Beyond Expectations, But Within Limits — The Theory of Coherent Risk Measures

Yuxi Liu

October 2019

A thesis submitted for the degree of Bachelor of Science (Advanced) (Honours)
of the Australian National University

For the glory of Celestia,
And Her dominion come.

Declaration

The work in this thesis is my own except where otherwise stated.

Yuxi Liu

Acknowledgements

I would like to thank the following people for the successful completion of my thesis.

Main supervisor, Robert Williamson, who used his knowledge in the fields of risk measure and philosophy to give valuable guidance for me. The weekly meetings were vital for keeping me afloat in an unfamiliar ocean of knowledge.

Co-supervisor, Markus Hegland, who, while I have not met for many times, gave good feedback during the writing phase of my thesis.

My parents, who gave unwavering material support.

Best friends, Artline, Ira, and Lurid, who kept me alive during the years.

Abstract

Expectation lies at the foundation of probability, but its centrality belies its contingency. Abstractly, the expectation of a random variable is a real number that measures some information about the variable, and one may well explore the consequences of replacing expectation by some alternative.

Certain alternatives to expectation, termed "coherent risk measures", have been well-investigated in financial engineering, but they are relatively unknown in the field of machine learning. Here, we collect and prove some fundamental properties of coherent risk measures that we believe would be applicable to machine learning.

In Chapter 1, we give a guide to the thesis, then we review the concept of risk measures, point out possible deficiencies of the expectation as a risk measure, and provide a historical overview of the study of risk measures in finance and other areas.

In Chapter 2, we review basic probability concepts, then define the concept of coherent risk measures and study the geometric properties of their envelope representations. Armed with geometric insight, we prove a Kusuoka representation theorem when the underlying sample space is finite and uniform, and construct counterexamples when it is finite but nonuniform.

In Chapter 3, we generalize some basic probability inequalities and concentration inequalities from expectation to conditional value at risk. Then we review statistical learning theory and generalize its basic concepts and its fundamental theorem by replacing expectation with spectral risk measures.

In Chapter 4, we review limit theorems in probability, give a new and intuitive proof of the central limit theorem for the empirical estimator of CVaR. We also prove the uniform strong law of large numbers for the empirical estimator of spectral risk measures. We provide numerical evidence to support our results and generate conjectures.

In Chapter 5, we summarize the main theorems and conjectures of the thesis,

review the literature on applications of general risk measures to machine learning, and point to possible future research directions.

Contents

Notation, convention, and terminology

Notation and convention

Conventions are choices that are purely for convenience and disambiguation, with no deep significance.

$\mathbb{N}$	The set of natural numbers is $\{1, 2, ...\}$.
n	A positive integer, unless otherwise noted.
$[n]$	$\{1, 2, ..., n\}$, with caveat that $n \geq 1$
c	A real constant, unless otherwise noted.
x^+	Positive part of x, that is, $\max(x, 0)$.
∂A	Topological boundary of A.
$\mathrm{cl}(A)$	The topological closure of A.
$\mathrm{co}(A)$	The convex hull of A, where A is a subset of some real vector space.
$(S, \mathcal{B}, \nu)$	Probability space as formalized in Kolmogorov probability theory. See Definition 2.1. When S is countable, $\mathcal{B}$ is its power set. When S has a topology, $\mathcal{B}$ is its Borel σ-algebra.
$\mu, \nu, ...$	Probability measures are written in Greek minuscule.
$Pr(F)$	Probability of event F.
$X, Y, ...$	Random variables are written in Latin majuscule. All random variables are real-valued unless otherwise stated.

$\mathbb{E}_\nu(X)$	Expectation of random variable X. The subscript ν means that X is based on a probability space $(S, \mathcal{B}, \nu)$, and is often omitted.
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2. See definition 4.2.
δ_x	The Dirac delta distribution at $x \in \mathbb{R}$. For any Borel subset A of $\mathbb{R}$, $\delta_x(A) = 1_{x \in A}$.
$\sum_{i=1}^n p_i \delta_{x_i}$	A discrete probability distribution, with $p_i \geq 0$, $\sum_{i=1}^n p_i = 1$.
μ_X	Given random variable X, μ_X is the probability measure on $\mathbb{R}$, such that for any Borel subset A of $\mathbb{R}$, $\mu_X(A) = Pr(X \in A)$.
$X \sim \mu$	X has the probability measure μ. That is, $\mu_X = \mu$.
$X \overset{\mathrm{d}}{=} Y$	X, Y have the same distribution. That is, $\mu_X = \mu_Y$.
$X \equiv Y$	$X = Y$ almost surely.
$X_n \overset{\mathrm{d}}{\to} Y$	X_n converges to Y in distribution.
$X_n \overset{\mathrm{Pr}}{\to} Y$	X_n converges to Y in probability.
$X_n \overset{\mathrm{a.s.}}{\to} Y$	X_n converges to Y almost surely.
$\mathbb{1}$	A random variable that has constant value 1. When no confusion could arise, $c\mathbb{1}$ may be written as c.
α	A real number in $[0, 1]$, unless otherwise noted.
$\bar{\alpha} = 1 - \alpha$	This simplifies some equations.
ρ	A probability density function.
$\mathscr{L}(S)$	The set of all random variables on S. If no confusion would arise, S is omitted.
$\mathscr{L}^p$	The set of all random variables with finite p-moment.
$\mathscr{L}_+$	The set of all random variables that are almost surely nonnegative.
$\mathcal{F}, \mathcal{R}, \mathcal{V}...$	Letters in calligraphic font are risk measures on $\mathscr{L}^2$.

$\mathscr{F}, \mathscr{R}, \mathscr{V} \ldots$	Letters in the script font are subsets of $\mathscr{L}^2$. See Notation 2.21 and Equation 2.23.
$\langle X, Y \rangle$	The inner product on $\mathscr{L}^2$, defined as $\mathbb{E}(XY)$.
(F_n)	Empirical cumulative distribution function. See Definition 2.14.
(L_n)	Empirical process. See Definition 2.14.

Terminology

risk measure	A function of type $A \to B$, where $A \subseteq \mathscr{L}, B \subseteq [-\infty, +\infty]$.
coherent	A possible property of a risk measure. Other possible properties include sublinear, subadditive, risk averse, etc. See Definition 2.22.
CRM	Coherent risk measure.
IID	Indepedent and identically distributed.
PDF	Probability density function.
CDF	Cumulative distribution function.
CLT	Central limit theorem.
SLLN	Strong law of large numbers.
WLLN	Weak law of large numbers.
VaR_α	Value at risk at level α. See Example 2.12
CVaR_α	Conditional value at risk at level α. See Definition 2.25.
EVaR_α	Entropic value at risk at level α. See Definition 4.25
SLT	Statistical learning theory.
ERM	Empirical risk minimization. See Definition 3.19.

PAC-learning	Probably approximately correct learning. See Definition 3.18.
VCdim	Vapnik–Chervonenkis dimension. See Definition 3.25.

Chapter 1

Introduction

1.1 How to read the thesis

The abstract provides an overview of the whole thesis.

Start with Chapter 1, which should be accessible for general readers with basic college mathematics.

Proceed to the first half of Chapter 2, up to Section 2.2, which gives a compressed introduction to probability and risk measures, necessary for understanding the rest of the paper.

The reader should then move to Chapter 5 for a summary of the main results in the paper, then study whichever appears the most interesting. This is assisted by the fact that there is little interdependence in the rest of the paper.

Most of the proofs in the paper may be skipped, as they are either trivial (such as that of Proposition 3.4) or highly technical (such as that of 3.12), and thus unlikely to be of general mathematical interest. However, we believe that our new proof of the central limit theorem of CVaR (Section 4.2.1), while technical and computational, is interesting, and recommend that any reader who has expertise in large deviation theory may profitably study it in detail.

1.2 Start with a problem

In online commerce, fraudulent accounts pose a constant threat. As such, softwares are written that can automatically detect suspicious accounts and suspend them. Such a software works by taking an account's activity log, and computing a judgment based on it: "fraudulent" or "honest".

This is a concrete example of the problem of classification. Similar problems include detecting suspicious activities in social media accounts, classifying images into categories, and handwriting recognition.

The common way in which classification problems are formalized is by defining a feature space $\mathcal{X}$ and a label space $\mathcal{Y}$, and a probability distribution μ on the space $\mathcal{X} \times \mathcal{Y}$. For example, in the case of handwritten digit recognition, the feature space could be the space of all grayscale images with resolution 256×256, and the sample space could be the set of all numerals: $\{0, 1, ..., 9\}$. For any feature-label pair $(I, n) \in \mathcal{X} \times \mathcal{Y}$, $\mu(\{(I, n)\})$ is the probability of encountering such a feature-label pair, and $\mu(\{(I, 1) | I \in \mathcal{X}\})$ is the probability of encountering any handwritten digit 1.

1.2.1 What is the right thing to do?

A classification problem is a special case of a rational decision problem, and most of the current theory on rational decision is formalized after the 1940s. The most influential model of rational decision is that of **expected utility maximization**, propounded by von Neumann, Morgenstein, Savage, and many others.

Simply put, expected utility maximization states that a rational person would have a "utility function" that assigns a **utility**, that is, a real number symbolizing how much they prefer a certain outcome, to every possible outcome of their decision. They would take the decision that **maximizes** the **expectation** of utility.

Convention 1.1. Throughout this thesis, we will only talk about **loss** instead of utility. This is just a sign convention, as loss is the negative of utility. Expected utility maximization becomes expected loss minimization.

According to this theory, to solve the question of classification, a rational agent would start by deciding on a loss function, then find the classification algorithm that minimizes the expectation of loss.

Concretely, the loss function can be defined as $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, such that $\ell(y, y')$ measures how bad it is to classify an object as y when it is actually y'.

A classification algorithm is some $f : \mathcal{X} \to \mathcal{Y}$.

Let $\mathcal{M}$ be the set of all classification algorithms that the agent can think of, then the best classification algorithm is

$$f = \underset{f \in \mathcal{M}}{\arg\min}\, \mathbb{E}_{(X,Y)\sim\mu}(\ell(f(X), Y)) \tag{1.1}$$

This is usually how the solution is given, but this is not necessarily the best solution. The issue lies within the use of expectations.

1.2.2 Is expectation the right thing to calculate?

There are two kinds of decision theories: **descriptive** and **normative**. Descriptive theories predict how people actually decide, while normative theories tell how people should decide. Expected utility maximization used to be both a descriptive and normative theory. Psychologists and economists considered it an accurate description of how humans decide when they are thinking clearly, and philosophers considered it the correct standard for rationality.

This has come under attack on both fronts.

On the descriptive front, the work of Kahneman and Tversky since 1970s [TK74] has made it clear that humans do not minimize expectations of loss. They perceive probabilities in a distorted way. They regard losing 1 dollar as a lot worse than gaining 1 dollar.

On the normative front, directly against Kahneman, who recommends that people attempt to avoid making such irrational deviations in decision-making, Gigerenzer since 1990s [GB09] has proposed that these deviations from expected loss minimization are shortcuts in reasoning that are vital for real humans, who do not have unlimited time and thinking ability.

Regardless of one's normative stance, the requirements for an accurate description of how people make decisions means we would do well to not restrict ourselves to expected loss minimization formalism.

1.2.3 The deadly long tail

> The climate system is an angry beast and we are poking it with sticks.
>
> Wallace Broecker

Another criticism to the normative theory of expected loss minimization is that expectation is a very impoverished standard with which to measure the desirability of possible outcomes. In particular, it does not adequately account for extremely bad outcomes, in certain situations where the expectation of loss is not as relevant as the possibility of a very large loss.

Consider the problem of designing a drug, with the loss function being the total number of lives lost after administering the drug, then the expectation can be lowered by making the drug marginally better in most situations, but greatly worse in a few situations.

Such loss is said to have a "long tail", that is, it has a small but non-negligible chance of causing great harm.

Returning briefly to human psychology, the anthropologist Jared Diamond argues that [Dia13, Chapter 7] humans in traditional ("primitive") societies exercise "constructive paranoia" towards the long tail. An activity (sleeping under a dead tree) that has a tiny chance of great danger (the tree falling down) is avoided whenever possible. That even if it would happen on average once every 1000 years, they would not do it. In effect, their decisions are less about expected risk minimization, but more about extreme tail risk minimization.

In fact, the flurry of activities on studies of rational behavior and game theory after 1940s was motivated in no small part by the specter of thermonuclear war and human extinction, the greatest of all extreme tail risks.

While the cold war has ended in 1990s, the tail risk of nuclear war has not been eliminated. One survey [15] among nuclear policy experts found that the national security experts give on average a 7% chance of nuclear war killing more people than World War Two in the next 25 years, which roughly corroborates with the result from another survey [BS08].

The other extreme risk faced by modern humans is climate change. Without swift action to limit atmospheric concentration of greenhouse gas, the global average temperature is expected to rise by more than two degrees Celsius by 2100. While this expected value is tolerable, the temperature rise has a deadly long tail, up to more than six degrees. [Wei09] in particular proposes that, since the loss is so great at the tail end of climate change, and the tail end is so uncertain, merely minimizing expectation of loss is unwise.

Or in simpler words: it matters a great deal if there is a small, uncertain chance of warming by 10 degrees Celsius, even if we don't know if it is 1% or 0.01%, because it would be the end of human civilization.

In such cases, especially cases where the long tail is uncertain, mere expectation appear to be deficient in giving a full picture of the risk, and thus risk measures that are sensitive to more details of the shape of the risk would be useful.

1.2.4 Further reading

For more history on expected utility theory, see [Fis89]. For a philosophical analysis of normative expected utility theory, see [Bri19].

For an overview of Kahneman and Tversky's research into descriptive decision theory, [Kah11] is a very readable book, in which Kahneman repeatedly advises the reader to control their hardwired irrationality and follow the expected loss minimization principle. On the other side of the spectrum is Gigerenzer's popular book [Gig07], which praises gut feeling as more practical than rational calculations.

The intuition that a risk manager should pay more attention to avoiding catastrophic outcomes has been formalized, from a legal point of view, as the (catastrophic) precautionary principle [Sun07].

1.3 Traditions of risk measurement

1.3.1 Financial mathematics

Financial mathematicians, while unconcerned with classification problems in machine learning, have been heavy users of risk measures. Crudely, investment could be thought of as a binary classification problem: given a portfolio constructed from financial products (stocks, bonds, foreign currencies, etc), one must judge whether this portfolio is an "acceptable" or "unacceptable" investment.

Portfolio optimization

In more detail, the problem of portfolio optimization is to construct the "best" portfolio, subject to certain constraints, and financial mathematicians traditionally formalize this problem thus [AF19, section 3]: Consider an investor who wishes to optimize their net worth in one year, and has n financial products with which to construct their portfolio. A portfolio is then formalized as a real vector $X = (x_1, ..., x_n) \in \mathbb{R}^n$, with x_i denoting that the portfolio contains x_i units of product i.

There are many possible constraints to consider. For example, suppose the investor cannot hold negative amount of products ("shorting" in financial jargon), then $x_i \geq 0$ for all i. Suppose the investor currently can invest no more than P, then $\sum_i x_i p_i \leq P$, where p_i is the price of a unit of product i. In general, such constraints are represented as a set $D \subseteq \mathbb{R}^n$ of possible portfolios.

After the year is up ("the portfolio has reached maturity"), the movements of the market during the year would determine the outcome of the portfolio. Formally, let $Y = (y_1, ...y_m)$ be a real vector of all the relevant facts about the market, then the outcome of the portfolio is a function of X and Y. Let it be

$L(X, Y)$. In order to cast this in the language of loss minimization, L represents loss, so if the portfolio earned money at the end of the year, it gives a negative L.

Since Y is uncertain, it is modeled as a random variable. Thus, for each pick of portfolio X, $L(X, Y)$ is a random variable representing the outcome at maturity.

Then, the problem of portfolio optimization is to find

$$\arg\min_{X \in D} \mathcal{F}(L(X, Y)) \tag{1.2}$$

where $\mathcal{F}$ is a risk measure that the investor chose to represent how they feel about possible losses. A very risk-neutral investor could choose $\mathcal{F} = \mathbb{E}$, while a very risk-averse investor could choose $\mathcal{F} = \max$.

Modern portfolio theory

Modern portfolio theory, or mean-variance analysis, was initiated by [Mar52], and postulates that the investor is interested in only two numbers: the expectation and variance of investment returns. An investor, in this theory, always chooses the portfolio with the least variance out of all portfolios that have the same expectation.

In the language of risk measures, let the investment return be denoted by the random variable L. To cast it in the language of loss minimization, let L be the amount of money *lost* in the investment. The goal is then to minimize variance of L, under the constraint that the expectation of L is lower than some fixed constant, representing the investor's tolerable expectation of loss.

Then, we can represent this as minimization of $\mathcal{F}(L)$, where $\mathcal{F}(L) = \mathbb{E}(L) + \lambda\sigma(L)$, where $\mathbb{E}$ is the expectation, σ is the standard deviation, and λ is a constant that represents how variance-averse the investor is. Here, σ instead of σ^2 is used, since the unit of risk measure should be in dollars, while the unit of σ^2 is $(\text{dollar})^2$.

A big λ represents a strong desire to keep the standard deviation down, while λ close to zero represents an investor that is indifferent to variance, and behaves similar to a classical rational agent who only aims to maximize expectation. A negative λ represents an investor who prefers variance, the opposite of what modern portfolio theory assumes, but in no way invalid. Indeed, some investors recommend limited risk-seeking investment as a wise way to benefit from unexpected boons [Tal12].

Criticisms

There are many criticisms of modern portfolio theory, which is not surprising considering it is over 60 years old now.

One main criticism is that variance and expectation do not characterize a distribution sufficiently. For example, consider a Gaussian distribution and a distribution with density $\mathcal{F}(x) \approx \frac{1}{x^4}$ for large x. They may have the same expectation and variance, but one decays far faster than the other. Put it more explicitly, if human height were distributed like $\mathcal{F}(x) \approx \frac{1}{x^4}$, then the tallest man in the world would very likely be several meters high at least. This does not happen, as human height, conditional on sex, is almost Gaussian distributed.

Value-at-risk (VaR)

Other than the mean-and-variance risk measure used by modern portfolio theory, the quantile, or value-at-risk (VaR), is another risk measure that is popular in finance. For any real random variable X, any $0 \leq \alpha \leq 1$, the α-VaR of X is the α-quantile* of X.

Banks do not just keep their customers' money in a vault. They might loan money for interest, or trade stocks for profit. However, each investment exposes banks to risks, and to protect themselves from failing, banks are required to keep a certain amount of money in its vaults so that they are considered sufficiently prudent. The intuition of "sufficiently prudent", again, relies on a risk measure.

Given all the investments of a bank, its negative net worth in a year can be considered as a random variable X, and for X to be seen as sufficiently prudent, some kind of judge must examine it, and give a verdict of "prudent" or "imprudent". Just as before, this can be formalized as a risk measure $\mathcal{F}$, such that $\mathcal{F}(X) > 0$ denotes imprudence, and $\mathcal{F}(X) \leq 0$ denotes prudence.

Many international banks follow the Basel Accords, a sequence of recommendations on bank regulations. In particular, they describe risk measures for banks to evaluate their prudence. In Basel II, published in 2004, the risk measure was VaR, which cemented its position in financial risk management up until the crisis of 2008.

There are widespread criticisms of VaR [Dan+01], among which, the most basic one is its insensitivity to extreme losses. For example, suppose a financial

*Annoyingly, there exists a subtly different convention, where the α-VaR of X is the negative of $1-\alpha$-quantile of $-X$. This convention is used by, for example, [Art+99]. The distinction has no bearing on the mathematical content.

product has a 95%-VaR of \$10000, meaning that out of all the possible outcomes, among the worst 5%, the *best* outcome is losing \$10000. It might very well be that in the worst 1% cases, the product would lose a billion dollars, an extreme risk that is completely invisible in the 95%-VaR.

