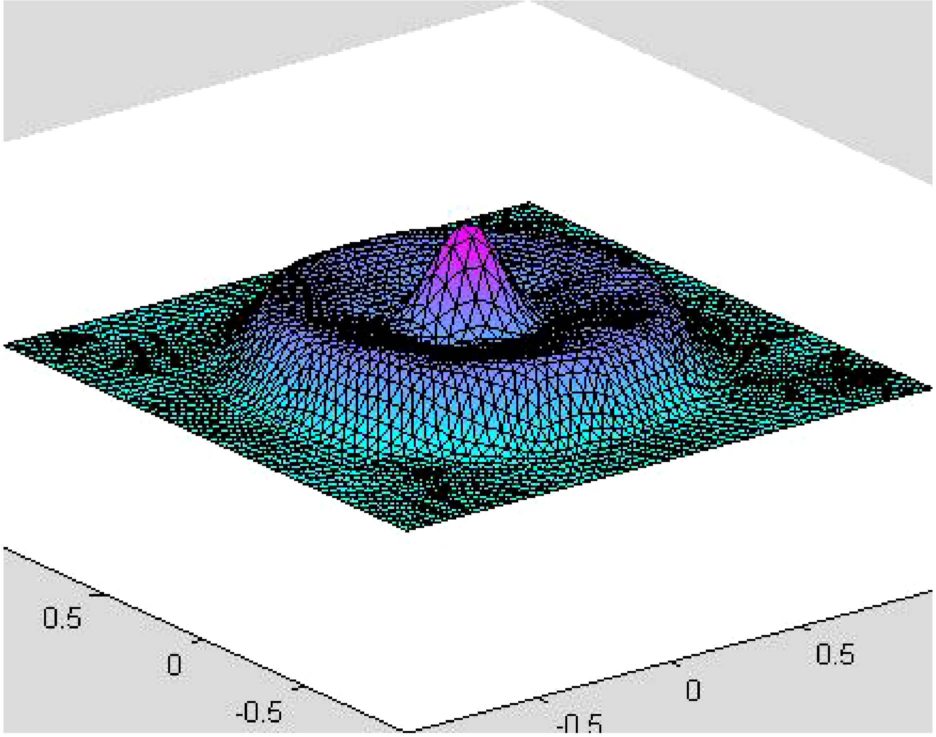

Figure 0.4: The circular dam at 100 timesteps

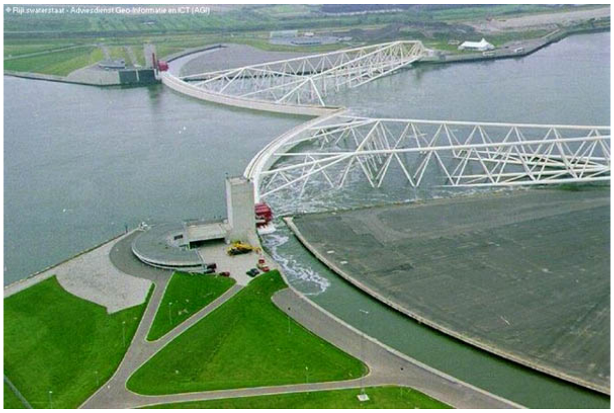
© Rijkswaterstaat - Adviesdienst Geo-Informatie en ICT (AGI)

Figure 0.5: The Maeslantkering fully closed along the Nieuwe Waterweg. (Source Rijkswaterstaat)