A more subtle argument against VaR is its non-convexity. That means that two low-risk products, when combined, can appear high-risk, which is a direct contradiction to the dogma that diversification reduces risk.

Coherent risk measures

In response to the criticisms, [Art+99] proposed axioms that any reasonable risk measure should satisfy, and they called such measures coherent risk measures, which is the main topic of this thesis.

The most commonly used coherent risk measure is the conditional value-at-risk (CVaR)[†], defined as the expectation of loss, conditional on the loss being worse than a certain level. So, for example, if a product has 95%-CVaR of \$10000, then among the worst 5% outcomes, the average is a loss of \$10000. In particular, this means that the probability of losing a huge sum of money must be small. The probability of losing over a million dollars, for example, must be less than 0.05%.

After the financial crisis of 2008, there was great suspicion that the use of VaR encouraged risky investments that contributed to the financial crisis, even resulting in a congressional hearing [09]. In reaction to this, VaR was changed to CVaR in Basel III, published in 2010.

For further reading, [Che14] is a detailed report on the history of VaR and CVaR in the Basel Accords.

1.3.2 Other traditions

Closely related to the financial tradition is the actuary tradition, where the study of tail risk is often called **ruin theory**. Ruin, in this context, denotes bankrupcy, often caused by rare but great losses. An insurance company can be ruined if a great earthquake struck all houses in a province. A bank can be ruined by a financial panic.

Further afield is the tradition of **reliability engineering**. In building a reliable house, the random variable X could stand for whether the house would

[†] Other names include "expected shortfall" (ES), "average value at risk" (AVaR), "conditional tail expectation" (CTE), "tail-VaR", and "mean excess".

fall down in the next earthquake. $X = 0$ for no and $X = 1$ for yes. Certainly, it is important to keep X as close to 0 as possible, but since reliability is not free, and there are competing priorities, such as budget limit, the architect cannot make X infinitely close to 0.

What can be done is then to define a risk measure $\mathcal{F}$, such that $\mathcal{F}(X)$ measures the risk measure from X, and the architect would tweak the design so as to minimize

$$\mathcal{F}(X) + (\text{risk measure from other risk factors}).$$

Chapter 2

The geometry of coherent risk measures

In this chapter, we first carefully set up the probability notation for the rest of the thesis. The impatient reader can skip over the section and refer to it only upon encountering difficulties in comprehension.

Then, we describe a geometric way to represent risk measures on random variables. This geometric viewpoint is then used to prove the Kusuoka representation for finite dimensional probability spaces with uniform probability distributions.

2.1 Basic probability definitions and notations

This section sets down and discusses basic probability definitions and notations.

2.1.1 Probability space

We set up the notations for Kolmogorov's probability axiomatization. For a detailed introduction to probability along this axiomatization, the reader is referred to [Chu01].

Notation 2.1. $(S, \mathcal{B}, \nu)$ is a **probability space**, with nonempty S as **state space**, $\mathcal{B}$ a **σ-algebra** on S, and ν a **probability measure** on $(S, \mathcal{B})$. The **events** of S are the elements of $\mathcal{B}$.

Convention 2.2. If S is countable, then unless otherwise noted, we assume $\mathcal{B} = 2^S$, that is, all subsets of S are measurable, and $\forall \omega \in S, \nu(\{\omega\}) > 0$, that is, all states have nonzero probability. This is an assumption of **nondegeneracy**.

Definition 2.3. Any $B \in \mathcal{B}$ with $\nu(B) > 0$ is **non-atomic** if and only if $\exists C \in \mathcal{B}$, such that $C \subseteq B, \nu(B) > \nu(C) > 0$. In other words, it has a subevent with smaller, but still nonzero, probability. S is **atomless** if and only if all of its subsets with nonzero probability are non-atomic.

Remark 2.4. In particular, if S is countable, then it is not atomless. In fact, it is the opposite of atomless, as every $B \in \mathcal{B}$ with $\nu(B) > 0$, any singleton subset of B is atomic.

Convention 2.5. Random variables, unless otherwise noted, are real, that is, they are real measurable functions on S. In order to precisely define real measurability, we must specify a σ-algebra on $\mathbb{R}$, which we choose to be the set of Borel measurable sets of $\mathbb{R}$. In general, whenever necessary to formalize $\mathbb{R}^n$-measurability, we choose the set of Borel measurable sets of $\mathbb{R}^n$.

Convention 2.6. Unless otherwise noted, (X_n) denotes a sequence of independent and identically-distributed real random variables, indexed by n, that has the same distribution as X. The index n ranges over $\mathbb{N}$.

Definition 2.7. (X_n) is the **IID process** of X. In general, any sequence of random variables is a **stochastic process**.

Notation 2.8. The constant-one random variable is $\mathbb{1}$, such that $\forall \omega \in S, \mathbb{1}(\omega) = 1$. We abuse notation slightly, so that if there is no confusion, any constant c can also denote the corresponding constant random variable $c \cdot \mathbb{1}$.

2.1.2 Probability distributions

In probability theory, the underlying probability space is often immaterial for the problem that is being studied. As Terence Tao noted [Tao10], probability theory can be said to be the study of measure spaces with measure one, but that is like saying number theory is the study of finite strings.

In particular, consider a coin, and let X represent the number of heads that come up if it is flipped once. Then, if the coin is fair, we have $Pr(X = 0) = Pr(X = 1) = \frac{1}{2}$. This can be formalized by defining an underlying probability space $S = \{1, 2\}$, with $\nu(1) = \nu(2) = \frac{1}{2}$, and $X(i) = i - 1$. However, all this formality accomplishes little in the way of understanding the probabilistic behavior of X.

The part of X that is of concern in probability theory is its distribution, that is, the probability of events involving X. For this particular X, its distribution is completely determined by $Pr(X = 0) = Pr(X = 1) = \frac{1}{2}$.

In general, a real random variable is well-described by its probability measure:

Definition 2.9. Given a real random variable X, its associated **probability measure** μ_X is defined by

$$\mu_X(E) = Pr(X \in E) \tag{2.1}$$

for any Borel measurable $E \subseteq \mathbb{R}$.

It is a standard result in measure theory that μ_X is determined by the cumulative distribution function F_X:

Definition 2.10. Given a real random variable X, its **cumulative distribution function** (CDF) is defined by

$$F_X(x) = Pr(X \leq x) = Pr(X \in (-\infty, x]) = \nu(X^{-1}((-\infty, x])) \tag{2.2}$$

It is a standard result in probability that the set of all possible CDF is the set of monotonically increasing, right-continuous real functions $F : \mathbb{R} \to [0,1]$, such that $\lim_{x \to -\infty} F(x) = 0, \lim_{x \to +\infty} F(x) = 1$.

Due to possible jump discontinuities, the inverse of F_X is not uniquely defined. The standard disambiguation is by demanding F_X^{-1} to be left-continuous:

Definition 2.11. Given a real random variable X, its **quantile function** F_X^{-1} : $[0,1] \to [-\infty, \infty]$ is defined by

$$F_X^{-1}(q) = \inf\{t : F_X(t) \geq q\} \tag{2.3}$$

Note that $F_X^{-1}(0) = -\infty$, and $F_X^{-1}(1) = \operatorname{ess\,sup}(X)$, which could be $+\infty$. For all $0 < \alpha < 1$ cases, $F_X^{-1}(\alpha)$ is real-valued.

Example 2.12. For any $0 \leq \alpha \leq 1$, the **value-at-risk at level** α of a random variable X is $\mathrm{VaR}_\alpha(X) = F_X^{-1}(\alpha)$.

Definition 2.13. Given X, let (X_n) be its IID process, then for any $n \in \mathbb{N}$, the n-th **empirical cumulative distribution function** of X is a random CDF defined by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,x]}(X_i) \tag{2.4}$$

where for any set C, 1_C is its indicator function.

Definition 2.14. The n-th **empirical measure** of X is the discrete measure

$$\mu_{X,n} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \tag{2.5}$$

where (X_n) is the IID process of X, and for any element x, δ_x is the Dirac delta measure, defined by

$$\forall E \subseteq \mathcal{B}, \quad \delta_x(E) = \begin{cases} 1 & \text{if } x \in E, \\ 0 & \text{else.} \end{cases}$$

For each $\mu_{X,n}$, let L_n be a discrete random variable that has probability measure equaling $\mu_{X,n}$, then the sequence of (L_n) defines the **empirical process** of X.

Remark 2.15. Since the empirical CDF $F_{X,n}$, and the empirical measure $\mu_{X,n}$, of X, are based on the IID process of X, which is itself random, $F_{X,n}$ and $\mu_{X,n}$ are thus random functions, while F_X and μ_X are deterministic.

Convention 2.16. Some functionals $\mathcal{F}$ on real random variables X, such as the expectation, are defined purely by the probability measure μ_X of X, which is determined by F_X. As such, we can unambiguously abuse notation:

$$\mathcal{F}(X) = \mathcal{F}(F_X) = \mathcal{F}(\mu_X) \tag{2.6}$$

As will be defined in Definition 2.22, this is equivalent to saying $\mathcal{F}$ is **law invariant**.

As an example, for any real random variable $X \in \mathscr{L}^1$, with CDF F_X and corresponding probability measure μ_X on $\mathbb{R}$, we have three equivalent definitions of expectation:

$$\begin{aligned} \mathbb{E}(X) &= \int_S X(\omega) d\nu(\omega) \\ = \mathbb{E}(F_X) &= \int_{[0,1]} F_X^{-1}(\alpha) d\alpha \\ = \mathbb{E}(\mu_X) &= \int_{\mathbb{R}} x d\mu_X(x) \end{aligned} \tag{2.7}$$

Example 2.17. The n**-th empirical mean** of X is

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}(F_{X,n}).$$

To say X satisfies the Strong Law of Large Numbers (SLLN) is to say that

$$\mathbb{E}(F_{X,n}) \overset{\text{a.s.}}{\to} \mathbb{E}(F_X) = \mathbb{E}(X),$$

where $\overset{\text{a.s.}}{\to}$ denotes almost-sure convergence.

This is the first example of the general phenomenon, where, if X is a "well-behaved" random variable, $F_{X,n}$ converges in some sense to F_X as $n \to \infty$. That is, the empirical distribution $F_{X,n}$ approximates certain aspects of the true distribution F_X with arbitrary precision, as $n \to \infty$.

In Chapter 4, we will prove several generalizations of the Central Limit Theorem that exhibit this general phenomenon.

2.1.3 Hilbert space of real random variables

Notation 2.18. Given a probability space $(S, \mathcal{B}, \nu)$, let $\mathscr{L}(S)$ be the space of all real random variables over S. Let $\mathscr{L}^p(S)$ be the space of all real random variables with finite p-moment. When no confusion could arise, S is omitted.

The case of $p = 2$ is special, as $\mathscr{L}^2$ is a Hilbert space of square-integrable real functions of S.

Convention 2.19. Unless otherwise noted, real random variables have finite variance. That is, we restrict our attention to elements of $\mathscr{L}^2(S)$

Notation 2.20. The inner product on $\mathscr{L}^2$ is $\langle X, Y\rangle = \mathbb{E}(XY)$.

Notation 2.21. Certain special sets are:

- $\mathscr{E}_{=c} = \{X \in \mathscr{L}^2 : \mathbb{E}(X) = c\}$.
- $\mathscr{L}^2_+ = \{X \in \mathscr{L}^2 : X \geq 0\}$. This is called the nonnegative quadrant of $\mathscr{L}^2$.
- $\mathscr{D} = \mathscr{E}_{=1} \cap \mathscr{L}^2_+$. This is the set of all random variables that are nonnegative, and have expectation one. If S is finite, then it is a simplex, which is often written with the letter Δ ("Delta", the Greek letter that looks like a simplex).
- For any $p \in [1, \infty)$, $\mathscr{U}_p = \{X \in \mathscr{L}^2 : \mathbb{E}(|X|^p) \leq 1\}$. This is the unit ball in p-norm. By Hölder's inequality, for all $p \geq q \geq 1$, $\mathscr{U}_p \subseteq \mathscr{U}_q$.
- $\mathscr{U}_\infty = \{X \in \mathscr{L}^2(S) : -1 \leq X \leq 1 \text{ almost surely}\}$. It can be thought of as the limit that is,

$$\mathscr{U}_\infty = \lim_{p\to\infty} \mathscr{U}_p = \bigcap_{p\geq 1} \mathscr{U}_p$$

.

2.1.4 Functionals on random variables

Definition 2.22. We define special properties of functional $\mathcal{F}$ on $\mathscr{L}^2$. In the following list, these conditions are added to the front: $\forall c \in \mathbb{R}, Z, Z' \in \mathscr{L}^2$.

(1) **Subadditivity.** $\mathcal{F}(Z + Z') \leq \mathcal{F}(Z) + \mathcal{F}(Z')$.

(2) **Positive homogeneity.** $\mathcal{F}(\lambda Z) = \lambda \mathcal{F}(Z)$, for all $\lambda \geq 0$. Note that this implies $\mathcal{F}(0) = 0$

(3) **Convexity.** $\mathcal{F}((1-\lambda)Z + \lambda Z') \leq (1-\lambda)\mathcal{F}(Z) + \lambda \mathcal{F}(Z')$, for all $0 \leq \lambda \leq 1$.

(4) **Sublinearity.** Subadditive and positive homogeneous. This implies convexity.

(5) **Monotonicity.** $\mathcal{F}(Z) \leq \mathcal{F}(Z')$ whenever $Z \leq Z'$ ν-a.s..
That is, when $\nu(\{s \in S | Z(s) \leq Z'(s)\}) = 1$

(6) **Translation invariance.** $\mathcal{F}(Z + c) = \mathcal{F}(Z) + c$.

(7) **Coherence.** Sublinear, monotone, and translation invariant.

(8) **Closedness.** $\{Z \in \mathscr{L}^2 | \mathcal{F}(Z) \leq c\}$ is closed. Note that the topology on $\mathscr{L}^2$ is defined by its inner product.

(9) **Risk aversity.** $\mathcal{F} \geq \mathbb{E}$.

(10) **Strict risk aversity.** $\mathcal{F} \geq \mathbb{E}$, with equality reached *only* for almost surely constant random variables.

(11) **Law invariance.** $X \overset{\mathrm{d}}{=} Y$ implies $\mathcal{F}(X) = \mathcal{F}(Y)$.

Each of these properties can be interpreted as formalizing practical properties of risk measures:

(1) **Subadditivity**: “merger does not create extra risk”. See [Art+99] for a detailed discussion.

(2) **Positive homogeneity**: doubling the outcome in all cases doubles the risk. This is sometimes called “scale invariance”.

(3) **Convexity**: iversification can only decrease risk, that is, holding stocks in a certain proportion has less risk compared to holding them separately in the same proportion.

(4) **Sublinearity**: Positively homogeneous and convex.

(5) **Monotonicity**: if in all cases, the outcome is not better, then the risk is not lower.

(6) **Translation invariance**: adding a sure loss of c increases risk by c.

(7) **Coherence**: to be interpreted in Section 2.2.

(8) **Closedness**: A technical assumption. A closed risk measure has convenient analytical properties, such as being lower semicontinous, and satisfying a Fatou's lemma. See [Kus01, Theorem 2] for details.

(9) **Risk aversity**: As noted in Section 1.3.2, in classical decision theory, a rational agent maximizes its expectation of utility, and is unconcerned with variances, no matter how extreme. Humans, in contrast, are often "risk averse", that is, they often give up a higher expectation if the variance is too great, indicating that they think such situations have a higher risk than the mere expectation.

(10) **Strict risk aversity**: a more exacting risk aversity. If X has any non-determinancy in its outcome, it is regarded as more risky than a sure loss of $\mathbb{E}(X)$.

(11) **Law invariance**: As noted in Section 2.1.2, the distribution functions of random variables are all that matters in probability, and so any risk measure meaningful in probability must be insensitive to any change of random variable that leaves its distribution unchanged, that is, it must be law invariant.

For more extensive interpretations of these properties, the reader is directed to the references in [Gia06].

2.2 Coherent risk measures (CRM)

2.2.1 Importance of coherence

The idea of coherence in risk measurement was first proposed in the context of financial mathematics by [Art+99]. Artzner explicitly argued against using VaR, due to its incoherence.

Example 2.23 (VaR is incoherent)**.** Let $Pr(X = -1) = Pr(X = +1) = 0.5$, and Y be an independent copy of X, then

$$\mathrm{VaR}_{0.49}(X + Y) = 0 > \mathrm{VaR}_{0.49}(X) + \mathrm{VaR}_{0.49}(Y) = -2.$$

Requiring a risk measure to be coherent incorporates several intuitions in judging the risk of financial products, and by extension, risky non-financial situations. Detailed interpretation of these risk measurement intuitions are found in [Art+99].

2.2.2 Conditional VaR (CVaR)

The conditional value-at-risk (CVaR) was proposed to be a coherent alternative to VaR, and has achieved a preeminent position in financial risk management. Intuitively, the CVaR at level α of a random loss X is the expectation of loss, conditional on the loss being the worst $(1 - \alpha)$ cases. That is,

$$\mathrm{CVaR}_\alpha(X) = \mathbb{E}(X|X > F_X(\alpha)) \tag{2.8}$$

for any $0 \leq \alpha < 1$. Note that when $\alpha = 0$, $F_X(\alpha) = -\infty$, and so $\mathbb{E}(X|X > F_X(\alpha)) = \mathbb{E}(X)$.

Remark 2.24. We will concentrate on the cases of $0 < \alpha < 1$ when discussing CVaR_α, since $\alpha = 0$ gives expectation, and $\alpha = 1$ gives essential supremum, both cases being often easier to handle.

This naive definition unfortunately does not work when X is atomic, because in such cases, F_X has jump discontinuities where the value of $F_X(\alpha)$ is ambiguous. Fortunately, there is a more general definition that overcomes such problems [RU02, Definition 3]:

Definition 2.25. For any $0 \leq \alpha < 1$.

$$\mathrm{CVaR}_\alpha(X) = \mathbb{E}(X^{(\alpha)}) \tag{2.9}$$

where $X^{(\alpha)}$ is a random variable with the CDF

$$F_{X^{(\alpha)}}(x) = \left(\frac{F_X(x) - \alpha}{1 - \alpha}\right)^+ \tag{2.10}$$

Intuitively, to get the graph of $F_{X^{(\alpha)}}$, take the graph of F_X, truncate it above the $y = \alpha$ line, and stretch it down to fill the $0 \leq y \leq 1$ stripe again.