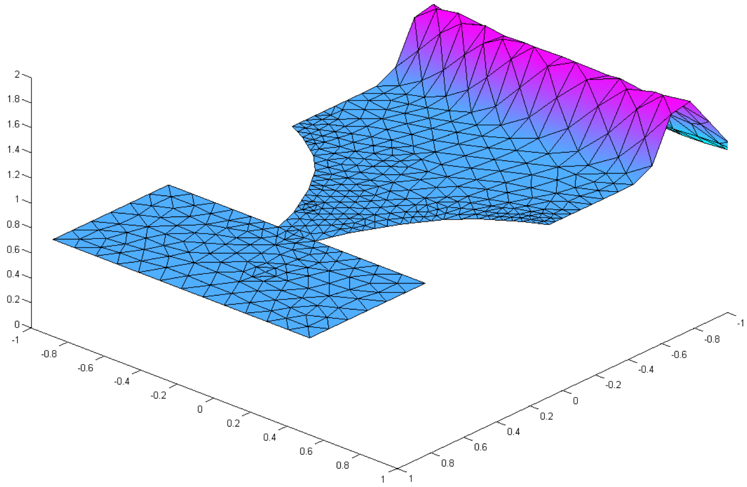

Figure 0.6: The initial wave sent in from the sea towards the Maeslantkering

For $x_j \leq 0$ we have that

$$\begin{aligned}
u_j^1 &= u_j^0 - \frac{\Delta t}{\Delta x} u_j^0 (u_j^0 - u_{j-1}^0) \\
&= 1 - \frac{\Delta t}{\Delta x}(1-1) \\
&= 1
\end{aligned}$$

Thus after one iteration, the scheme has not changed and the profile will always stay the same. Clearly, as can be verified by the method of characteristics, there is evolution, and for this specific example, shock waves will form. Thus the limitation of finite difference schemes requires a different scheme is apparent. In this paper we consider equations very similar to Burger's equation. We note that we can re-write the Burger's equation in the following form

$$\frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial u^2}{\partial x} = 0$$

which is referred to as a conservation law as it can be interpretted as the conservation of the solution.

Shallow Water Waves

The shallow water equations describe the motion of incompressible fluids in situations where the vertical depth of the system is much smaller than the relevant horizontal length scale. Although simplified, the shallow water equations are very powerful and can accurately model the motion of water in a puddle to the entire ocean. The starting point of the derivation is given by the Navier-Stokes equations

$$\rho\left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{v}\right) = -\nabla p + \nu\nabla^2\mathbf{v} + \mathbf{f}$$

Since the Navier-Stokes equations are a very complicated non-linear set of partial differential equations, we make a few simplifying assumptions. Firstly we assume inviscid flow ($\nu = 0$), a reasonable assumption far away from the boundary. However, since we are modelling a physical system, in actuality there will be boundary layer effects since water is not actually inviscid, but these effects are small and negligible in this approximation. Additionally, any turbulent effects are neglected. We also assume that the geometry is only two dimensional since vertical length scale H, is much less than the horizontal scale L, giving $H \ll L$ (which is valid in the case of a long canal since the depth is about 10-15 m while the length is on the order of hundreds of meters), and thus vertical velocity of the fluid can be neglected since the horizontal velocities will dominate the dynamics (see [Ray] for an elementary derivation). A more rigorous argument can be made by appealing to the orders of magnitude of the various terms in the Navier-Stokes equations [Ach90] [KC04] in which the same result is shown. This approximation forms the basis of the shallow water wave which can be in conservative form with $h(x,y,t)$, the height of the water and $u(x,y,t), v(x,y,t)$ the horizontal components of velocity

$$\begin{aligned}
\frac{\partial h}{\partial t} + \frac{\partial (hu)}{\partial x} + \frac{\partial (hv)}{\partial y} &= 0 \\
\frac{\partial (hu)}{\partial t} + \frac{\partial}{\partial x}(hu^2 + gh^2/2) + \frac{\partial (huv)}{\partial y} &= 0 \qquad (0.1)\\
\frac{\partial (hv)}{\partial t} + \frac{\partial (huv)}{\partial x} + \frac{\partial}{\partial y}(hv^2 + gh^2/2) &= 0
\end{aligned}$$

on some domain Ω with boundary $\partial\Omega$. It is possible to rewrite this set of equations in a vector form:

$$\frac{\partial \varphi}{\partial t} + \frac{\partial J^x}{\partial x} + \frac{\partial J^y}{\partial y} = 0 \qquad (0.2)$$

$$\varphi = \begin{pmatrix} h \\ hu \\ hv \end{pmatrix} \quad J^x = \begin{pmatrix} hu \\ hu^2 + \frac{g}{2}h^2 \\ huv \end{pmatrix} \quad J^y = \begin{pmatrix} hv \\ huv \\ hv^2 + \frac{g}{2}h^2 \end{pmatrix} \tag{0.3}$$

If the "floor" of the system is flat and there is no sourcing term, this equation correctly describes the motion of the fluid if there are no external forces.