Theorem 2.26 (CVaR is coherent, strictly risk averse, and law invariant)**.** *For any* $0 \leq \alpha \leq 1$, CVaR_α *is coherent, strictly risk-averse, and law invariant.*

The only difficulty is in proving the subadditivity of CVaR. To show it, we utilize equivalent ways to define CVaR, offering different perspectives on it. The most useful ones for our purpose are

Proposition 2.27 (Equivalent formulations of CVaR)**.** *For any real random variable* X*, and any* $0 \leq \alpha < 1$*, we have*

$$\mathrm{CVaR}_\alpha(X) = \min_{s\in\mathbb{R}} \left(s + \frac{1}{\bar{\alpha}} \mathbb{E}\left((X-s)^+\right)\right) = \frac{1}{\bar{\alpha}} \int_\alpha^1 F_X^{-1}(q)dq \tag{2.11}$$

Proof. See [RU02, Theorem 10] and [AT02, Proposition 3.2]. □

As an example of the power of such representation, the subadditivity of CVaR is now immediate:

Proof. For any $\alpha \in (0,1)$, and real random variables X, Y, let

$$s_1 \in \arg\min_{s\in\mathbb{R}} \left(s + \frac{1}{\bar{\alpha}} \mathbb{E}\left((X-s)^+\right)\right), s_2 \in \arg\min_{s\in\mathbb{R}} \left(s + \frac{1}{\bar{\alpha}} \mathbb{E}\left((Y-s)^+\right)\right),$$

and

$$s_0 = s_1 + s_2,$$

then we use the minimization definition of CVaR (Equation 2.11):

$$\begin{aligned}
\mathrm{CVaR}_\alpha(X+Y) &= \min_{s\in\mathbb{R}} \left(s + \frac{1}{\bar{\alpha}} \mathbb{E}\left((X+Y-s)^+\right)\right) \\
&\leq s_0 + \frac{1}{\bar{\alpha}} \mathbb{E}\left((X+Y-s_0)^+\right) \\
&= s_1 + s_2 + \frac{1}{\bar{\alpha}} \mathbb{E}\left((X-s_1+Y-s_2)^+\right) \\
&\leq \left(s_1 + \frac{1}{\bar{\alpha}} \mathbb{E}\left((X-s_1)^+\right)\right) + \left(s_2 + \frac{1}{\bar{\alpha}} \mathbb{E}\left((Y-s_2)^+\right)\right) \\
&= \mathrm{CVaR}_\alpha(X) + \mathrm{CVaR}_\alpha(Y),
\end{aligned}$$

where we used the fact that for any two real numbers x, y, $(x+y)^+ \leq x^+ + y^+$. □

Proposition 2.28 (Continuity of CVaR)**.** *For any real random variable* X, $\mathrm{CVaR}_\alpha(X)$ *is a continuous function on* $0 \leq \alpha \leq 1$.

Proof. By the integral definition of CVaR, $\mathrm{CVaR}_\alpha(X)$ is continuous on $0 \leq \alpha < 1$.

If $\operatorname{ess\,sup}(X) < \infty$, then for any $\epsilon > 0$, there exists α_0 such that any $\alpha > \alpha_0$, $F_X^{-1}(\alpha) > \operatorname{ess\,sup}(X) - \epsilon$, and so $\mathrm{CVaR}_\alpha(X) > \operatorname{ess\,sup}(X) - \epsilon$.

If $\operatorname{ess\,sup}(X) = \infty$, the proof is similar, with an arbitrarily big M replacing $\operatorname{ess\,sup}(X) - \epsilon$. □

2.2.3 The significance of CVaR

There are representation theorems, often named "Kusuoka representation", with the following format: On any "nice" probability space S, any law invariant CRM $\mathcal{F}$ that is "nice" can be represented by convex integrals of CVaR.

Now we state this rigorously.

Definition 2.29. Given a closed and sublinear functional $\mathcal{F}$ on $\mathscr{L}^2$, we say that it has a **Kusuoka representation** if and only if it can be represented as

$$\mathcal{F}(X) = \sup_{\theta \in \Theta} \int_{[0,1]} \mathrm{CVaR}_\alpha(X) dm_\theta(\alpha) \quad \text{for all } X \in \mathscr{L}^2 \tag{2.12}$$

where $\{m_\theta : \theta \in \Theta\}$ is a family of probability measures on $[0, 1]$.

Kusuoka representation theorems in general state that, if S and $\mathcal{F}$ satisfies certain conditions, then $\mathcal{F}$ has a Kusuoka representation [PR08, Section 2.2.4]. The original theorem has been generalized to dizzying heights of abstraction, which we will not review.

If the sup sign is removed, we obtain the class of spectral risk measures[*], proposed in [Ace02]:

Definition 2.30. A **spectral risk measure** is any $\mathcal{F}$ defined by a probability distribution m on $[0, 1]$, and

$$\mathcal{F} = \int_0^1 \mathrm{CVaR}_\alpha \, dm(\alpha) \tag{2.13}$$

Definition 2.31. Two random variables $X, Y : S \to \mathbb{R}$ are **comonotone** if for all $\omega, \omega' \in S$,

$$(X(\omega) - X(\omega'))(Y(\omega) - Y(\omega')) \geq 0, \tag{2.14}$$

that is, X, Y rises and falls together.

A risk measure $\mathcal{F}$ is comonotone additive if for any comonotone X, Y,

$$\mathcal{F}(X + Y) = \mathcal{F}(X) + \mathcal{F}(Y). \tag{2.15}$$

Proposition 2.32 (Kusuoka representation)**.** *Any risk function with Kusuoka representation is a law invariant coherent risk measure.*

If the probability space S is atomless, then the converse also holds.

A risk functional is a law invariant coherent and comonotone additive functional if and only if it is a spectral risk measure.

Proof. See [NR15, Theorem 3.1]. □

[*] Such representation is sometimes called Choquet representation, for example in [PR08, Definition 2.48].

2.3 The envelope representation of risk measures

In math, there is a common duality between analysis and geometry. The risk measures, being analytical, have the dual representation as risk envelopes.

Definition 2.33. Given any nonempty subset $\mathscr{F} \subseteq \mathscr{L}^2$, its associated **support function** is

$$\sigma_{\mathscr{F}}(X) = \sup\{\langle X, F\rangle \,|F \in \mathscr{F}\}$$

which we often write as $\mathcal{F}$.

Given a functional $\mathcal{F}$, if there exists some $\mathscr{F} \subseteq \mathscr{L}^2$, such that $\mathcal{F} = \sigma_{\mathscr{F}}$, then we say that it has an **envelope representation** $\mathscr{F}$.

2.3.1 The symmetry group on $\mathscr{L}^2$

To describe the geometry of law-invariance, we define:

Definition 2.34. The **symmetry group** on $\mathscr{L}^2$ is

$$\mathbb{G} = \{(\circ f) | f \text{ is a measure-preserving bijection on } S\} \tag{2.16}$$

Then $\mathbb{G}$ is a group that acts on $\mathscr{L}^2$ on the right. This group action preserves distribution, that is,

$$\forall X \in \mathscr{L}^2, (\circ f) \in \mathbb{G}, X \circ f \text{ and } X \text{ have the same distribution.}$$

In particular, $\mathbb{G}$-action preserves:

- law invariant functionals on $\mathscr{L}^2$, such as the inner product and $\mathbb{E}$;
- certain elements and subsets of $\mathscr{L}^2$, such as $\mathbb{1}$, $\mathscr{L}^2_+$, $\mathscr{E}_{=1}$, and $\mathscr{D}$.

If we think of $\mathscr{L}^2$ as a linear subspace of $\mathbb{R}^S$, then each element of $\mathbb{G}$ acts on $\mathscr{L}^2$ by a permutation of the coordinates.

2.3.2 Geometry of risk envelopes

There is a bijection between certain risk measures and risk envelopes [HL01, Section C.3]:

$$\{\mathcal{F} \mid \mathcal{F} : \mathscr{L}^2 \to (-\infty, \infty],\ \text{sublinear and closed}\} \tag{2.17}$$

$$\leftrightarrow \{\mathscr{F} \mid \mathscr{F} \subseteq \mathscr{L}^2,\ \text{nonempty, closed, and convex}\} \tag{2.18}$$

with the bijection given explicitly by

$$\mathcal{F}(X) = \sup_{Q \in \mathscr{F}} \langle X, Q \rangle \quad \mathscr{F} = \{Q \in \mathscr{L}^2 : \langle X, Q \rangle \leq \mathcal{F}(X), \forall X \in \mathscr{L}^2\}. \tag{2.19}$$

Remark 2.35. The sublinear and closed risk measures can be partially ordered by

$$\mathcal{F} \geq \mathcal{G} \quad \text{iff} \quad \forall X \in \mathscr{L}^2, \mathcal{F}(X) \geq \mathcal{G}(X) \tag{2.20}$$

In terms of their risk envelopes,

$$\mathcal{F} \geq \mathcal{G} \quad \text{iff} \quad \mathscr{G} \subseteq \mathscr{F} \tag{2.21}$$

With this ordering, the set of all sublinear and closed risk measures becomes a lattice [WM19a, Section 5.4]. Its maximal element is the essential supremum, and its minimal element is the expectation. This explains the thesis title.

The following proposition enumerates the correspondence between analytic properties of risk measure and geometric properties of its risk envelope. It is [WM19a, Proposition 7].

Proposition 2.36. *Suppose $\mathcal{F} : \mathscr{L}^2 \to (-\infty, \infty]$ is a sublinear, closed functional, with envelope representation*

$$\mathcal{F}(X) = \sup_{Q \in \mathscr{F}} \langle X, Q \rangle \quad \mathscr{F} = \{Q \in \mathscr{L}^2 : \langle X, Q \rangle \leq \mathcal{F}(X), \forall X \in \mathscr{L}^2\}$$

Then

(1) $\mathcal{F}$ is monotonic if and only if $\mathscr{F} \subseteq \mathscr{L}^2_+$.

(2) $\mathcal{F}$ is translation invariant if and only if $\forall c \in \mathbb{R}, \mathcal{F}(c) = c$, if and only if $\mathscr{F} \subseteq \mathscr{E}_{=1}$. In particular, if $\mathcal{F}$ is risk averse, then it is translation invariant.

(3) $\mathcal{F}$ is coherent if and only if $\mathscr{F} \subseteq \mathscr{L}^2_+ \cap \mathscr{E}_{=1}$.

(4) $\mathcal{F}$ is risk averse if and only if $\mathbb{1} \in \mathscr{F}$.

(5) $\mathcal{F}$ is strictly risk averse if and only if $\mathscr{F} \subseteq \mathscr{E}_{=1}$, and $\mathbb{1}$ is in the interior of $\mathscr{F}$ relative to $\mathscr{E}_{=1}$.

(6) If $\mathcal{F}$ is monotonic and risk averse, with finite $\mathcal{F}(0)$, then let $\mathscr{R} = \mathscr{F} \cap \mathscr{E}_{=1}$, and $\mathcal{R} = \sigma_{\mathscr{R}}$ be its support function, then $\mathcal{R}$ is closed, risk averse, and coherent. Furthermore,

$$\mathcal{R}(X) = \inf_{c \in \mathbb{R}} \mathcal{F}(X - c) + c \quad \forall X \in \mathscr{L}^2.$$

(7) (a) If $\mathcal{F}$ is law invariant, then $\mathscr{F}$ is invariant under the action of $\mathbb{G}$.

(b) If S is finite with uniform distribution, then the converse is true.

Proof. (1)

$$\begin{aligned} & X \le Y \Rightarrow \mathcal{F}(X) \le \mathcal{F}(Y) \\ \Leftrightarrow \quad & X \le Y \Rightarrow \sup_{Q\in\mathscr{F}} \langle X, Q\rangle \le \sup_{Q\in\mathscr{F}} \langle Y, Q\rangle \\ \Leftrightarrow \quad & X \ge 0 \Rightarrow \sup_{Q\in\mathscr{F}} \langle X, Q\rangle \ge 0 \\ \Leftrightarrow \quad & \mathscr{F} \subseteq \mathscr{L}^2_+. \end{aligned}$$

(2) We show that the three conditions imply each other in a circle.

If $\mathcal{F}$ is translation invariant, then since it is also positively homogeneous by assumption,

$$\forall c \in \mathbb{R}, \mathcal{F}(c) = c + \mathcal{F}(0) = c.$$

If

$$\forall c \in \mathbb{R}, \mathcal{F}(c) = c + \mathcal{F}(0) = c,$$

then

$$\begin{aligned} & \mathcal{F}(\pm\mathbb{1}) = \pm 1 \\ \Rightarrow \quad & \sup_{Q\in\mathscr{F}} \mathbb{E}(Q) = 1, \sup_{Q\in\mathscr{F}} \mathbb{E}(-Q) = -1 \\ \Rightarrow \quad & \forall Q \in \mathscr{F}, \mathbb{E}(Q) = 1 \\ \Rightarrow \quad & \mathscr{F} \subseteq \mathscr{E}_{=1}. \end{aligned}$$

If $\mathscr{F} \subseteq \mathscr{E}_{=1}$, then $\forall X \in \mathscr{L}^2_+,\ c \in \mathbb{R}$,

$$\begin{aligned} \mathcal{F}(X + c) &= \sup_{Q\in\mathscr{F}} (\langle X, Q\rangle + c\,\langle \mathbb{1}, Q\rangle) \\ &= \sup_{Q\in\mathscr{F}} (\langle X, Q\rangle + c) \\ &= c + \sup_{Q\in\mathscr{F}} (\langle X, Q\rangle) \\ &= c + \mathcal{F}(X). \end{aligned}$$

So $\mathcal{F}$ is translation invariant.

If $\mathcal{F}$ is strictly risk averse, then by definition of strict risk aversity, $\forall c \in \mathbb{R}, \mathcal{F}(c) = c$, so it is translation invariant.

(3) By parts (1), (2).

(4) If $\mathbb{1} \in \mathscr{F}$, then

$$\mathcal{F}(X) = \sup_{Q \in \mathscr{F}} \langle X, Q \rangle \geq \langle X, \mathbb{1} \rangle = \mathbb{E}(X)$$

Conversely, if $\mathcal{F}$ is risk averse, then by definition of $\mathscr{F}$, $\mathbb{1} \in \mathscr{F}$.

(5) Let

$$\mathscr{F}_0 = \mathscr{F} - \mathbb{1},\ \mathcal{F}_0 = \sigma_{\mathscr{F}_0} = \mathcal{F} - \mathbb{E}$$

then the desired result is equivalent to

$$\begin{aligned} & \forall X \in \mathscr{E}_{=0}, X \not\equiv 0 \Rightarrow \exists Q_X \in \mathscr{E}_{=0}, \langle X, Q_X \rangle > 0 \\ \Leftrightarrow \quad & 0 \text{ is in the interior of } \mathscr{F}_0 \text{ relative to } \mathscr{E}_{=0} \end{aligned}$$

If 0 is in the interior of $\mathscr{F}_0$ relative to $\mathscr{E}_{=0}$, then $\forall X \in \mathscr{E}_{=0}, X \not\equiv 0$, there exists $\theta > 0, \theta X \in \mathscr{F}_0$, such that

$$\mathcal{F}_0(X) \geq \langle X, \theta X \rangle = \theta \|X\|^2 > 0$$

so $\mathcal{F}$ is strictly risk averse. If not, then either $0 \notin \mathscr{F}_0$, in which case $\mathcal{F}$ is not even risk averse, or 0 is in the boundary of $\mathscr{F}_0$ relative to $\mathscr{E}_{=0}$.

Then, by [HL01, Lemma 4.2.1], there exists $X_0 \in \mathscr{E}_{=0}$, $X \not\equiv 0$, and a supporting hyperplane h with normal vector X_0, such that it supports $\mathscr{F}_0$ at 0. But this implies that

$$\begin{aligned} \mathcal{F}(X_0) &= \sup_{Q \in \mathscr{F}} \langle X_0, Q \rangle \\ &= \sup_{Q \in \mathscr{F}_0} \langle X_0, Q \rangle + \mathbb{E}(X_0) \\ &= 0 + 0 = 0 = \mathbb{E}(X_0) \end{aligned}$$

so $\mathcal{F}$ is not strictly averse.

(6) By [HL01, Proposition 2.1.2], since $\mathcal{R}$ is a support function, it is closed, convex, and positive homogeneous.

$\mathcal{F}$ is monotone $\Rightarrow \mathscr{F} \subseteq \mathscr{L}^2_+ \Rightarrow \mathscr{R} \subseteq \mathscr{L}^2_+ \Rightarrow \mathcal{R}$ is monotone.

Thus $\mathscr{R} \subseteq \mathscr{E}_{=1} \Rightarrow \mathcal{R}$ is translation invariant.

By [HL01, Equation 3.3.1],

$$\begin{aligned} \mathcal{R}(X) &= \sigma_{\mathscr{F} \cap \mathscr{E}_{=1}}(X) \\ &= \mathrm{cl}(\sigma_{\mathscr{F}} \uplus \sigma_{\mathscr{E}_{=1}})(X) \end{aligned}$$

where $\uplus$ denotes the infimal convolution, such that

$$\sigma_{\mathscr{F}} \uplus \sigma_{\mathscr{E}_{=1}}(X) = \inf_{Y \in \mathrm{dom}(\sigma_{\mathscr{E}_{=1}})} \left(\sigma_{\mathscr{F}}(X-Y) + \sigma_{\mathscr{E}_{=1}}(Y)\right).$$

$\sigma_{\mathscr{F}} \uplus \sigma_{\mathscr{E}_{=1}}$ is finite, since

$$\sigma_{\mathscr{F}} \uplus \sigma_{\mathscr{E}_{=1}}(X) \leq \sigma_{\mathscr{F}}(0) + \sigma_{\mathscr{E}_{=1}}(X) = \sigma_{\mathscr{F}}(0) + \mathbb{E}(X) < \infty.$$

Since $\mathrm{dom}(\sigma_{\mathscr{E}_{=1}}) = \{c\mathbb{1} : c \in \mathbb{R}\}$, and $\sigma_{\mathscr{E}_{=1}}(c\mathbb{1}) = c$, we have

$$\sigma_{\mathscr{F}} \uplus \sigma_{\mathscr{E}_{=1}}(X) = \inf_{c \in \mathbb{R}} (\mathcal{F}(X-c) + c).$$

By [HL01, Proposition 2.3.2], $\sigma_{\mathscr{F}} \uplus \sigma_{\mathscr{E}_{=1}}$ is convex. Since a convex and finite function is continuous, it is closed, so it equals its closure, and we obtain the result.

(7) (a) If $\mathcal{F}$ is law invariant, then for all $(\circ f) \in \mathbb{G}$, since $X \circ f \overset{\mathrm{d}}{=} X$, we have $\mathcal{F}(X \circ f) = \mathcal{F}(X)$, and by definition of $\mathscr{F}$,

$$\mathcal{F}(X \circ f) = \sigma_{\mathscr{F} \circ f^{-1}}(X)$$

so $\mathscr{F} \circ f^{-1} = \mathscr{F}$. So $\mathscr{F}$ is invariant under the action of $\mathbb{G}$.

(b) If S is finite with uniform distribution, then any $X, Y \in \mathscr{L}^2_+$ with the same distribution, there exists permutation f on S such that $X \circ f = Y$, and so

$$\mathcal{F}(X \circ f) = \mathcal{F}(Y) = \mathcal{F}(X).$$

□

From this, we obtain another dual representation of coherent risk measures:

Proposition 2.37. *For any closed, coherent risk measure $\mathcal{F}$ on $\mathscr{L}^2$, there exists a family Θ of probability measures on $(S, \mathcal{B})$, such that for any $X \in \mathscr{L}^2$,*

$$\mathcal{F}(X) = \sup_{\theta \in \Theta} \mathbb{E}_\theta(X) \tag{2.22}$$

Proof. Take $\mathscr{Q}$, the envelope representation of $\mathcal{F}$. By Proposition 2.36,

$$\mathscr{Q} \subseteq \mathscr{L}^2_+ \cap \mathscr{E}_{=1}$$

so for each $Q \in \mathscr{Q}$, we can define a new probability measure on $(S, \mathcal{Q})$ by

$$\mu_Q(A) = \int_{x \in A} Q(x) d\mu(x)$$

To see that μ_Q is a probability measure, note that:

- μ_Q is nonnegative, since $\mathscr{Q} \subseteq \mathscr{L}^2_+$.
- $\mu_Q = \int_{x \in S} Q(x) d\mu(x) = \mathbb{E}_\mu(Q) = 1$, since $\mathscr{Q} \subseteq \mathscr{E}_{=1}$.

Then let $\Theta = \{\mu_Q : Q \in \mathscr{Q}\}$. □

2.3.3 Envelope representation of CVaR

It is shown in [WM19a, Section 5.8] that, if we let

$$\mathscr{C}_\alpha = \begin{cases} \frac{1}{\bar{\alpha}} \mathscr{U}_\infty \cap \mathscr{D}, \text{ when } 0 \leq \alpha < 1 \\ \mathscr{D}, \text{ when } \alpha = 1 \end{cases} \tag{2.23}$$

then $\mathrm{CVaR}_\alpha = \sigma_{\mathscr{C}_\alpha}$. Thus, the properties of CVaRcan be studied by studying the geometry of $\mathscr{C}_\alpha$.

By algebraic manipulation, we have

$$\mathscr{C}_\alpha - \mathbb{1} = \left(-\frac{\alpha}{\bar{\alpha}}(\mathscr{D} - \mathbb{1})\right) \cap (\mathscr{D} - \mathbb{1}) \tag{2.24}$$

Thus, $\mathscr{C}_\alpha$ is the intersection of $\mathscr{D}$ with a homothetic copy of itself, with homothety center $\mathbb{1}$.

Given any risk measure $\mathcal{F}$ with envelope $\mathscr{F}$ and Kusuoka representation

$$\mathcal{F} = \sup_{\theta \in \Theta} \int_0^1 \mathrm{CVaR}_\alpha \, dm_\theta(\alpha),$$

we use [HL01, Table 3.3.1] to obtain the corresponding Kusuoka representation of its envelope:

$$\mathscr{F} = \mathrm{cl}\,\mathrm{co} \bigcup_{\theta \in \Theta} \int_0^1 \mathscr{C}_\alpha dm_\theta(\alpha). \tag{2.25}$$

2.4 Finite dimensional Kusuoka representation

In Proposition 2.32, it is stated that when the probability space S is atomless, any law invariant CRM over $\mathscr{L}^2$ has a Kusuoka representation. This does not hold in general when S has atoms.

In Section 2.4.1, we geometrically prove that when $S = [n]$ with uniform probability measure, then there exists Kusuoka representation for any closed, law invariant CRM.

In Section 2.4.2, we show that in a nonuniform atomic probability space, there exist closed, law invariant, strictly risk averse, coherent risk measures that can be arbitrarily close to $\mathbb{E}$, and yet has no Kusuoka representation.

Note that the original Kusuoka representation theorem does not cover this case, as $S = [n]$ is the opposite of atomless.

2.4.1 The case of uniform probability on S

Take a closed, law invariant coherent risk measure $\mathcal{F}$, and let its envelope representation be $\mathscr{F}$.

Let $S = [n] = \{1, 2, \cdots, n\}$, where n is an integer at least 2. ν is uniform on S. $\mathscr{D}$ is an n-simplex, with vertices

$$X_1 = (n, 0, \cdots, 0), \cdots, X_n = (0, \cdots, 0, n)$$

and center $\mathbb{1} = (1, \cdots, 1, 1)$. Define the centers of sub-simplices of $\mathscr{D}$ as:

$$C_i = \frac{1}{n-i}(X_{i+1} + \cdots + X_n), \text{ where } i = 0, 1, \cdots n-1$$

The case of $n = 3$ is sufficiently illustrative. As such, it will be used in drawing the figures in this section. In such case, $\mathscr{L}^2$ is just $\mathbb{R}^3$, and $\mathscr{D}$ is a triangle, with vertices $X_1 = (3, 0, 0), X_2 = (0, 3, 0), X_3 = (0, 0, 3)$, and center $\mathbb{1} = (1, 1, 1)$.

Since ν is uniform on S, the possible measure-preserving bijections on S are all permutations π of $[n]$, and its symmetry group $\mathbb{G}$ is isomorphic to the symmetric group on n elements, with $n!$ elements. The action of $\mathbb{G}$ is to permute the coordinates of $\mathbb{R}^n$, and for any $X = (x_1, \cdots x_n) \in \mathbb{R}^n$, the orbit of X under the action of $\mathbb{G}$ is

$$\mathbb{G}(x) = \{(x_{\pi(1)}, \cdots x_{\pi(n)}) : \pi \text{ permutes } [n]\}.$$

For a generic $X \in \mathbb{R}^n$, its orbit has $n!$ elements, but if some components of X are equal, its orbit would have fewer elements. In particular, the orbit of $\mathbb{1}$ has only one element. Out of this orbit of X, a unique element in it has its components arranged in nondecreasing order. That is, $\exists \pi \in \mathbb{G}$, such that $X \circ \pi = (x_{\pi(1)}, \cdots x_{\pi(n)})$, and $x_{\pi(1)} \leq \cdots \leq x_{\pi(n)}$.

Thus, we define the fundamental domain $\mathscr{D}_0$ of $\mathscr{D}$ under the symmetry group $\mathbb{G}$:

$$\mathscr{D}_0 := \{X \in \mathscr{D} : X = (x_1, \cdots, x_n), x_1 \leq \cdots \leq x_n\} \tag{2.26}$$

$\mathscr{D}_0$ is the convex hull of its n vertices, which are

$$\{C_0, C_1, ..., C_{n-1}\}. \tag{2.27}$$

This is illustrated in Figure 2.1.

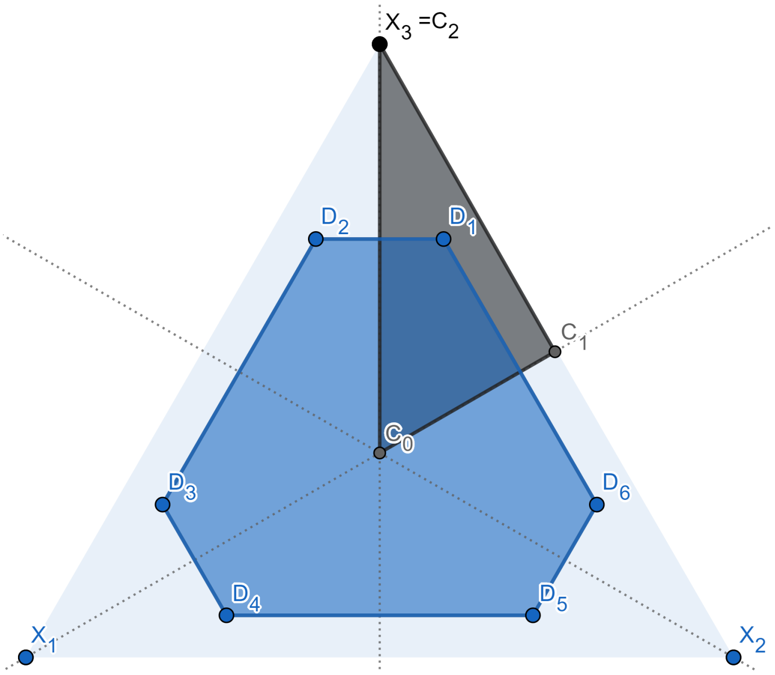


Figure 2.1: The triangle $\mathscr{D}$, when $n = 3$. Its three reflection symmetry axes are marked by dashed lines. Its three vertices are $X_1 = (3, 0, 0), X_2 = (0, 3, 0), X_3 = (0, 0, 3)$. Its center is $C_0 = \mathbb{1} = (1, 1, 1)$. Its fundamental domain is the dark triangle $C_0C_1C_2$. The hexagon $D_1D_2D_3D_4D_5D_6$ is generated by the point D_1 in the fundamental domain.

$\mathscr{D}_0$ can be regarded as a set of representatives from $\mathscr{D}/\mathbb{G}$, and so, for any $X \in \mathscr{D}$, there exists a unique element from its orbit that is in $\mathscr{D}_0$. For any $\mathscr{F} \subseteq \mathscr{D}$, if $\mathscr{F}$ is invariant under the actions of $\mathbb{G}$, that is, $\mathcal{F}$ is law invariant, then it is determined by $\mathscr{F} \cap \mathscr{D}_0$, since it can be reconstructed by

$$\mathscr{F} = \bigcup_{\pi \in \mathbb{G}} (\mathscr{F} \cap \mathscr{D}_0) \circ \pi.$$

Let $\Theta = \mathscr{F} \cap \mathscr{D}_0$. For each $p \in \Theta$, let $\mathscr{P}_p$ be the convex hull of $\mathbb{G}(p)$, which is shown in Figure 2.1 as a hexagon. Then it is geometrically clear that

$$\mathscr{F} = \bigcup_{p \in \Theta} \mathscr{P}_p = \operatorname{cl} \operatorname{co} \bigcup_{p \in \Theta} \mathscr{P}_p. \tag{2.28}$$

So by [HL01, Table 3.3.1], since $\mathcal{F}$ is the support function of $\cup_{p \in \Theta} \mathscr{P}_p$, which is itself convex and closed, we have

$$\mathcal{F} = \sup_{p \in \Theta} \sigma_{\mathscr{P}_p}. \tag{2.29}$$

Finally, since each $p \in \Theta$ is in $\mathscr{D}_0$, it is a convex sum of the vertices

$$\{C_0, C_1, ..., C_{n-1}\} = \{q_0, q_1, ...q_{n-1}\}.$$

Thus, there exists a tuple $\theta_p = (\theta_{p,0}, ..., \theta_{p,n-1})$, such that each $\theta_{p,i} \geq 0$, $\sum_{i=0}^{n-1} \theta_{p,i} = 1$, and

$$p = \sum_{i=0}^{n-1} \theta_{p,i} C_i,$$

which implies

$$\mathscr{P}_p = \operatorname{co}(\mathbb{G}(p)) = \sum_{i=0}^{n-1} \theta_{p,i} \operatorname{co}(\mathbb{G}(q_i)),$$

where the second equality is illustrated in Figure 2.2.

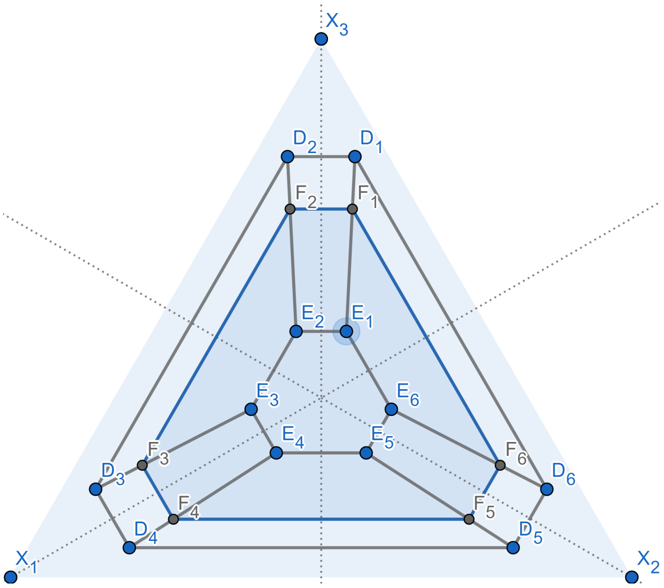


Figure 2.2: The triangle $\mathscr{D}$, with three hexagons generated by the points D_1, E_1, F_1 in the fundamental domain. $F_1 = \theta D_1 + (1-\theta)E_1$. We have that $\theta\mathbb{G}(D_1) + (1-\theta)\mathbb{G}(E_1) = \mathbb{G}(\theta D_1 + (1-\theta)E_1)$. In other words, the following two operations commute: interpolating between points in the fundamental domain, and generating a hexagon from a point in the fundamental domain.

Then, since $\operatorname{co}(\mathbb{G}(q_i)) = \mathscr{C}_{\frac{i}{n}}$, as illustrated in Figure 2.3, we have

$$\mathscr{F} = \bigcup_{p\in\Theta} \mathscr{P}_p = \bigcup_{p\in\Theta} \left(\sum_{i=0}^{n-1} \theta_{p,i} \mathscr{C}_{\frac{i}{n}} \right).$$

Taking the support function on both sides, by [HL01, Table 3.3.1] again,

$$\mathcal{F} = \sup_{p\in\Theta} \sum_{i=0}^{n-1} \theta_{p,i} \operatorname{CVaR}_{\frac{i}{n}}.$$

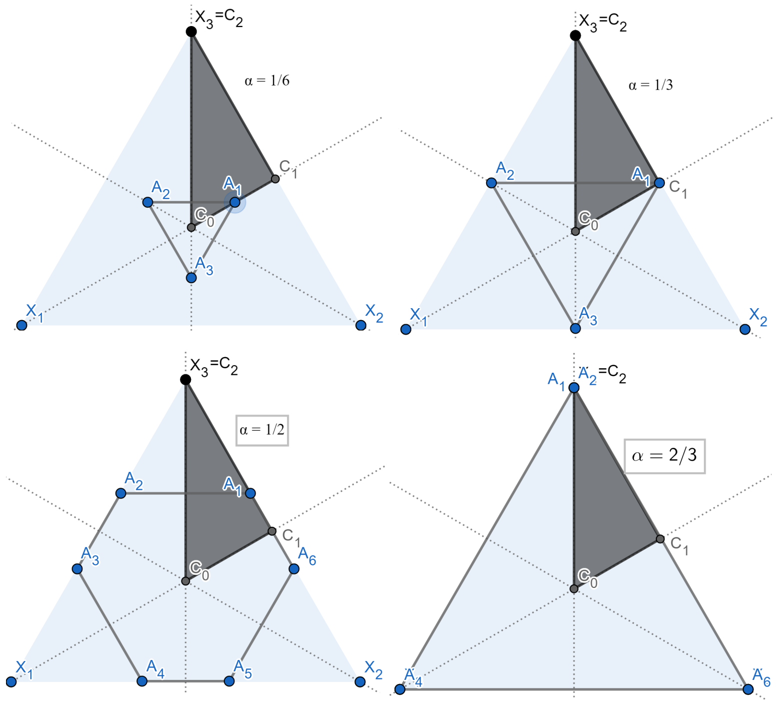


Figure 2.3: $\mathscr{C}_\alpha$ is generated by A_1, which moves in the fundamental domain as α increases. When $\alpha \in [\frac{k}{n}, \frac{k+1}{n}]$, A_1 is on the segment $C_k C_{k+1}$.

We summarize the result in a proposition:

Proposition 2.38. *If $S = [n]$, $n \geq 2$, and ν is uniform on S, then any closed, sublinear, translation invariant, law invariant risk measure $\mathcal{F}$ is coherent.*

Let its risk envelope be $\mathscr{F}$. Let $\mathscr{D}_0$ be a fundamental domain of $\mathscr{D}$, defined as in Equation 2.26.

Define $\Theta = \mathscr{D}_0 \cap \partial\mathscr{F}$, then for each $p \in \Theta$, let $\mathscr{P}_p$ be the convex hull of $\mathbb{G}(p)$, then there exists a tuple $\theta_p = (\theta_{p,0}, ..., \theta_{p,n-1})$, such that each $\theta_{p,i} \geq 0$, $\sum_{i=0}^{n-1} \theta_{p,i} = 1$, and

$$\mathscr{P}_p = \sum_{i=0}^{n-1} \theta_{p,i} \mathscr{C}_{\frac{i}{n}}, \tag{2.30}$$

and we have the Kusuoka representation

$$\mathcal{F} = \sup_{p \in \Theta} \sum_{i=0}^{n-1} \theta_{p,i} \, \mathrm{CVaR}_{\frac{i}{n}} . \tag{2.31}$$

2.4.2 The case of nonuniform probability on S

Let $S = [n] = \{1, 2, \cdots, n\}$, where n is an integer at least 2. Let ν be a nonuniform probability measure on S.

For this section, we do a specific example, which is a sufficiently illustrative example for the general $n \geq 2$ cases.

Let $n = 4$; let ν be defined by

$$\nu : 1 \mapsto \frac{1}{6}, 2 \mapsto \frac{1}{6}, 3 \mapsto \frac{1}{3}, 4 \mapsto \frac{1}{3}$$

Thus, its symmetry group $\mathbb{G}$ has 4 elements, and breaks S into two orbits: $\{1, 2\}$ and $\{3, 4\}$.

Let the centroids of the orbits be $A = \frac{X_1+X_2}{2}, B = \frac{X_3+X_4}{2}$, and let $C = \mathbb{1}$.

Take any closed convex $\mathscr{F} \subseteq \mathscr{D}$ that contains C in its interior (relative to $\mathscr{D}$). Connect A, B, making a line passing C and intersecting $\partial\mathscr{F}$ at two points D, E. This is shown in Figure 2.4.

Define the **line ratio** of $\mathscr{F}$ be

$$LR(\mathscr{F}) = \overline{DC} : \overline{CE},$$

and let $r_0 = LR(\mathscr{D}) = \overline{AC} : \overline{CB}$. Without loss of generality, assume that $\overline{AC} \geq \overline{CB}$, so that $r_0 \geq 1$.

In our particular example, it happens that $\overline{AC} = \sqrt{10}, \overline{CB} = \sqrt{10}/2$, so $r_0 = 2$.

For small $\alpha > 0$, $\mathscr{C}_\alpha$ is an inverted homothetic image of $\mathscr{D}$, so $LR(\mathscr{C}_\alpha) = 1/r_0$. For big $\alpha < 1$, we have $D = A$ and $E = B$, so $LR(\mathscr{C}_\alpha) = r_0$. Between them, we have

$$\forall \alpha \in (0, 1], LR(\mathscr{C}_\alpha) \in [1/r_0, r_0].$$

This is in fact true in general for all Kusuoka-representable sets:

Proposition 2.39 (Kusuoka set line ratio)**.** *Take any nontrivial Kusuoka set, as defined in Equation 2.25:*

$$\mathscr{F} = \operatorname{cl}\operatorname{co} \bigcup_{\theta \in \Theta} \int_0^1 \mathscr{C}_\alpha dm_\theta(\alpha),$$

and assume it is nontrivial, that is, $\mathscr{F} \neq \{\mathbb{1}\}$. Then $LR(\mathscr{F}) \in [1/r_0, r_0]$.

Proof. (Sketch) Refer to Figure 2.4.

Translate the origin of the coordinate frame to point C. So, for example, A in the new coordinate frame is $(2, 2, -1, -1)$.

For each point $P \in \mathbb{R}^n$, let $f(P)$ be its projection onto the line AB, and define $d(P)$ to be the directed distance from C to $f(P)$. So, for example, $d(A) = \sqrt{10}, d(B) = -\sqrt{10}/2$.

Since the affine subspaces spanned by $\{X_1, X_2\}$ and spanned by $\{X_3, X_4\}$ are perpendicular to line AB, the projection of $\mathscr{D}$ onto the line AB is the line segment $[A, B]$.

Then, for any $\alpha \in (0, 1]$,

$$f(\mathscr{C}_\alpha) = [D_\alpha, E_\alpha] = AB \cap \mathscr{C}_\alpha,$$

where we append the subscript α to D, E to denote its association with $\mathscr{C}_\alpha$. Then the line ratio of $\mathscr{C}_\alpha$ is

$$LR(\mathscr{C}_\alpha) = -\frac{d(D_\alpha)}{d(E_\alpha)} \in [1/r_0, r_0]$$

For any probability measure m_θ on $[0, 1]$, let $\mathscr{P}_\theta = \int_0^1 \mathscr{C}_\alpha dm_\theta(\alpha)$. Then $f(\mathscr{P}_\theta) = [D_\theta, E_\theta]$, where

$$d(D_\theta) = \int_0^1 d(D_\alpha) dm_\theta(\alpha), \quad d(E_\theta) = \int_0^1 d(E_\alpha) dm_\theta(\alpha).$$

Thus,

$$LR(\mathscr{P}_\theta) = -\frac{d(D_\theta)}{d(E_\theta)} \in [1/r_0, r_0].$$

Finally, let $\mathscr{F} = \operatorname{cl}\operatorname{co}\bigcup_{\theta\in\Theta} \mathscr{P}_\theta$, we have

$$d(\mathscr{F}) = \left[\inf_{\theta\in\Theta} d(E_\theta), \sup_{\theta\in\Theta} d(D_\theta)\right],$$

and so

$$LR(\mathscr{F}) \in [1/r_0, r_0].$$

□

To violate this line ratio, simply take a closed ball of very small radius r, in $\mathscr{D}$ around $\mathbb{1}$, then shift it by $(1-\epsilon)r$ in the direction of $\frac{X_1+X_2}{2} - \frac{X_3+X_4}{2}$. Let the resulting set be $\mathscr{F}$, then $\mathcal{F} = \sigma_{\mathscr{F}}$ is a closed, law invariant, strictly risk averse, coherent risk measure that can be arbitrarily close to $\mathbb{E}$, and yet has no Kusuoka representation.