Although not discussed in this paper, if we want to examine a system where the bed of the water body is not flat (this situation is certainly more physical) a simple forcing term on the right hand side can be added, see [KC04] [CDMW98] [AC97]

Finite Volume Scheme

Since solving systems described by these equations analytically is infeasible, except in a few very few simple cases [Ach90], a numerical approach is needed. However, as demontrated above, a simple finite difference scheme is not well suited. As can be seen from (0.2) we can write the equation in a well known conservation law form

$$\frac{\partial \varphi}{\partial t} + \nabla \cdot \mathbf{F} = 0 \tag{0.4}$$

These types of equations arise in many places in physics and applied mathematics, especially in fluid mechanics. When written in this form, the conservation law is classified as a hyperbolic problem and is well suited for finite volume techniques [Lev04]. The idea of finite volume method is to break the domain up into cells and describe the changes in a cell over by considering fluxes through the cell boundary. For example, suppose we have cell 1 and cell 2 and we approximate the solution in cell 1 to be constant and the solution in cell 2 to be another, typically different, constant. Depending on the values of the constants the flux through the cell boundary will be different. Consider the simple fluid dynamical example where the flux is simply the velocity of the water. If the velocity of the water in cell 1 is greater then cell 2, the water will want to flow to the right. If it was greater in cell 2, the water would want to flow to the left. For more complicated situations, there can be more possible combinations.

Collectively, the problem of solving a PDE with a constant solution with a single discontinuity is known as a Riemann problem. The example given in the previous section is a very simple example of a Riemann problem. A more complete discussion can be found in [Lev04].

There are many choices for how to calculate this flux for a given Riemann problem. In particular we want to investigate the shallow water equations in situations that lend themselves well to description by triangular tilings. One potential way of determining the flux through a cell boundary is to simply take the average. However, if we do this, and write out the resulting scheme, we get an equation that is similar to the one derived above for finite difference which fails to capture the relevant important evolution behaviour. Instead, we modify the average by adding a correction term that involves the normal component of the flux through a cell which is effectively a viscous correction. This solution to the Riemann problem, is a variation of the Godunov scheme called a Roe solver. For a full derivation for the 1D shallow water equations, see [Lev04].

To derive a finite volume scheme, we first transform the conservation law form of the shallow wave equations into a form

$$\frac{\partial}{\partial t}\int_{\Omega} \varphi d\Omega + \oint_{\partial\Omega} dS\mathbf{J}\cdot\hat{\mathbf{n}} = 0 \qquad \mathbf{J}\cdot\hat{\mathbf{n}} = J^x n_x + J^y n_y$$

which is obtained by integrating over the domain and applying the divergence theorem. To proceed, we now break the integrals into components over each of the cell domains

$$\sum_i \frac{\partial}{\partial t}\int_{\Omega_i} \varphi_i d\Omega_i + \oint_{\partial\Omega_i} dS_i\mathbf{J}_i\cdot\hat{\mathbf{n}}_i = 0$$

where we are now integrating each of the functions over a cell labelled by i. However, we now make the approximation that in each of the cells, the variables are constant. In other words we make the approximation that

$$\int_{\Omega_i} \varphi_i d\Omega_i \approx \varphi_i d\Omega_i$$

$$\oint_{\partial\Omega_i} \mathbf{J}_i \cdot \hat{\mathbf{n}}_i dS \approx \sum_{j=k(i)} J_{i,j} \Delta l_j \qquad J_{i,j} = J_i^x n_j^x + J_i^y n_j^y$$

where Δl_j labels the length of the boundary and $k(i)$ is a list of the edges of the cell. Thus we now have an equation for each cell i of the form

$$\frac{\partial \varphi_i}{\partial t} = \frac{1}{\Delta\Omega_i} \sum_{j=k(i)} J_{i,j} \Delta l_j$$

We now make a simple finite difference approximation to the time derivative to obtain the scheme

$$\varphi_i^{n+1} = \varphi_i^n - \frac{\Delta t}{\Delta\Omega_i} \sum_{j=k(i)} J_{i,j} \Delta l_j \tag{0.5}$$

This is the finite volume scheme we will be applying in this paper. So far, we have left the geometry of the cells unspecified but we will now assume a triangular mesh. It is now important to note how this differs from the simpler rectangular finite volume meshes. For rectangular finite volume meshes this expression greatly simplifies since the area of each cell is identical as are the lengths. With a triangular grid, this is no longer true. Additionally, for a rectangular grid the normal vectors are very simple and one component of the flux $J_{i,j}$ will cancel out, however in a triangular grid, the normal vector varies from triangle to triangle and even each side of the triangle. While not more conceptually difficult, this procedure means that much more bookkeeping must be done.