Also, note that if we shift the ball by $(1+\epsilon)r$ instead, then we obtain a closed, law invariant, coherent risk measure that is not risk averse, and can be arbitrarily close to $\mathbb{E}$. This contrasts with the case of uniform probability on S, where any law invariant, coherent risk measure is either $\mathbb{E}$ or strictly risk-averse.

We record this formally as:

Proposition 2.40. *If $S = [n]$, $n \geq 2$, and ν is nonuniform on S, then there exists closed, law invariant, risk averse, coherent risk measures that are arbitrarily close to $\mathbb{E}$, and yet have no Kusuoka representations.*

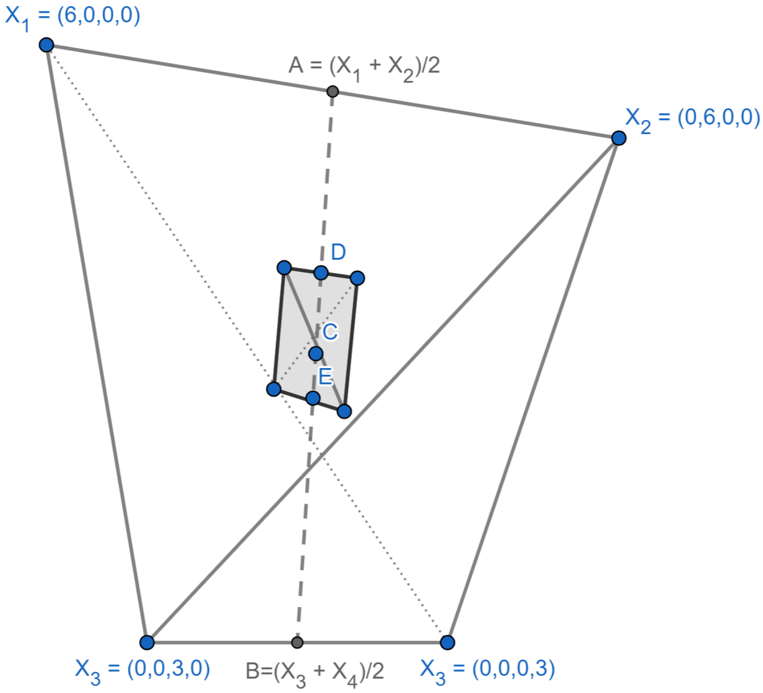


Figure 2.4: When $n = 4$, $\mathscr{D}$ is a tetrahedron, centered at $C = \mathbb{1} = (1, 1, 1, 1)$. With the ν defined in the text, the tetrahedron is not regular. For a small α, $\mathscr{C}_\alpha$ is a small homothetic copy of $\mathscr{D}$, here drawn in the center in gray. The "line ratio" is obtained by connecting $(X_1 + X_2)/2$ and $(X_3 + X_4)/2$, cutting $\mathscr{C}_\alpha$ at D, E, then defining the ratio as $\overline{DC} : \overline{CE}$.

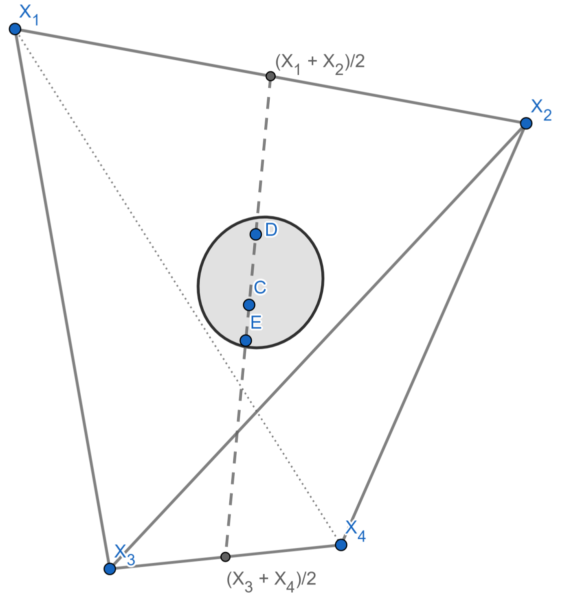


Figure 2.5: To construct a closed, law invariant, strictly risk averse, coherent risk measure that has no Kusuoka representation, take a small closed ball around $\mathbb{1}$, then shift it along the line connecting $(X_1 + X_2)/2$, $(X_3 + X_4)/2$, until its line ratio moves out of the bound.

Chapter 3

Inequalities of coherent risk measures

Inequalities are the foundation of any analytical treatment of a subject. Here, we collect and prove many inequalities for coherent risk measures. Most of them are generalizations to inequalities that involve expectation.

We cannot find in the literature the inequalities that follow, so we believe these are new results. However, in [ZU16, Chapter 3], the authors presented similar probability inequalities for general risk measures.

3.1 Elementary inequalities

Proposition 3.1 (Cauchy–Schwarz inequality)**.** *For any sublinear, monotonic functional $\mathcal{F}$ on $\mathscr{L}^2$, and any $X, Y \in \mathscr{L}^2$, if $\mathcal{F}(XY) \leq 0$, then*

$$\mathcal{F}(XY)^2 \leq \mathcal{F}(X^2)\mathcal{F}(Y^2) \tag{3.1}$$

Similarly, if $\mathcal{F}(-XY) \leq 0$, then

$$\mathcal{F}(-XY)^2 \leq \mathcal{F}(X^2)\mathcal{F}(Y^2) \tag{3.2}$$

Proof. For the first case, let $\lambda = \mathcal{F}(XY)/\mathcal{F}(Y^2)$, and expand the right side of

$$0 \leq \mathcal{F}((X - \lambda Y)^2)$$

.

The second case is converted to the first case by changing the sign of X. □

Jensen's inequality does not generalize easily. [CHK13, Theorem 3] shows that any coherent risk measures that is a g-expectation obeys Jensen's inequality, but since we have no use for g-expectation in this thesis, this result will be skipped.

Two inequalities for expectations extend trivially via the dual representation of coherent risk measures (Proposition 2.37).

Proposition 3.2 (Hölder inequality)**.** *For any closed, coherent risk measure $\mathcal{F}$ on $\mathscr{L}^2$, and any $p, q > 0$ such that $\frac{1}{p} + \frac{1}{q} = 1$, for any $X, Y \in \mathscr{L}^2$ with finite $\mathbb{E}(|X|^p), \mathbb{E}(|Y|^q)$, then*

$$\mathcal{F}(|XY|) \leq \mathcal{F}(|X|^p)^{1/p} \mathcal{F}(|Y|^q)^{1/q}. \tag{3.3}$$

Proof. As in Equation 2.22, take the dual representation of $\mathcal{F}$:

$$\mathcal{F}(Z) = \sup_{\theta \in \Theta} \mathbb{E}_\theta(Z)$$

Then

$$\begin{aligned}
\mathcal{F}(|XY|) &= \sup_{\theta \in \Theta} \mathbb{E}_\theta(|XY|) \\
&\leq \sup_{\theta \in \Theta} \left(\mathbb{E}_\theta(|X|^p)^{1/p} \mathbb{E}_\theta(|Y|^q)^{1/q}\right) \qquad \text{(Hölder inequality)} \\
&\leq \left(\sup_{\theta \in \Theta} \mathbb{E}_\theta(|X|^p)^{1/p}\right) \left(\sup_{\theta \in \Theta} \mathbb{E}_\theta(|Y|^q)^{1/q}\right) \\
&= \left(\sup_{\theta \in \Theta} \mathbb{E}_\theta(|X|^p)\right)^{1/p} \left(\sup_{\theta \in \Theta} \mathbb{E}_\theta(|Y|^q)\right)^{1/q} \\
&= \mathcal{F}(|X|^p)^{1/p} \mathcal{F}(|Y|^q)^{1/q}.
\end{aligned}$$

□

Proposition 3.3 (Minkowski inequality)**.** *Under the same assumptions as above,*

$$\mathcal{F}(|XY|) \leq \mathcal{F}(|X|^p)^{1/p} + \mathcal{F}(|Y|^q)^{1/q}. \tag{3.4}$$

Proof. Essentially the same as the previous proof. □

3.2 Concentration inequalities

A "nice" random variable should be "far" to its expectation with a small probability, that is, its value should be concentrated around its expectation. Concentration inequalities quantify this vague statement in various ways.

In this section, we prove generalizations of Markov, Chebyshev, Chernoff, Hoeffding, Bennett, and Bernstein's inequalities. These inequalities are the most important tail inequalities, heavily used in all parts of statistics, probability, and machine learning. As one example, the Hoeffding inequality is the basis of the popular Monte Carlo tree search algorithm UCT[KS06].

Proposition 3.4 (Markov's inequality)**.** *Given any monotonically increasing $\phi : [0, \infty) \to [0, \infty)$, any $a \geq 0, X \in \mathscr{L}$, and monotonic, positively homogenous functional $\mathcal{F}$,*

$$\mathcal{F}(\phi(|X|)) \geq \phi(a)\mathcal{F}(1_{|X|\geq a}) \tag{3.5}$$

Similarly, if $\phi : \mathbb{R} \to [0, \infty)$ is monotonically increasing, then for any a,

$$\mathcal{F}(\phi(X)) \geq \phi(a)\mathcal{F}(1_{X\geq a}) \tag{3.6}$$

Proof. For the first case, since $\phi(|X|) \geq \phi(a)1_{|X|\geq a}$, by monotonicity of ϕ and positive homogeneity of $\mathcal{F}$,

$$\mathcal{F}(\phi(|X|)) \geq \mathcal{F}(\phi(a)1_{|X|\geq a}) = \phi(a)\mathcal{F}(1_{|X|\geq a}).$$

Similarly for the second case. □

Proposition 3.5 (Chebyshev's inequality)**.** *Given any $a > 0, X \in \mathscr{L}$, and monotonic, positively homogenous functional $\mathcal{F}$,*

$$\mathcal{F}(1_{|X|\geq a}) \leq \frac{1}{a^2}\mathcal{F}(X^2) \tag{3.7}$$

Proof. Let $\phi(a) = a^2$ in the first Markov's inequality. □

Proposition 3.6 (Chernoff bound)**.** *Given any $t > 0, X \in \mathscr{L}$, and monotonic, positively homogenous functional $\mathcal{F}$,*

$$\ln \mathcal{F}(1_{X\geq t}) \leq -\sup_{a\geq 0}(ta - \ln \mathcal{F}(e^{aX})) \tag{3.8}$$

Proof. Let $\phi(x) = e^{tx}$ in the second Markov's inequality, to get

$$\mathcal{F}(1_{X\geq t}) \leq e^{-ta}\mathcal{F}(e^{tX}) \quad \text{for any } a > 0.$$

Then take logarithm on both sides, and take the infimum over $a > 0$ on the right side.

Note that the $a = 0$ case gives $\ln \mathcal{F}(1_{X\geq t}) \leq 0$, which is trivially true, since

$$\mathcal{F}(1_{X\geq t}) \leq \mathcal{F}(1) = 1,$$

so it can be incorporated safely into the infimum. □

We record a particularly useful case for CVaR:

Corollary 3.7 (Chernoff bound for CVaR)**.** *Given any real random variable X with finite first moment, any $\alpha \in [0,1]$, and any $a > \mathrm{VaR}_\alpha(X)$, we have*

$$Pr(X \geq a) \leq \bar{\alpha} \inf_{t>0} \mathrm{CVaR}_\alpha(e^{tX})e^{-ta} \tag{3.9}$$

Proof. In Chebyshev's inequality, replace $\mathcal{F}, a, X$ by $\mathrm{CVaR}_\alpha, 1, e^{t(X-a)/2}$, to obtain

$$\mathrm{CVaR}_\alpha(1_{X\geq a}) = \mathrm{CVaR}_\alpha(1_{e^{t(X-a)/2}\geq 1}) \leq \mathrm{CVaR}_\alpha(e^{t(X-a)}) = \mathrm{CVaR}_\alpha(e^{tX})e^{-ta}$$

for any $t > 0$.

Since $a > \mathrm{VaR}_\alpha(X)$, we have $Pr(X \geq a) \leq \bar{\alpha}$, and so

$$\mathrm{CVaR}_\alpha(1_{X\geq a}) = \min(1, Pr(X \geq a)/\bar{\alpha}) = Pr(X \geq a)/\bar{\alpha}.$$

Optimize over t to obtain the result. □

Proposition 3.8 (Hoeffding's lemma)**.** *For any coherent risk measure $\mathcal{F}$ on $\mathscr{L}$, and X such that $Pr(a \leq X \leq b) = 1$, let*

$$c(t) = \frac{\mathcal{F}(X) + \mathcal{F}(-X)}{e^{t(b-a)} - 1},$$

then for any $t \in \mathbb{R}, t \neq 0$,

$$\mathcal{F}(\exp[t(X - \mathcal{F}(X) - c(t))]) \leq \exp\left(\frac{1}{8}(b-a)^2t^2\right) \tag{3.10}$$

Proof. Rewrite X as a convex sum of a, b:

$$X = \theta b + (1-\theta)a.$$

By convexity of $x \mapsto e^{tx}$,

$$e^{tX} \leq \theta e^{tb} + (1-\theta)e^{ta} = \frac{1}{b-a}((X-a)e^{tb} + (b-X)e^{ta})$$

Then by the coherence of $\mathcal{F}$,

$$\mathcal{F}(e^{tX}) \leq \frac{e^{tb}}{b-a}(\mathcal{F}(X) - a) + \frac{e^{ta}}{b-a}(\mathcal{F}(-X) + b)$$

Replace X by $X - \mathcal{F}(X) - c(t)$ in the above equation, and simplify, using translation invariance of $\mathcal{F}$, to cancel out the $\mathcal{F}$ terms on the right. This yields

$$\mathcal{F}(\exp[t(X - \mathcal{F}(X) - c(t))]) \leq \frac{be^{ta} - ae^{tb}}{b-a} = e^{g(u)}$$

where

$$\begin{cases} u = t(b-a) \\ g(u) = -\gamma u + \ln(1-\gamma+e^u\gamma) \\ \gamma = -\frac{a}{b-a} \end{cases}$$

Since

$$\begin{cases} g(0) = 0 \\ g'(0) = 0 \\ g''(u) \leq \frac{1}{4} \quad \text{for } u \geq 0 \end{cases}$$

we have

$$g(u) \leq \frac{1}{8}u^2 = \frac{1}{8}t^2(b-a)^2$$

yielding the result. □

Hoeffding's lemma is often used to prove Hoeffding's inequality. However, the proof does not generalize immediately to general CRM, or even CVaR, because it depends on that for any two independent $X, Y \in \mathscr{L}^1$, $\mathbb{E}(XY) \leq \mathbb{E}(X)\mathbb{E}(Y)$, which is false for CVaR in general:

Example 3.9. Given independent X, Y, it is not the case that $\mathcal{F}(XY) = \mathcal{F}(X)\mathcal{F}(Y)$. In fact, even for CVaR_α, it is possible for one side to be arbitrarily greater than the other.

Let X, Y both have the distribution on $\{-T, T\}$ defined by

$$Pr(X = -T) = \frac{3}{4}, \quad Pr(X = T) = \frac{1}{4}$$

T being a positive constant, then

$$\mathrm{CVaR}_{1/2}(X) = \mathrm{CVaR}_{1/2}(Y) = 0$$

$$\mathrm{CVaR}_{1/2}(XY) = T^2$$

so $\mathrm{CVaR}_{1/2}(XY)$ can be arbitrarily greater than $\mathrm{CVaR}_{1/2}(X)\,\mathrm{CVaR}_{1/2}(Y)$.

However, when $X, Y \geq 0$ almost surely, we do have:

Proposition 3.10. *For any two independent $X, Y \in \mathscr{L}^1$ that are almost surely non-negative, and any $\alpha \in [0,1]$,*

$$\mathrm{CVaR}_\alpha(XY) \leq \mathrm{CVaR}_\alpha(X)\,\mathrm{CVaR}_\alpha(Y). \tag{3.11}$$

Proof. By continuity of CVaR_α with respect to α, it suffices to prove this for all rational $\alpha \in (0,1)$.

By closedness of CVaR_α, it suffices to prove this for all X, Y in a dense subset of $\mathscr{L}^1$.

For any $\alpha = \frac{m}{n}$, with $0 < m < n$ integers, we prove this proposition for the dense subset of $\mathscr{L}^1$:

$$\left\{ X \in \mathscr{L}^1 : X \text{ has distribution } \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}, n \text{ divides } N, \text{ and all } x_i \geq 0 \right\}.$$

This subset is chosen, because it makes $\mathrm{CVaR}_\alpha(X)$ easy to calculate:

$$\mathrm{CVaR}_\alpha(X) = \frac{1}{\bar{\alpha}N} \sum (\text{the biggest } \bar{\alpha}N \text{ terms of } x_i)$$

for any such X. That is, the average of the largest $\bar{\alpha}N$ terms of the sequence. We write this as $\bar{x}_{\bar{\alpha}N}$

Now, it remains to show that for any two such sequences $(x_i)_{i=1}^N, (y_i)_{i=1}^N$, we have

$$\overline{(xy)}_{\alpha N^2} \leq \bar{x}_{\bar{\alpha}N} \bar{y}_{\bar{\alpha}N}$$

The proof proceeds by modifying the sequences X, Y until they become trivial, preserving the inequality along the way.

Let $0 \leq y_1 \leq \cdots \leq y_N$, so that

$$\bar{y}_{\bar{\alpha}N} = \frac{1}{\bar{\alpha}N} \sum_{i=\alpha N+1}^{N} y_i.$$

Now let $\epsilon = y_N - y_{\alpha N+1}$, and decrease y_N by ϵ. This decreases $\mathrm{CVaR}_\alpha(X)\,\mathrm{CVaR}_\alpha(Y)$ by $\frac{\epsilon}{\bar{\alpha}N}\,\mathrm{CVaR}_\alpha(X)$.

Since it decreases the entries in the list $\{x_i y_j\}_{i,j\in[N]}$ by $\epsilon x_1, ..., \epsilon x_N, 0, 0, ...0$, it decreases $\mathrm{CVaR}_\alpha(XY)$ by at most

$$\frac{\epsilon}{\bar{\alpha}N^2}(x_1 + \cdots + x_N) \leq \frac{\epsilon}{\bar{\alpha}N}\,\mathrm{CVaR}_\alpha(X).$$

Thus, each such reduction decreases $\mathrm{CVaR}_\alpha(X)\,\mathrm{CVaR}_\alpha(Y)$ more than $\mathrm{CVaR}_\alpha(XY)$. So if after such decrease, the inequality holds, the inequality still holds before the decrease.

After doing this decrease for $\bar{\alpha}N$ times on y, then on X, we have the largest $\bar{\alpha}N$ entries of x equal, and the same for Y, and so

$$\mathrm{CVaR}_\alpha(X)\,\mathrm{CVaR}_\alpha(Y) = \max(X)\max(Y) = \max(XY) \geq \mathrm{CVaR}_\alpha(XY).$$

□

Now we are ready to prove generalized Hoeffding, Bennett, and Bernstein's inequalities.

Definition 3.11. The **cumulant generating function** of CVaR_α is

$$K_{\alpha X}(t) = \log \mathrm{CVaR}_\alpha(e^{tX}). \tag{3.12}$$

Proposition 3.12 (Hoeffding's inequality). *Let $X_1, ...X_n$ be a sequence of independent real random variables, with each $X_i \in [a_i, b_i]$ almost surely. Let $\alpha \in [0,1)$, then $\forall t > 0$,*

$$\begin{gathered} Pr\left(\sum_{i=1}^n (X_i - \mathrm{CVaR}_\alpha(X_i)) \geq t\right) \leq \\ \bar{\alpha} \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_1 - a_i)^2} + \sum_{i=1}^n \frac{\mathrm{CVaR}_\alpha(X_i) + \mathrm{CVaR}_\alpha(-X_i)}{b_i - a_i}\right). \end{gathered} \tag{3.13}$$

Proof. Define for all $i \in [n]$,

$$c_i(s) = \frac{\mathrm{CVaR}_\alpha(X_i) + \mathrm{CVaR}_\alpha(-X_i)}{e^{s(b_i - a_i)} - 1}$$

For any t, and any $s > 0$, such that $t + \sum_{i=1}^n (\mathrm{CVaR}_\alpha(X_i) + c_i(s)) > \mathrm{VaR}_\alpha(\sum_{i=1}^n X_i)$, we have by Chernoff bound,

$$\begin{aligned} \frac{1}{\bar{\alpha}} Pr\Bigg(\sum_{i=1}^n X_i \geq t + \sum_{i=1}^n (\mathrm{CVaR}_\alpha(X_i) + c_i(s))\Bigg) \\ &\leq \mathrm{CVaR}_\alpha(e^{s\sum_{i=1}^n X_i}) e^{-st - s\sum_{i=1}^n (\mathrm{CVaR}_\alpha(X_i) + c_i(s))} \\ &= \mathrm{CVaR}_\alpha\left(\prod_{i=1}^n e^{s(X_i - \mathrm{CVaR}_\alpha(X_i) - c_i(s))}\right) e^{-st} \\ &\leq e^{-st} \prod_{i=1}^n \mathrm{CVaR}_\alpha\left(e^{s(X_i - \mathrm{CVaR}_\alpha(X_i) - c_i(s))}\right) \\ &\leq \exp\left(-st + \frac{s^2}{8}\sum_{i=1}^n (b_i - a_i)^2\right) \end{aligned}$$

So for any $t > \mathrm{VaR}_\alpha(\sum_{i=1}^n X_i)$, and any $s > 0$,

$$\frac{1}{\bar{\alpha}} Pr\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-st + s\sum_{i=1}^n \mathrm{CVaR}_\alpha(X_i) + s\sum_{i=1}^n c_i(s) + \frac{1}{8}s^2 \sum_{i=1}^n (b_i - a_i)^2\right)$$

Since for all $s > 0$,

$$sc_i(s) \leq \frac{\mathrm{CVaR}_\alpha(X_i) + \mathrm{CVaR}_\alpha(-X_i)}{b_i - a_i}$$

we have

$$\frac{1}{\bar{\alpha}} Pr\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-st + s\sum_{i=1}^{n} \mathrm{CVaR}_\alpha(X_i) + \frac{1}{8}s^2 \sum_{i=1}^{n}(b_i - a_i)^2 + \right.$$
$$\left. \sum_{i=1}^{n} \frac{\mathrm{CVaR}_\alpha(X_i) + \mathrm{CVaR}_\alpha(-X_i)}{b_i - a_i}\right).$$

So for all $t > \mathrm{VaR}_\alpha(\sum_{i=1}^n X_i)$,

$$Pr\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \bar{\alpha} \inf_{s>0} \exp\left(-st + s\sum_{i=1}^{n} \mathrm{CVaR}_\alpha(X_i) + \frac{1}{8}s^2 \sum_{i=1}^{n}(b_i - a_i)^2 + \right.$$
$$\left. \sum_{i=1}^{n} \frac{\mathrm{CVaR}_\alpha(X_i) + \mathrm{CVaR}_\alpha(-X_i)}{b_i - a_i}\right).$$

When $t \leq \sum_{i=1}^n \mathrm{CVaR}_\alpha(X_i)$, the right side is $\geq \bar{\alpha}$, which makes the inequality useless, as the left side is $\leq \bar{\alpha}$ from $t > \mathrm{VaR}_\alpha(\sum_{i=1}^n X_i)$.

So assume that $t > \sum_{i=1}^n \mathrm{CVaR}_\alpha(X_i)$, which implies $t > \mathrm{VaR}_\alpha(\sum_{i=1}^n X_i)$. Then, taking the minimum on the right side gives

$$Pr\left(\sum_{i=1}^{n} X_i \geq t\right) \leq$$
$$\bar{\alpha} \exp\left(-\frac{2\left(t - \sum_{i=1}^n \mathrm{CVaR}_\alpha(X_i)\right)^2}{\sum_{i=1}^n (b_i - a_i)^2} + \sum_{i=1}^{n} \frac{\mathrm{CVaR}_\alpha(X_i) + \mathrm{CVaR}_\alpha(-X_i)}{b_i - a_i}\right)$$

which is equivalent to what we need to prove. □

Proposition 3.13 (Bennett's inequality). *Let $X_1, ...X_n$ be a sequence of independent real random variables, with each $\mathrm{CVaR}_\alpha(X_i) = 0$, and $X_i \leq a$ almost surely. Let $\alpha \in [0, 1)$, and let*

$$v_\alpha = \sum_{i=1}^{n} \mathrm{CVaR}_\alpha(X_i^2),$$

then $\forall t > 0$,

$$Pr\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \bar{\alpha} \exp\left(-\frac{v_\alpha}{a^2} h\left(\frac{at}{v_\alpha}\right)\right) \tag{3.14}$$

where

$$h(t) = (1 - t)\ln(1 + t) - t \tag{3.15}$$

Proof. Let

$$\psi(x) = e^x - 1 - x$$

so that

$$\psi(x) = \begin{cases} \leq \frac{1}{2}x^2 & \text{when } x \leq 0 \\ \geq \frac{1}{2}x^2 & \text{when } x \geq 0 \end{cases}$$

As in the previous proof, for all $s > 0$, $t > \mathrm{VaR}_\alpha(\sum_{i=1}^n X_i)$,

$$\frac{1}{\bar{\alpha}} Pr\left(\sum_{i=1}^n X_i \geq t\right) \leq e^{-st} \prod_{i=1}^n \mathrm{CVaR}_\alpha\left(e^{sX_i}\right)$$

Then

$$\begin{aligned} \mathrm{CVaR}_\alpha\left(e^{sX_i}\right) &= \mathrm{CVaR}_\alpha\left(1 + sX_i + \psi(sX_i)\right) \\ &\leq 1 + s\,\mathrm{CVaR}_\alpha(X_i) + \mathrm{CVaR}_\alpha(\psi(sX_i)) \\ &= 1 + \mathrm{CVaR}_\alpha(\psi(sX_i)) \\ &\leq \exp\left(\mathrm{CVaR}_\alpha(\psi(sX_i))\right) \end{aligned}$$

So,

$$Pr\left(\sum_{i=1}^n X_i \geq t\right) \leq \bar{\alpha} \exp\left(-st + \sum_{i=1}^n \mathrm{CVaR}_\alpha(\psi(sX_i))\right)$$

Let $X_i^+ = \max(X_i, 0), X_i^- = \max(-X_i, 0)$, so that $X_i = X_i^+ - X_i^-$, and by convexity of ψ,

$$\psi(sX_i) \leq \psi(sX_i^+) + \psi(-sX_i^-)$$

By bounds on ψ,

$$\psi(-sX_i^-) \leq \frac{1}{2}s^2(X_i^-)^2 \leq \frac{(X_i^-)^2}{a^2}\psi(as)$$

For any $s > 0, x \in [0, 1]$,

$$\psi(sx) = \frac{1}{2}s^2x^2 + \frac{1}{6}s^3x^3 + \cdots \leq x^2(\frac{1}{2}s^2 + \frac{1}{6}s^3 + \cdots) = x^2\psi(s)$$

So

$$\psi(sX_i^+) = \psi(as(X_i/a)^+) \leq \frac{(X_i^+)^2}{a^2}\psi(as)$$

So

$$\psi(sX_i) \leq \psi(as)\frac{1}{a^2}\left((X_i^+)^2 + (X_i^-)^2\right) = \frac{\psi(as)}{a^2}X_i^2$$

By monotonicity and positive homogeneity of CVaR_α,

$$\mathrm{CVaR}_\alpha(\psi(sX_i)) \leq \frac{\psi(as)}{a^2}\,\mathrm{CVaR}_\alpha(X_i^2)$$

so

$$Pr\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \bar{\alpha}\exp\left(-st + \frac{v_\alpha \psi(as)}{a^2}\right)$$

for any $s > 0$. Optimizing the right side by $s = \frac{1}{a}\ln(1+at/v_\alpha)$, we get the desired inequality. □

Using the simple inequality

$$\forall t \geq 0, \quad h(t) \geq \frac{t^2}{2+2t/3}$$

we obtain

Corollary 3.14 (Bernstein's inequality)**.** *Let $X_1, ...X_n$ be a sequence of independent real random variables, with each $\mathrm{CVaR}_\alpha(X_i) = 0$, and $X_i \leq a$ almost surely. Let $\alpha \in [0, 1)$, and let*

$$v_\alpha = \sum_{i=1}^{n} \mathrm{CVaR}_\alpha(X_i^2),$$

then $\forall t > 0$,

$$Pr\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \bar{\alpha}\exp\left(-\frac{t^2}{2v_\alpha + 2at/3}\right) \tag{3.16}$$

3.2.1 A conjecture

The conditional CVaRcan be defined in the same way as conditional expectation. So for example, if $X, Y \in \mathscr{L}^1$ are independent, then $\mathrm{CVaR}_\alpha(X|Y) = \mathrm{CVaR}_\alpha(X)$.

The Law of Total Expectation states that

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)) \tag{3.17}$$

The following conjecture was numerically found and verified for several million examples of discrete X, Y:

Conjecture 3.15 (Law of total CVaR)**.** *For any two independent $X, Y \in \mathscr{L}^1$ that are almost surely non-negative, and any $\alpha \in [0, 1)$,*

$$\mathrm{CVaR}_\alpha(X) \leq \frac{1}{\bar{\alpha}}\mathrm{CVaR}_\alpha(\mathrm{CVaR}_\alpha(X|Y)). \tag{3.18}$$

Further, this is sharp in that for any $\epsilon \in (0, 1)$, $\frac{1}{\bar{\alpha}}$ cannot be replaced by $\frac{1}{\bar{\alpha}^{1-\epsilon}}$

In proofs of martingale inequalities of expectation, such as McDiarmid's inequality and Doob's martingale inequality, the law of total expectation is used, so it stood to reason that a law of total CVaRwould be required for a proof of martingale inequalities for CVaR.

Unfortunately, even assuming the conjecture to be true, no valuable generalization of martingale inequalities to CVaRwere discovered.

3.3 Statistical learning theory (SLT)

Learning theory is the theoretical counterpart of practical machine learning, and one of its main branches is statistical learning theory (SLT). For a historical overview up to 1999, see [Vap00, Introduction]. For a detailed introduction to SLT, see [BBL03].

SLT models the problem of learning by positing a **learner** who is trying to find the best **model** out of a **class of models**, given some **training data**.

3.3.1 Overview of SLT

Formally, let $\mathcal{X}$ be a nonempty set called **feature space**, and $\mathcal{Y}$ be another nonempty set called **label space**. Let there be a probability distribution μ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. This specifies the learning problem.

For example, suppose the problem is to predict the height based on biological sex only, the feature space would be $\{\text{male}, \text{female}\}$, and the sample space $[0, \infty)$. The distribution μ is the probability distribution of a random person from the population having the specified sex and height. In particular, $\mu(\text{male}, [150, 160])$ is the probability that a random person would be male and having height between 150 and 160 cm.

The learner is given a sequence of n **training data** sampled from μ, that is, it is given

$$Z^n = (Z_1, \cdots, Z_n) \in \mathcal{Z}^n$$

The learner already knows $\mathcal{X}, \mathcal{Y}$, so it can start with the **hypothesis class** $\mathcal{H}$, a set of **hypotheses** $h : \mathcal{X} \to \mathcal{Y}$.

The problem of the learner, then, is to find the best h_{Z^n}, given training data Z^n.

To formalize "best", a **loss function** $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ is defined, such that $\ell(y, y')$ measures how different y, y' are. Given a hypothesis h, and a new data point (x, y), $\ell(h(x), y)$ measures how badly the hypothesis errs on the data point.

The simplest loss function is:

Definition 3.16 (0-1 loss)**.** The **0-1 loss** is

$$\ell_{01}(y, y') = \begin{cases} 1 & \text{if } y \neq y', \\ 0 & \text{else.} \end{cases} \tag{3.19}$$

More complex loss functions are often used, such as the quadratic loss, widely used in problems where the label space is continuous rather than discrete:

$$\ell_{01}(y, y') = (y - y')^2 \tag{3.20}$$

however, we have not generalized our results to such loss functions.

To judge the overall performance of hypothesis h, the expectation of loss is used:

Definition 3.17 (Expected loss function)**.**

$$\mathrm{Loss}(h, \mu) = \mathbb{E}_{(X,Y)\sim\mu}(\ell(h(X), Y)) \tag{3.21}$$

where the subscript of $\mathbb{E}_{(X,Y)\sim\mu}$ means that (X, Y) are random variables with joint probability measure $\mu_{(X,Y)} = \mu$.

Suppose the learner knows μ, then it could just return the optimal solution (sometimes called the **Bayes classifier**)

$$h^* = \arg\min_{h\in\mathcal{H}} \mathrm{Loss}(h, \mu),$$

with minimal loss (the **Bayes risk**)

$$L^* = \min_{h\in\mathcal{H}} \mathrm{Loss}(h, \mu) = \mathrm{Loss}(h^*, \mu).$$

But almost always, μ is inaccessible, and the learner may only access Z^n, with which it constructs an approximation of μ, with which it selects a good hypothesis h_{Z^n}, such that $(\mathrm{Loss}(h_{Z^n}, \mu) - L^*)$ is small.

However, if Z^n happens to be a very unlucky draw, it would give a bad approximation of μ, from which the learner has no chance of learning well. To deal with such unlucky cases, instead of perfectly reliable learning, the concept of probably approximately correct (PAC) learning * is used:

*As a historical note, PAC-learning is proposed by Leslie Valiant in 1984 [Val84], and Valiant is quite enamored with it, recently proposing that the concept of evolvability, a fundamental notion of biological evolution and a significant ingredient in the explanation of life on earth, is a special case of PAC-learnability [Val09].

Definition 3.18 (PAC-learnability)**.** Given a hypothesis class $\mathcal{H}$ on the spaces $\mathcal{X}, \mathcal{Y}$, the hypothesis class is **PAC-learnable** if there exists a **learner function** $Z^n \mapsto h_{Z^n}$, such that for any probability measure μ over $\mathcal{X} \times \mathcal{Y}$, and any $\epsilon > 0, \delta > 0$, there exists some $N \in \mathbb{N}$, such that for any integer $n > N$,

$$Pr_{(X,Y)\sim\mu}(\text{Loss}(h_{Z^n}, \mu) - L^* < \epsilon) > 1 - \delta \tag{3.22}$$

where $L^* = \min_{h\in\mathcal{H}} \text{Loss}(h, \mu)$.

In other words, given enough samples, the probability of learning a bad hypothesis is low, regardless of the truth μ. This can be reformulated as a convergence in probability:

$$\min_{h\in\mathcal{H}} \text{Loss}(h, \mu_{Z^n}) \overset{\text{Pr}}{\to} \min_{h\in\mathcal{H}} \text{Loss}(h, \mu) \tag{3.23}$$

In other words,

$$\min_{h\in\mathcal{H}} \text{Loss}(h, \mu_{Z^n})$$

is a **consistent estimator** of $\min_{h\in\mathcal{H}} \text{Loss}(h, \mu)$.

Just from the definition, it is unclear whether any nontrivial learning problem is PAC-learnable. However, as it turns out, even the most naive learner, the empirical risk minimizer (ERM), can PAC-learn some nontrivial problems, as will be demonstrated in Theorem 3.26.

Definition 3.19 (Empirical risk minimizer)**.** Given a loss function ℓ and a hypothesis class $\mathcal{H}$, its associated empirical risk minimizer is defined by

$$h_{Z^n} = \arg\min_{h\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i) \tag{3.24}$$

where $Z^n = ((x_1, y_1), ..., (x_n, y_n))$. It is often written as $ERM_{\mathcal{H}}$, with ℓ elided.

In other words, upon receiving the training samples Z^n, it constructs μ_{Z^n} as the empirical approximation of μ, then chooses the hypothesis $h \in \mathcal{H}$ that minimizes $\text{Loss}(h, \mu_{Z^n})$,

Notation 3.20. For any $Z^n \in \mathcal{Z}^n$, μ_{Z^n} is the empirical distribution on $\mathcal{X} \times \mathcal{Y}$ defined by

$$\mu_{Z^n} = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Y_i)} \tag{3.25}$$

So we can more succinctly write Equation 3.24 as

$$h_{Z^n} = \arg\min_{h\in\mathcal{H}} \mathrm{Loss}(h, \mu_{Z^n}) \tag{3.26}$$

For an ERM learner to perform well, it must receive a training sample Z^n that closely represents μ. This can be formalized by the concept of

Definition 3.21 (ϵ-representative sample)**.** A sample Z^n is **ϵ-representative sample** with respect to $\mathcal{H}$ if

$$\sup_{h\in\mathcal{H}} |\,\mathrm{Loss}(h, \mu_{Z^n}) - \mathrm{Loss}(h, \mu)| \le \epsilon \tag{3.27}$$

Certain hypothesis classes can be easily ϵ-represented, while others cannot. For example, with respect to the trivial class $\mathcal{H} = \{h\}$ with only one hypothesis allows any Z^n to well, *trivially* 0-represent μ. In contrast, if $\mathcal{H}$ is big, it will be difficult to represent μ well, since there is likely a $h \in \mathcal{H}$ that fits Z^n very well, but fits μ poorly. This is essentially one manifestation of the **problem of overfitting**.

To formalize this notion of certain hypothesis classes being easier to allow good representation, we define:

Definition 3.22 (Uniform convergence)**.** A hypothesis class $\mathcal{H}$ has uniform convergence property if there exists $n_{\mathcal{H}} : (0,1) \times (0,1) \to \mathbb{N}$, such that for any distribution μ over $\mathcal{X} \times \mathcal{Y}$, for any $\epsilon, \delta \in (0,1)$, and any $n > n_{\mathcal{H}}(\epsilon, \delta)$, we have

$$Pr_{Z^n \sim \mu^n}(Z^n \text{ is } \epsilon\text{-representative}) > 1 - \delta \tag{3.28}$$

To say that $\mathcal{H}$ has uniform convergence is to say that a uniform weak law of large numbers holds. In detail, for any hypothesis $h \in \mathcal{H}$ and any probability distribution μ over $\mathcal{X} \times \mathcal{Y}$, let $(z_n)_n$ be an IID sequence sampled from μ, then

$$\mathrm{Loss}(h, \mu_{Z^n}) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i) \xrightarrow{\Pr} \mathrm{Loss}(h, \mu)$$

and this convergence in probability is uniform over μ and h.