Triangular Meshes

In order to implement triangular meshes we used the toolbox `pdetool` in MATLAB. This toolbox allows for the automatic creation triangular meshes when given a certain geometry. For the two problems we are considering, we have two different geometries given by Figures 1 and 2.

In order to translate to be implemented in the finite volume scheme we use the command `initmesh` which allows the creation of three matrices that encode all the information about the geometry. Using these three matrices we are able to encode all the vital information about the mesh. Unfortunately MATLAB does not order the triangles in any particular order so we had to implement a method of ordering the triangles. This was achieved by using a simple search over all the triangles and matching triangle vertices. Special care had to be paid for the boundary edges as MATLAB does not keep track of whether the triangle is a boundary point or not. From this bookkeeping, the normal vectors, lengths, and areas of the triangles were calculated and stored. For more information about how this is done specifically, see the Appendix.

The Riemann Problem

In the finite volume class of methods, finding the value of the fluxes at the interface is of primary importance, as this allows us to advance the system in time. For systems of nontrivial complexity, determining this value exactly is very difficult or impossible. A variety of approximation techniques have been developed to allow efficient calculation of the solution to the Riemann problem. One of these is the Roe solver, developed by Philip Roe. The Roe solver linearises the Jacobian of the fluxes over the normal vector. This allows us to calculate the flux at the interface relatively easily while still preserving features such as shocks. We write, following [AC97]

$$J_{i,j} = \frac{1}{2}\left[J(\varphi^{+}_{i,j}) + J(\varphi^{-}_{i,j}) - |A|(\varphi^{+}_{i,j} - \varphi^{-}_{i,j})\right] \tag{0.6}$$

where i, j refer to the i-th cell's j-th interface, φ^{+} refers to the values of φ in the current cell, and φ^{-} refers to the values of φ in the adjacent cell being considered.

$$A = \frac{\partial(\mathbf{J}\cdot\mathbf{n})}{\partial\varphi} \tag{0.7}$$

$$= \begin{bmatrix} 0 & \mathbf{n}\cdot\hat{\mathbf{x}} & \mathbf{n}\cdot\hat{\mathbf{y}} \\ (gh-u^2)\mathbf{n}\cdot\hat{\mathbf{x}} - uv\mathbf{n}\cdot\hat{\mathbf{y}} & 2u\mathbf{n}\cdot\hat{\mathbf{x}} + v\mathbf{n}\cdot\hat{\mathbf{y}} & u\mathbf{n}\cdot\hat{\mathbf{y}} \\ (gh-v^2)\mathbf{n}\cdot\hat{\mathbf{y}} - uv\mathbf{n}\cdot\hat{\mathbf{x}} & v\mathbf{n}\cdot\hat{\mathbf{x}} & u\mathbf{n}\cdot\hat{\mathbf{x}} + 2v\mathbf{n}\cdot\hat{\mathbf{y}} \end{bmatrix} \tag{0.8}$$

As we can see, we compute the average flux with a viscous correction of the form $-|A|(\varphi^{+}_{i,j} - \varphi^{-}_{i,j})$. We can split diagonalize this matrix to find $|A|$:

$$|A| = R|\Lambda|L \tag{0.9}$$

$$\Lambda = \begin{bmatrix} u\mathbf{n}\cdot\hat{\mathbf{x}}) + v\mathbf{n}\cdot\hat{\mathbf{y}} & 0 & 0 \\ 0 & u\mathbf{n}\cdot\hat{\mathbf{x}} + v\mathbf{n}\cdot\hat{\mathbf{y}} - \sqrt{gh} & 0 \\ 0 & 0 & u\mathbf{n}\cdot\hat{\mathbf{x}} + v\mathbf{n}\cdot\hat{\mathbf{y}} + \sqrt{gh} \end{bmatrix} \tag{0.10}$$

$$R = \begin{bmatrix} 0 & 1 & 1 \\ \mathbf{n}\cdot\hat{\mathbf{y}} & u - \sqrt{gh}\mathbf{n}\cdot\hat{\mathbf{x}} & u + \sqrt{gh}\mathbf{n}\cdot\hat{\mathbf{x}} \\ -\mathbf{n}\cdot\hat{\mathbf{x}} & v - \sqrt{gh}\mathbf{n}\cdot\hat{\mathbf{y}} & v + \sqrt{gh}\mathbf{n}\cdot\hat{\mathbf{y}} \end{bmatrix} \tag{0.11}$$

$$L = \begin{bmatrix} -(u\mathbf{n}\cdot\hat{\mathbf{y}} - v\mathbf{n}\cdot\hat{\mathbf{x}}) & \mathbf{n}\cdot\hat{\mathbf{y}} & -\mathbf{n}\cdot\hat{\mathbf{x}} \\ (u\mathbf{n}\cdot\hat{\mathbf{x}} + v\mathbf{n}\cdot\hat{\mathbf{y}})/2\sqrt{gh} + \frac{1}{2} & -\mathbf{n}\cdot\hat{\mathbf{x}}/2\sqrt{gh} & -\mathbf{n}\cdot\hat{\mathbf{y}}/2\sqrt{gh} \\ -(u\mathbf{n}\cdot\hat{\mathbf{x}} + v\mathbf{n}\cdot\hat{\mathbf{y}})/2\sqrt{gh} + \frac{1}{2} & \mathbf{n}\cdot\hat{\mathbf{x}}/2\sqrt{gh} & \mathbf{n}\cdot\hat{\mathbf{y}}/2\sqrt{gh} \end{bmatrix} \tag{0.12}$$

Where R is the right eigenvector matrix and L is the left eigenvector matrix.

Applications

We now have all the tools needed to solve the shallow water equations using a finite volume scheme. Using (0.5) we apply the Roe solver to $J_{i,j}$ and compute over each of the three cell sides.

Circular Dam

To demonstrate a simple application of the scheme we consider the classical problem of a circular dam using the mesh given by Figure 1. We now need to apply the proper boundary and initial conditions. Since we are dealing with an inviscid flow we have the no-slip boundary condition $\mathbf{u}\cdot\hat{n} = 0$ at the boundary. For the simple case of a rectangular box these conditions are obtained by choosing an appropriate flux through the boundary. This is given by

$$h_{-} = h_{+} \qquad (uh)_{-} = \pm(uh)_{+} \qquad (vh)_{-} = \mp(vh)_{+}$$

where the minus sign is chosen when the uh or vh are perpendicular to the wall. For initial condition, we simply chose an initial height of $h = 1.9$, see Figure 2. The simulation was run on three settings with varying numbers of triangles for 1000 timesteps. The plotted figures demonstrate the most refined mesh used which has roughly 5000 triangles. As can be seen in Figure 4, the triangular mesh is able to deal with the circular geometry very easily and no loss of the flux is obtained due to edge effects. For a video of the dam see [BH12a]

Rotterdam

For this scenario, we want to examine the behaviour of the Nieuwe Waterweg in Holland. This is a shipping canal created to allow passage of ships from the North Sea to the Europoort in Rotterdam, which is one of the world's busiest ports. The Netherlands is at great risk of flooding from the North Sea, and a surge down the Nieuwe Waterweg would prove disasterous to Rotterdam and the surrounding region. As part of Holland's Delta Works plan to construct flood barriers, dams, and surge protectors, a movable barrier was constructed in the Nieuwe Waterweg - the Maeslantkering (see Figure). This barrier sits in drydock most of the time, but when a surge is imminent its halves will swing out to save Rotterdam. The arms of the barrier take about 2 hours to close, and begin closing if the North Sea is likely to generate surges of 3 metres or more. First the drydocks are flooded and the wedges float out into the water and begin to move towards each other. After the gates have closed, they are filled with water (causing them to submerge) and then function to block the surges. In this simulation, we want to investigate the situation in which the Maeslantkering is hit with a wave while being hit from the sea. As opposed to the image of the Maeslantkering where it is fully opened, we consider ours midway closed to demonstrate the effects of how the structure breaks the waves. Additionally, we have simplified the geometry of the structure for modelling purposes.