That $\mathcal{H}$ is uniformly convergent is a strictly stronger hypothesis than that $ERM_{\mathcal{H}}$ is a PAC-learner. To see this, note that the ERM learner is a PAC-learner if

$$\min_{h\in\mathcal{H}} \mathrm{Loss}(h, \mu_{Z^n}) \xrightarrow{\Pr} \min_{h\in\mathcal{H}} \mathrm{Loss}(h, \mu)$$

uniformly over μ. This has one less uniformity condition, and thus weaker.

Proposition 3.23. *If a hypothesis class is uniformly convergent, then the $ERM_{\mathcal{H}}$ learner is a PAC-learner.*

Proof. Given any distribution μ, let the Bayes hypothesis be h^* and the Bayes loss be L^*. Now, for any $\epsilon, \delta \in (0, 1)$, and any $n > n_{\mathcal{H}}(\epsilon, \delta)$, let Z^n be the training data sampled out of μ^n, and h_{Z^n} be the hypothesis produced by the $ERM_{\mathcal{H}}$ learner.

With probability above $1 - \delta$, Z^n is ϵ-representative, then we have

$$\mathrm{Loss}(\mu, h_{Z^n}) \leq \mathrm{Loss}(\mu_{Z^n}, h_{Z^n}) + \epsilon \quad \leq \mathrm{Loss}(\mu_{Z^n}, h^*) + \epsilon \leq \mathrm{Loss}(\mu, h^*) + 2\epsilon \quad = L^* + 2\epsilon$$

Thus

$$|\,\mathrm{Loss}(\mu, h_{Z^n}) - L^*| \leq 2\epsilon.$$

Thus

$$Pr_{Z^n \sim \mu^n}(|\,\mathrm{Loss}(\mu, h_{Z^n}) - L^*| \leq 2\epsilon) \geq 1 - \delta.$$

□

In general, any learner must balance between too much overlearning and underlearning. Underlearning occurs when the learner fails to extract sufficient information from its training data, while overlearning occurs when the learner extracts too much from its training data.

Consider a human pupil learning from a textbook on arithmetics. An underlearner might read through the textbook and remain the same as before, while an overlearner might memorize every single word, but fail to do any problem that does not appear in the book.

For the $ERM_{\mathcal{H}}$ learner, its hypothesis class $\mathcal{H}$ is the deciding factor for whether it would underlearn or overlearn. Intuitively, if $\mathcal{H}$ is small, for example, having only size 2, then no matter how rich the training samples are, the learner could only encode one bit of information in its final choice of hypothesis. One might say that the learner has only one bit of memory. This is a severe underlearner. If $\mathcal{H}$ is too big, for example, containing every single function of type $\mathcal{X} \to \mathcal{Y}$, then the $ERM_{\mathcal{H}}$ learner would simply "memorize" the whole training sample, with no regard for generalization outside it. Figure 3.1 depicts these two pathologies of learning.

The key to a good $ERM_{\mathcal{H}}$ learner is a hypothesis class $\mathcal{H}$ with the right level of complexity, [†] not so big as to cause overlearning, but not so small as to cause underlearning.

[†] Other words for this nebulous concept include "capacity", "expressiveness", "richness", "expressive power", and "flexibility".

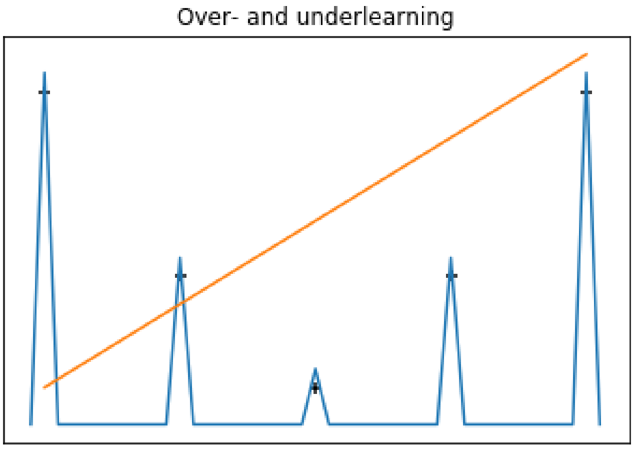


Figure 3.1: Illustration of over- and underlearning. The black crosses are the training samples. The straight line is the hypothesis learned by an underlearner, while the spiky line is the hypothesis learned by an overlearner.

One useful definition of the complexity of a hypothesis class is its Vapnik–Chervonenkis dimension (VCdim), first proposed in 1968 [VC71].

Definition 3.24. Given the feature space $\mathcal{X}$, binary label space $\mathcal{Y} = \{0, 1\}$, and a set of feature points $x_1, ...x_n \in \mathcal{X}$, a hypothesis class $\mathcal{H}$ **shatters** the set of feature points if, for any set of labels $y_1, ..., y_n \in \mathcal{Y}$, there exists a hypothesis $h \in \mathcal{H}$ such that

$$h(x_i) = y_i, \quad \forall i \in [n]$$

Definition 3.25. For any hypothesis class $\mathcal{H}$ on the spaces $\mathcal{X}, \mathcal{Y}$, the **Vapnik–Chervonenkis dimension** of $\mathcal{H}$, written as $VCdim(\mathcal{H})$, is the size of the biggest subset of $\mathcal{X}$ that can be shattered by $\mathcal{H}$. If there is no maximal size, then $VCdim(\mathcal{H}) = \infty$.

3.3.2 The fundamental theorem of SLT

Speculations about the meaning of life aside, the theory of PAC-learnability is very difficult, and the “fundamental theorem” concerns itself with only binary classification problems. That is, the label space $\mathcal{Y} = \{0, 1\}$, and the loss function ℓ is the 0-1 loss.

Under such restrictive conditions, we have [SB14, Theorem 6.7]:

Theorem 3.26. *Given any learning problem $\mathcal{X}, \mathcal{Y} = \{0, 1\}$, with hypothesis class $\mathcal{H}$, and loss function being 0-1 loss, the following conditions are equivalent:*

(1) $\mathcal{H}$ is uniformly convergent.

(2) $ERM_{\mathcal{H}}$ is a PAC-learner for this problem.

(3) This problem is PAC-learnable.

(4) $VCdim(\mathcal{H})$ is finite.

Proof. (1) $\Rightarrow$ (2): Proved in Proposition 3.23.

(2) $\Rightarrow$ (3): Trivial.

(3) $\Rightarrow$ (4): This step uses the *no free lunch theorem.*

(4) $\Rightarrow$ (1): There is a purely combinatorial proof, which we skip. □

Given any finite VC-dimensional hypothesis class, there are explicit bounds $n_{\mathcal{H}}(\epsilon, \delta)$ on how many samples the $ERM_{\mathcal{H}}$ learner need in order to accomplish PAC-learning, which are given quantitatively by combinatorial calculations in the (4) $\Rightarrow$ (1) step. We will not need explicit bounds in the proof.

In order to complete the (3) $\Rightarrow$ (4) step, we present the no free lunch theorem. The proof can be found in [SB14, Theorem 5.1]

Theorem 3.27 (no free lunch)**.** *Given any feature space $\mathcal{X}$, and binary sample space $\{0,1\}$, let the loss function ℓ be 0-1 loss, then for any learner $Z^n \mapsto h_{Z^n}$ and any positive integer $n \leq \frac{1}{2}|\mathcal{X}|$ there exists a probability distribution μ on $\mathcal{X} \times \{0,1\}$, so that there exists some*

$$h^* : \mathcal{X} \to \{0,1\})$$

such that $\mathrm{Loss}(h^*, \mu) = 0$*, and yet*

$$Pr_{Z^n \sim \mu^n}\left(\mathrm{Loss}(h_{Z^n}, \mu) \geq \frac{1}{8}\right) \geq \frac{1}{7} \tag{3.29}$$

Given the no free lunch theorem, we can complete the proof of Theorem 3.26:

Proof of (3) $\Rightarrow$ (4). Suppose (3) and not (4).

By (3), PAC-learning is possible, so there exists some n, such that for any probability distribution μ on $\mathcal{X} \times \{0,1\}$, we have PAC-learning

$$Pr_{Z^n \sim \mu^n}\left(\mathrm{Loss}(h_{Z^n}, \mu) \geq \operatorname*{arg\,min}_{h \in \mathcal{H}} \mathrm{Loss}(h, \mu) + \frac{1}{8}\right) < \frac{1}{7}.$$

By not (4), $VCdim(\mathcal{H}) = \infty$, and so there exists some $\{x_1, ..., x_n\} \in \mathcal{X}$ that is shattered by $\mathcal{H}$, that is, any partial hypothesis $h^* : \mathcal{X} \to \{0,1\}$ can be realized by some h in $\mathcal{H}$.

Then by no free lunch theorem, there exists a distribution μ on $\{x_1, ..., x_n\} \times \{0, 1\}$, such that, after extending μ to the rest of $\mathcal{X} \times \{0, 1\}$, satisfies

$$\arg\min_{h \in \mathcal{H}} \mathrm{Loss}(h, \mu) = 0,$$

$$Pr_{Z^n \sim \mu^n}\left(\mathrm{Loss}(h_{Z^n}, \mu) \geq \frac{1}{8}\right) \geq \frac{1}{7}.$$

This is a contradiction. □

3.3.3 Generalization to CVaR

The fundamental theorem of SLT can be easily generalized by replacing expectations with CVaR, after an appropriate generalization of PAC-learnability to arbitrary risk measures.

Definition 3.28 ($\mathcal{F}$-expected loss function)**.** For any risk measure $\mathcal{F} : \mathscr{L} \to \mathbb{R}$ on real random variables, not necessarily the expectation or CVaR, we define a generalized expected loss function. For any hypothesis h and distribution μ on $\mathcal{X} \times \mathcal{Y}$,

$$\mathrm{Loss}_{\mathcal{F}}(h, \mu) = \mathcal{F}(\ell(h(X), Y)) \tag{3.30}$$

where $\ell(h(X), Y)$ is a random variable with $(X, Y) \sim \mu$.

By replacing Loss with $\mathrm{Loss}_{\mathcal{F}}$ in their definitions, we can generalize PAC-learnability, ϵ-representativeness, uniform convergence, and empirical risk minimization for any $\mathcal{F}$, not just $\mathbb{E}$.

The no free lunch theorem generalizes almost for free, despite its name:

Theorem 3.29 (Generalized no free lunch)**.** *Let $\mathcal{F}$ be any risk averse risk measure, that is, $\mathcal{F} \geq \mathbb{E}$. Given any feature space $\mathcal{X}$, and binary sample space $\{0, 1\}$, let the loss function ℓ be 0-1 loss, then for any learner $Z^n \mapsto h_{Z^n}$ and any positive integer $n \leq \frac{1}{2}|\mathcal{X}|$ there exists a probability distribution μ on $\mathcal{X} \times \{0, 1\}$, so that there exists some*

$$h^* : \mathcal{X} \to \{0, 1\})$$

such that $\mathrm{Loss}_{\mathcal{F}}(h^*, \mu) = 0$*, and yet*

$$Pr_{Z^n \sim \mu^n}\left(\mathrm{Loss}_{\mathcal{F}}(h_{Z^n}, \mu) \geq \frac{1}{8}\right) \geq \frac{1}{7} \tag{3.31}$$

Proof. Since Loss $\leq \mathrm{Loss}_{\mathcal{F}}$, the original no free lunch theorem immediately gives the result. □

Finally, we define a temporary notation, used only in the next theorem:

Notation 3.30. For any law invariant risk measure $\mathcal{F} : \mathscr{L} \to \mathbb{R}$, let $f_{\mathcal{F}} : [0,1] \to \mathbb{R}$ be the function

$$f_{\mathcal{F}}(p) = \mathcal{F}(X)$$

where X is a random variable with

$$Pr(X = 0) = 1 - p, \quad Pr(X = 1) = p$$

Now, the fundamental theorem of SLT can be generalized to:

Theorem 3.31 (Generalized fundamental theorem of SLT)**.** *Given any law invariant risk measure $\mathcal{F}$ with continuous $f_{\mathcal{F}}$, any learning problem $\mathcal{X}, \mathcal{Y} = \{0,1\}$, with hypothesis class $\mathcal{H}$, and loss function being 0-1 loss, the following conditions are equivalent:*

(1) $\mathcal{H}$ is $\mathcal{F}$-uniformly convergent.

(2) $\mathcal{F}$-$ERM_{\mathcal{H}}$ is a $\mathcal{F}$-PAC-learner for this problem.

(3) This problem is $\mathcal{F}$-PAC-learnable.

(4) $VCdim(\mathcal{H})$ is finite.

Proof. (1) $\Rightarrow$ (2): This holds for all $\mathcal{F}$. The proof is the same as in Proposition 3.23.

(2) $\Rightarrow$ (3): This holds for all $\mathcal{F}$. The proof is immediate by definition of ρ-PAC-learnability.

(3) $\Rightarrow$ (4): This holds for all risk averse $\mathcal{F}$, that is, $\mathcal{F} \geq \mathbb{E}$, by the generalized no free lunch theorem.

(4) $\Rightarrow$ (1): By the original fundamental theorem of SLT, $\mathcal{H}$ is uniformly convergent, so it suffices to show it implies $\mathcal{H}$ is $\mathcal{F}$-uniformly convergent.

For any μ, $\text{Loss}_{\mathcal{F}}(h, \mu) = \mathcal{F}(\ell(h(X), Y))$, where $(X, Y) \sim \mu$. Let the binary random variable $\ell(h(X), Y)$ have $Pr(\ell(h(X), Y) = 1) = p$, then $\text{Loss}(\ell(h(X), Y)) = p$, and so

$$\text{Loss}_{\mathcal{F}}(\ell(h(X), Y)) = f_{\mathcal{F}}(p) = f_{\mathcal{F}}(\text{Loss}(\ell(h(X), Y)))$$

$f_{\mathcal{F}}$ is continuous on $[0,1]$, so it is uniformly continuous, so for any $\epsilon > 0$, there exists ϵ', such that any ϵ'-representative training data Z^n is ϵ-$\mathcal{F}$-representative:

$$\forall h \in \mathcal{H}, \ |\text{Loss}(h, \mu_{Z^n}) - \text{Loss}(h, \mu)| \leq \epsilon'$$

$$\Rightarrow \; |\operatorname{Loss}_{\mathcal{F}}(h, \mu_{Z^n}) - \operatorname{Loss}_{\mathcal{F}}(h, \mu)| = |f_{\mathcal{F}}(\operatorname{Loss}(h, \mu Z^n)) - f_{\mathcal{F}}(\operatorname{Loss}(h, \mu))| \leq \epsilon$$

Thus, if $\mathcal{H}$ is uniformly convergent, it has some $n_{\mathcal{H}} : (0,1) \times (0,1) \to \mathbb{N}$ such that $\forall \epsilon', \delta \in (0,1), n \geq n_{\mathcal{H}}(\epsilon', \delta)$,

$$Pr_{Z^n \sim \mu^n}(Z^n \text{ is } \epsilon'\text{-representative}) \geq 1 - \delta$$

$$\Rightarrow Pr_{Z^n \sim \mu^n}(Z^n \text{ is } \epsilon\text{-}\mathcal{F}\text{-representative}) \geq 1 - \delta$$

Thus $\mathcal{H}$ is $\mathcal{F}$-uniformly convergent. □

Moreover, given $\mathcal{F}$ and its function $f_{\mathcal{F}}$, an explicit bound on how many samples the ERM learner need in order to do PAC-learning can be given.

Lemma 3.32. *For any spectral risk measure* $\mathcal{F} = \int_0^1 \operatorname{CVaR}_\alpha dm(\alpha)$ *defined by a probability distribution* m *on* $[0,1)$*, its* $f_{\mathcal{F}}$ *is continuous.*

Proof. By definition of $\operatorname{CVaR}_\alpha$, we have $f_{\operatorname{CVaR}_\alpha} = f_\alpha$, where

$$f_\alpha(p) = \min\left(\frac{p}{1-\alpha}, 1\right)$$

In particular, since all f_α are concave and monotonically increasing on $[0,1]$, their integral $f_{\mathcal{F}}$ is also. Since $f_{\mathcal{F}}$ is concave on $(0,1)$, it is continuous there [Art15, Theorem 1.5].

For any $p \in (0,1)$, and any $n > 1$,

$$\begin{aligned}
f_{\mathcal{F}}\Big(\frac{p}{n}\Big) &= \int_{[0,1)} \min\left(\frac{\frac{p}{n}}{1-\alpha}, 1\right) dm(\alpha) \\
&= \int_{[0,1-p)} \frac{\frac{p}{n}}{1-\alpha} dm(\alpha) + \int_{\left[1-p,1-\frac{p}{n}\right)} \frac{\frac{p}{n}}{1-\alpha} dm(\alpha) + \int_{\left[1-\frac{p}{n},1\right)} 1 dm(\alpha) \\
&\leq \int_{[0,1-p)} \frac{\frac{p}{n}}{1-\alpha} dm(\alpha) + \int_{\left[1-p,1-\frac{p}{n}\right)} 1 dm(\alpha) + \int_{\left[1-\frac{p}{n},1\right)} 1 dm(\alpha) \\
&= \frac{p}{n} \int_{[0,1-p)} \frac{1}{1-\alpha} dm(\alpha) + m([1-p, 1))
\end{aligned}$$

Thus, $0 \leq \limsup_{p \to 0} f_{\mathcal{F}}(p) \leq \inf_{p \in (0,1)} m([1-p,1)) = 0$. So $f_{\mathcal{F}}$ is continuous at 0.

Finally, at $p \to 1$, since $f(p)$ is monotonically increasing, any discontinuity there can only be a jump upwards. Since $f_{\mathcal{F}}$ is also concave on $[0,1]$, it cannot have it, so $f_{\mathcal{F}}$ is continuous at 1. □

Corollary 3.33. *For any spectral risk measure $\mathcal{F} = \int_0^1 \mathrm{CVaR}_\alpha \, dm(\alpha)$ defined by a probability distribution m on $[0,1)$, it satisfies the generalized fundamental theorem of SLT.*

We give an explicit quantitative illustration:

Example 3.34. For any $\alpha \in (0,1)$, f_{CVaR_α} has the modulus of continuity $\frac{1}{1-\alpha}$, so any ϵ-representative sample Z^n is a $\frac{\epsilon}{1-\alpha}$-$\mathcal{F}$-representative sample.

By combining [SB14, Theorem 6.10, 6.11], if $\mathcal{H}$ has finite VC-dimension d, then for all $n \in \mathbb{N}$,

$$Pr_{Z^n \sim \mu^n}\left(Z^n \text{ is } g(n)\text{-representative}\right) \geq 1 - \delta$$

where

$$g(n) = \frac{4 + \sqrt{d(\ln(2n) - \ln d + 1)}}{\delta\sqrt{2n}}$$

Thus,

$$Pr_{Z^n \sim \mu^n}\left(Z^n \text{ is } \frac{g(n)}{1-\alpha}\text{-}\mathcal{F}\text{-representative}\right) \geq 1 - \delta$$

Since $\lim_n g(n) = 0$, for any $\epsilon, \delta \in (0,1)$, take any N such that $2g(N) < \frac{\epsilon}{1-\alpha}$, then for any $n > N$,

$$Pr_{Z^n \sim \mu^n}\left(Z^n \text{ is } \frac{\epsilon}{2}\text{-}\mathcal{F}\text{-representative}\right) \geq 1 - \delta$$

So, taking n training samples allows ERM to do $\mathcal{F}$-PAC-learning:

$$Pr_{Z^n \sim \mu^n}\left(\mathrm{Loss}_{\mathcal{F}}(h_{Z^n}, \mu) \leq \min_{h \in \mathcal{H}} \mathrm{Loss}_{\mathcal{F}}(h, \mu) + \epsilon\right) \geq 1 - \delta.$$

Chapter 4

Limit theorems of coherent risk measures

In this chapter, we start by reviewing limit theorems in probability and large deviation theory, then move on to a new proof of the central limit theorem (CLT) for an empirical process of CVaR, which generalizes the classical CLT for expectation.