0.0.1 Implementation

There are three boundaries we need to account for: the open water at the ends of the channel, the banks of the river, and the surge gate [CDMW98]. For the open sea, we fix u_B, which is the horizontal velocity coming from the North Sea. We are simulating a storm surge, so enforcing that the flow from the sea is always towards Rotterdam is reasonable. We also assume that the flow from the sea is always subcritical. In the field of fluid mechanics, a subcritical flow exists when the flow velocity is less than the wave velocity. Supercritical flows have the opposite property. A supercritical flow is analogous to a supersonic wave in air - we assume that a similar condition does not exist at the end of the waterway. At the other end, we assume that there is no flow from Rotterdam. This makes sense since the continent lies in that direction - the flows from the Rhine-Meuse-Scheldt delta are relatively small compared to those from the North Sea.

On triangles with edges facing the open sea, we fix the value for $\mathbf{u_r}$ as the boundary u_B, force $v_R = 0$ (we assume no perpendicular flow at the interface), and solve for h using:

$$u_r = u_L - \sqrt{g}(\sqrt{h_r} - \sqrt{h_l}) \tag{0.13}$$

We send the boundary condition to the Riemann solver by passing it as part of `adj_tri_info`, the matrix which contains the information about the triangles the current one interfaces with.

For the riverbank, we assume that the interface causes perfect reflection of the vertical flow and does not affect the horizontal flow (there are no eddy currents). Since we only examine inviscid flow, this is somewhat reasonable - in this case, there would be no boundary layer to affect the horizontal flows near the edge. However, it is not a very physical assumption. Water is not an ideal liquid and isn't inviscid, so in the Nieuwe Waterweg there will be boundary layers which our simulation doesn't take into account.

On the gate, we assume that if the height of the water is less than the height of the gate, then the gate also reflects the flows perfectly. However, we face an additional complication here - the gate faces are neither perfectly horizontal nor perfectly vertical. In order to find the resulting fluxes from the reflection, some algebra is necessary. Let $\mathbf{u}_r$ and $\mathbf{v}_r$ be the reflected flows at the interface with the gate. Since perfect reflection occurs, $|\mathbf{u}_r| = |\mathbf{u}|$ and $|\mathbf{v}_r| = |\mathbf{v}|$. We also specify that θ is the angle between the $\hat{x}$ direction and the normal vector at the interface with the gate.

$$\begin{aligned}
\mathbf{u}\cdot\mathbf{u}_r =& u^2\cos 2\theta & \mathbf{v}\cdot\mathbf{v}_r =& v^2\cos{(2\theta-\frac{\pi}{2})} = v^2\sin 2\theta \\
u\cdot -u_{r,x} =& u^2\cos 2\theta & v\cdot -v_{r,y} =& u^2\sin 2\theta \\
u_{r,x} =& -u\cos 2\theta & v_{r,y} =& -v\sin 2\theta \\
u =& \sqrt{u_{r,x}^2+u_{r,y}^2} & v =& \sqrt{v_{r,x}^2+v_{r,y}^2} \\
\Rightarrow u_{r,y} =& u\sin\theta & \Rightarrow v_{r,x} =& v\cos\theta
\end{aligned}$$

So that the interface fluxes u_R and v_R are:

$$u_R = -u\cos 2\theta - v\sin 2\theta \tag{0.14}$$

$$v_R = u\sin 2\theta + v\cos 2\theta \tag{0.15}$$

Where $\theta = \arctan(n_y/n_x)$.

When sending the wave down the waterway, we simulate a square wave by dropping the height of the sea after a few timesteps. Although the sea doesn't produce square waves, the wave "spreads out" due to our Riemann solver, creating a reasonable facsimile of an ocean wave. See Figure 6 for the wave after 15 timesteps

Results

For a video of the results, again see [BH12a]. Of particular interest is the reflective behaviour of the gates and the extremely dampened wave that makes it though the gap between them - the Maeslantkering seems to be an effective surge barrier, provided the gates themselves are not overwhelmed by a very tall wave.

Conclusion

We have implemented finite volume using triangular meshes with a Roe solver to obtain the evolution of a fluid in two geometries. The implementation provides a very robust numerical scheme that can be easily applied to many non-trivial geometries that simple rectangular finite volume methods are not able to handle well. Future work that can be done is to implement the shallow water wave equations with viscous effects as done in [AC97]. Additionally, we've neglected sea geometry and other force effects that might be present. All the code run, along with documentation can be found at [BH12b].

Appendix

In this section we describe how the bookkeeping methods operate. We have made an effort to get the code to run in the open source version of MATLAB, Octave, however Octave did not have the available tools that MATLAB has and we were unable to get it to run properly.

Creating the triangular mesh is done by using the MATLAB toolbox `pdetools` which, unfortunately, does not come standard with the student edition of MATLAB. In order to get around this, the matrices are provided. However, it will not be possible to experiment with other geometries without such toolboxes. Additionally, the way MATLAB does bookkeeping of the triangles is not very intuitive and much bookkeeping must be done before the numerical model is applied.

Initially we export the geometry in terms of three matrices, `p,e,t`. The `p` matrix contains the co-ordinates of the verticies of each the triangle. The matrix `t` contains the vertex labels of each the triangles in a counter-clockwise order. Thus in order to find the co-ordinates of a given triangle, one would look at the first three entries of `t` which tell us which entries of `p` to look at. For example, if we are considering the

15th triangle, the co-ordinates of the verticies are given by `p(:,t(1,15)),p(:,t(2,15)),p(:,t(3:15))`. Unfortunately, MATLAB does not order the triangles in any particular order, i.e. triangle 15 is not next to triangle 16 in `t`. Thus we must search through all the triangle verticies in order to determine which triangles are beside each other. This is done by a brute force search using the function `edgefind`. The way this function works is by simply taking the ith triangle and checking the vertices of one edge of the triangle with all the others. If the triangle is a boundary, this is handled approptiately. The resulting matrix `EdgeMatrix` tells us that if we have, say the ith triangle, it shares boundaries with triangles a,b,c. However the labelling is no particular order and it is now sorted via `order_triangles_b` (for triangles that lie on the boundary) and `order_trinagles_nb` otherwise. Thus we now have that in `EdgeMatrix` for the ith triangle, the normal vector of the first edge shares the boundary with first triangle in `EdgeMatrix` and so on. Finally we must assign, properly, the boundaries. This is contained in the matrix `e`. This matrix contains only the boundaries and the associated boundary number assigned to it by `pdetools`. Throughout we use this matrix to check whether or not the triangle is on the boundary. For more information on how MATLAB maintains all the triangle information see the help file.

The main parts of the programme are `TriInfo`,`TriData`. We have that

```
TriInfo(1) = Triangle Index
TriInfo(2,3) = components of n1 vector
TriInfo(4,5) = components of n2 vector
TriInfo(6,7) = components of n3 vector
TriInfo(8,) = whether the triangle considered is on the boundary
TriInfo(9,10,11) = lengths of n1,n2,n3
TriInfo(12) = area of triangle
```

which contains all the information about each triangle while `TriInfo` contains

```
TriData(1) = Triangle index
TriData(2) = h
TriData(3) = uh
TriData(4) = vh
```

which corresponds to φ.

References

[AC97] K. Anastasiou and C.T. Chan, *Solution of the 2d shallow water equations using the finite volume method on unstructured triangular meshes*, Int. J. Numer. Meth. Fluids **24** (1997), 1225–1245.

[Ach90] D.J. Acheson, *Elementary fluid dynamics*, Oxford Applied Mathematics and Computing Series, 1990.

[BH12a] L. Bovard and K.S. Hyatt, 2012.

[BH12b] ———, 2012.

[CDMW98] S. Chippada, C.N. Dawson, M.L. Martinez, and M.F. Wheeler, *A Godunov-type finite volume method for the system of shallow water equations*, Computer Methods in Applied Mechanics and Engineering **151** (1998).

[KC04] P. Kundu and I. Cohen, *Fluid mechanics*, Elsevier Academic Press, 2004.

[Lev04] R.J. Leveque, *Finite-volume methods for hyperbolic problems*, Cambridge University Press, 2004.

[Ray] D. Raymond.

ISSN 1927-1417 (Print) ISSN 1927-1425 (Online)