In the literature, there have been generalizations of limit theorems results to CVaR. In particular, [Che07] proves the generalized central limit theorem, and [GW11], the generalized Berry–Esseen inequality, law of iterated logarithm, and large and moderate deviation principles. However, the proof presented below of the CLT of CVaRis new, as far as we are aware.

4.1 Limit theorems in probability

In calculus, each convergent real sequence (x_n) has a limit x_∞ associated with it. Taking the limit is a lossy operation that loses most of the details, but preserves something essential about the sequence (x_n), namely, its *eventual* behavior.

In probability, given a stochastic process (Y_n), one may ask if there exists some random variable or real number (which is a degenerate random variable) as its limit. The answer is yes, in certain senses. Making this precise gives us the limit theorems.

The most important limit theorems are the central limit theorem (CLT), the weak law of large number (WLLN), and the strong law of large number (SLLN).

4.1.1 Central limit theorem (CLT)

There are many formulations of CLT, but we will only need the classical CLT [Bil12, Theorem 27.1]:

Theorem 4.1 (classical CLT). *Suppose that (X_n) is an independent sequence of random variables having the same distribution with finite mean μ and variance σ^2. If $\overline{X_n} = \frac{1}{n}(X_1 + \cdots + X_n)$, then*

$$\sqrt{n}\,(\overline{X_n} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \tag{4.1}$$

Here, $\mathcal{N}$ denotes the normal distribution:

Definition 4.2. For any $\mu \in \mathbb{R}, \nu > 0$, $\mathcal{N}(\mu, \nu)$ is the standard normal distribution with mean μ and variance ν, with PDF

$$\rho(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{4.2}$$

As noted in Example 2.17, the CLT can be rephrased in the language of empirical process (L_n) of X:

$$\sqrt{n}\,(\mathbb{E}(L_n) - \mathbb{E}(X)) \xrightarrow{d} \mathcal{N}(0, 1) \tag{4.3}$$

This immediately suggests the generalization

$$\sqrt{n}\,(\mathrm{CVaR}_\alpha(L_n) - \mathrm{CVaR}_\alpha(X)) \xrightarrow{d} \mathcal{N}(0, \sigma(\alpha)) \tag{4.4}$$

where $0 \leq \alpha < 1$, and $\sigma(\alpha)$ is a function that possibly depends on α and X.

At $\alpha = 0$, this reduces to the original CLT, and so $\sigma(0) = 1$. At $\alpha = 1$,

$$\mathrm{CVaR}_\alpha(L_n) - \mathrm{CVaR}_\alpha(X) = \max_{i \in [n]} X_i \operatorname{ess\,sup}(X) > 0$$

has probability zero, and so the generalized CLT cannot be true.

As will be shown in Theorem 4.11, except at points of α where $F_X^{-1}(\alpha)$ is discontinuous, this generalization (Equation 4.4) is indeed true.

4.1.2 Strong laws of large numbers (SLLN)

Theorem 4.3 (Laws of large numbers for expectations). *If $\mathbb{E}(|X|) < \infty$, that is, $X \in \mathscr{L}^1$, then $\mathbb{E}(L_n)$ converges to $\mathbb{E}(X)$, in the senses of:*

(1) (Weak law) convergence in probability.

(2) (Strong law) almost sure convergence.

Both CLT and SLLN are "stronger" than WLLN, in that any random variable X that satisfies CLT or SLLN would satisfy WLLN. However, CLT and SLLN do not imply each other in general. As it is a corollary of SLLN, we will not mention WLLN any longer.

Theorem 4.4 (SLLN for CVaR)**.** *For any real random variable X, and any $0 \leq \alpha \leq 1$, $\mathrm{CVaR}_\alpha(L_n)$ converges to $\mathrm{CVaR}_\alpha(X)$ almost surely.*

For the case of $\alpha = 0$, this is the usual case of the SLLN. For the case of $\alpha = 1, \mathrm{CVaR}_1 = \operatorname{ess\,sup}$, and the proof is easy. The general case is Theorem 4.14, deferred to Section 4.2.2.

Theorem 4.5 (SLLN for $\operatorname{ess\,sup}$)**.** *For any real random variable X, $\operatorname{ess\,sup}(L_n)$ converges to $\operatorname{ess\,sup}(X)$ almost surely.*

Proof. First, the case of $\operatorname{ess\,sup}(X) < \infty$: For any $n \in \mathbb{N}$, $\operatorname{ess\,sup}(L_n) = \max(X_1, ..., X_n)$, so $(\operatorname{ess\,sup}(L_n))_n$ is a non-decreasing sequence.

By definition of essential supremum, $Pr(\operatorname{ess\,sup}(L_n) \leq \operatorname{ess\,sup}(X)) = 1$, so the sequence almost surely converges to a limit less or equal to $\operatorname{ess\,sup}(X)$, and it suffices to show $Pr(\limsup_n(\operatorname{ess\,sup}(L_n) - \operatorname{ess\,sup}(X)) \geq 0) = 1$.

For any $\epsilon > 0$,

$$Pr(\operatorname{ess\,sup}(L_n) < \operatorname{ess\,sup}(X) - \epsilon) = Pr(X < \operatorname{ess\,sup}(X) - \epsilon)^n \to 0$$

Thus, $Pr(\lim_n(\operatorname{ess\,sup}(L_n) - \operatorname{ess\,sup}(X)) \geq -\epsilon) = 1$ for all $\epsilon > 0$, and so $Pr(\limsup_n(\operatorname{ess\,sup}(L_n) - \operatorname{ess\,sup}(X)) \geq 0) = 1$.

For the case of $\operatorname{ess\,sup}(X) = \infty$, the same proof applies, after replacing $\operatorname{ess\,sup}(X) - \epsilon$ by M, an arbitrarily big number. □

4.1.3 Law of the iterated logarithm (LIL)

Lying between SLLN and CLT is the law of the iterated logarithms (LIL). With no loss of generality, consider a random variable X with mean 0 and variance 1, and $(X_n)_n$ being its IID process. SLLN states that $\frac{1}{n}\sum_{i=1}^n X_i \overset{\text{a.s.}}{\to} 0$, that is, the distribution of $\frac{1}{n}\sum_{i=1}^n X_i$ converges to δ_0 "quickly". CLT states that the distribution of $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$ converges to $\mathcal{N}(0, 1)$.

Intermediate between them, LIL states that $\frac{1}{\sqrt{n \ln \ln n}}\sum_{i=1}^n X_i$ converges to δ_0, but slowly, so that $\limsup_n \frac{1}{\sqrt{n \ln \ln n}}\sum_{i=1}^n X_i = \sqrt{2}$ almost surely.

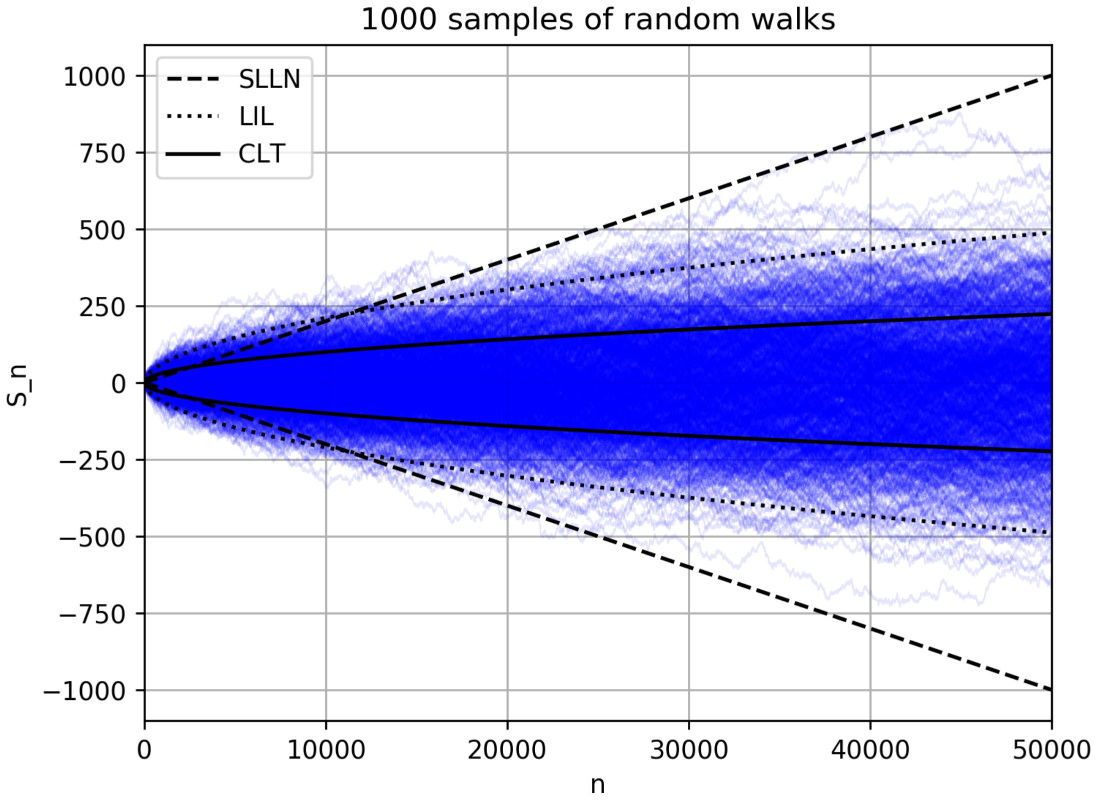


Figure 4.1: 1000 samples of random walks, with each step having mean 0 and variance 1. Almost all the walks eventually fall inside the outer dashed cone of $y = \pm\epsilon n$, demonstrating the SLLN. Most of the walks barely touch the edges of the cone $y = \pm\sqrt{2n \ln\ln n}$, demonstrating the LIL. About 68% of the walks are inside the cone $y = \pm\sqrt{n}$ at the right edge of the graph, demonstrating one instance of the CLT.

Theorem 4.6 (Law of the iterated logarithms). *Given random variable X with mean 0 and variance 1, and $(X_n)_n$ being its IID process, then*

$$\limsup_n \frac{1}{\sqrt{n \ln\ln n}} \sum_{i=1}^{n} X_i = \sqrt{2} \textit{ almost surely.} \tag{4.5}$$

By symmetry, we also have

$$\liminf_n \frac{1}{\sqrt{n \ln\ln n}} \sum_{i=1}^{n} X_i = -\sqrt{2} \textit{ almost surely.} \tag{4.6}$$

Pictorially, one can consider a random walk process $(S_n)_n$ defined by $S_n = \sum_{i=1}^n X_i$, shown in Figure 4.1.

SLLN states that for any $\epsilon > 0$, with probability one, a randomly chosen walk would eventually be contained in the cone $y = \pm\epsilon x$.

CLT states that for any $\epsilon > 0$, for big n, the probability that a randomly chosen path is within the cone $y = \pm\epsilon\sqrt{x}$ at the $x = n$ section (that is, $S_n \in (-\epsilon\sqrt{x}, \epsilon\sqrt{x})$) is $Pr(\mathcal{N}(0,1) \in (-\epsilon, \epsilon))$.

Intermediately, LIL states that with probability one, a random path would touch the edge of the cone $y = \pm\sqrt{2x \ln\ln x}$ infinitely many times, but for any $\epsilon > 0$, it would only touch the edge of $y = \pm(1+\epsilon)\sqrt{2x \ln\ln x}$ finitely many times.

While SLLN and CLT are extensively used in practical statistics, the LIL in comparison has little practical consequence [Van00]. One classic paper on applications of LIL to statistics is [Rob70].

Before plunging into large deviation theory, which we will use to derive the CLT for certain risk measures, we take note of the surrounding territory for context.

In the literature, one can often find mentions of **large/moderate/small deviation principles**. Among these, the *large* deviation principles are the most popular. Two standard references on this subject are [DZ09; Den08].

4.1.4 Deviation principles

Consider a random variable X with mean 0 and variance 1, and its IID process $(X_n)_n$. Let $S_n = X_1 + \cdots + X_n$. By the CLT, we have for any constant $\epsilon > 0$,

$$\lim_n Pr(S_n > \epsilon n^{\frac{1}{2}}) = 1 - \Phi(\epsilon), \tag{4.7}$$

where Φ is the CDF of the standard normal distribution.

If $|X|$ has finite third moment, then by the Berry–Esseen theorem [Dur10, Theorem 3.4.9], this convergence is uniform in the sense that

$$\lim_n \frac{Pr(S_n > x_n n^{\frac{1}{2}})}{1 - \Phi(x_n)} = 1 \tag{4.8}$$

for any sequence of $(x_n)_n$ that satisfies $x_n = O(1)$. This, though not often called so, is a *small* deviation principle.

One immediately considers generalization for x_n that may grow faster than $O(1)$.

The large deviation theorem would give an asymptotic expansion in the case where x_n is $O(n^{\frac{1}{2}})$. For example, Cramér's theorem states that if X is "nice", there exists a **rate function** $I : \mathbb{R} \to [0, \infty)$ such that

$$\forall \epsilon > 0,\ Pr(S_n > \epsilon n) \to e^{-nI(\epsilon)}.$$

More rigorously, the convergence is

$$\frac{1}{n}\ln Pr(S_n > \epsilon n) \to -I(\epsilon). \tag{4.9}$$

Between them, a moderate deviation theorem [CFS13] states that the asymptotic expansion is

$$\frac{Pr(S_n > x_n n^{\frac{1}{2}})}{1-\Phi(x_n)} = 1 + O(1)\frac{1+x_n^3}{\sqrt{n}}. \tag{4.10}$$

for $x_n = O(n^{\frac{1}{6}})$. In more details, it states that

$$\left(\frac{Pr(S_n > x_n n^{\frac{1}{2}})}{1-\Phi(x_n)} - 1\right)\frac{\sqrt{n}}{1+x_n^3}. \tag{4.11}$$

is bounded as $n \to \infty$.

In [RS65], a moderate deviation defined by $x_n = O(\sqrt{\ln n})$ is studied. In general, moderate deviation studies

$$O(1) < x_n < O(\sqrt{n}),$$

that is,

$$|x_n| \to \infty, \ \frac{x_n}{\sqrt{n}} \to 0. \tag{4.12}$$

A large part of modern probability consists of various generalizations of the deviation principles under assumptions on $(X_n)_n$ weaker than full independence [*]. The sequences of $(X_n)_n$ could also be generalized to be multidimensional, or graphs, or some other complicated mathematical objects.

4.1.5 The Gärtner–Ellis theorem

The Gärtner–Ellis theorem states a large deviation principle for sequences of not necessarily independent random variables. It uses a generalization of the cumulant generating function:

Definition 4.7. For any real random variable X, its **cumulant generating function** is

$$\forall t \in \mathbb{R}, \quad K(t) = \ln \mathbb{E}(e^{tX}). \tag{4.13}$$

[*] Such as “weakly dependent”, “strongly mixing”, “exchangeable”, “ergodic”, and many other highly technical weakenings.

Consider, for example, the empirical process (L_n) of X, and the sequence $\mathbb{E}(L_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$, then we have

$$\mathbb{E}\left(e^{nt\mathbb{E}(L_n)}\right) = \mathbb{E}\left(e^{tX}\right)^n$$

and so for any $t \in \mathbb{R}$,

$$K(t) = \lim_{n\to\infty} \frac{1}{n} \ln \mathbb{E}\left(e^{nt\mathbb{E}(L_n)}\right)$$

For a general sequence of real random variables, $(Y_n)_n$, let

$$K_n(t) = \frac{1}{n} \ln \mathbb{E}\left(e^{nt\mathbb{E}(L_n)}\right),$$

then if the following limit exists

$$\lim_{n\to\infty} K_n(t)$$

for all t in a neighborhood of 0, then the Gärtner–Ellis theorem gives the limit behavior of a properly scaled version of the sequence $(Y_n)_n$.

There are many versions of the Gärtner–Ellis theorem with varying generalities. The version that we use is [CG84, Lemma 1]:

Theorem 4.8 (Gärtner–Ellis theorem)**.** *For a general sequence of real random variables, $(Y_n)_n$, if the following limit exists*

$$K(t) = \lim_{n\to\infty} \frac{1}{n} \ln \mathbb{E}\left(e^{ntY_n}\right)$$

for t in a neighborhood (ϵ_-, ϵ_+) of 0, and if K is strictly convex and C^2 on (ϵ_-, ϵ_+), then, letting $\mu = K'(0), \sigma^2 = K''(0)$,

(1) (SLLN) $Y_n \overset{a.s.}{\to} \mu$.

(2) (CLT) If for all sufficiently large n, K_n is convex on $[0, \epsilon_+)$, and $\lim_n K_n''(0) = \sigma^2$, then

$$\frac{Y_n - \mathbb{E}(Y_n)}{\sqrt{n}} \overset{d}{\to} \mathcal{N}(0, \sigma^2). \tag{4.14}$$

4.2 Limit theorems of CVaR

In this section, we present calculations and numerical evidence that demonstrate, if not prove with full rigor, the CLT and SLLN of CVaR.

4.2.1 CLT for CVaR

Consider a real X with PDF ρ and CDF Φ. For any $h \in \mathbb{N}$, define

$$\begin{aligned}\mathbb{E}(\exp(nt\,\mathrm{CVaR}_\alpha(L_n))) &= \mathbb{E}\left(\exp\left(\frac{t}{\bar\alpha}\sum_{i=1}^{\bar\alpha n} X_{(i)}\right)\right)\\ &= \mathbb{E}\left(\exp\left(\frac{t}{\bar\alpha}\sum_{i=1}^{\bar\alpha n} X_i\right)\middle| X_1,...,X_{\bar\alpha n} > X_{\bar\alpha n+1} > X_{\bar\alpha n+2},...X_n\right)\end{aligned}$$

Here, $X_{(i)}$ denotes the i-th biggest term in the sequence $X_1,...X_n$. The cases where two or more X_i are equal having measure 0, thus ignored.

Note that the sum should not be taken literally, as in actuality, $\mathrm{CVaR}_\alpha(L_n)$ is the average of the biggest $\lfloor \bar\alpha n\rfloor$ terms of $X_1,...,X_n$, plus $\{\bar\alpha n\}$ times the next biggest term. However, as n grows, this little fudge factor will be swamped out, and therefore we ignore it.

That the expectation can be conditioned on a particular choice of ordering of X_i is because the sequence $(X_n)_n$ is an **exchangeable sequence of random variables**, that is, any finite permutation of the sequence creates a sequence with the same distribution.

Now we continue the calculation.

$$\begin{aligned}&= \int_{\mathbb{R}} Pr(X_{\bar\alpha n+1}\in dx | X_1,...,X_{\bar\alpha n} > X_{\bar\alpha n+1} > X_{\bar\alpha n+2},...X_n)\mathbb{E}(e^{\frac{tX}{\bar\alpha}}|X>x)^{\bar\alpha n}\\ &= \int_{\mathbb{R}} \frac{F_X(x)^{\alpha n-1}(1-F_X(x))^{\bar\alpha n}\rho(x)dx}{B(\alpha n, \bar\alpha n+1)}\mathbb{E}(e^{\frac{tX}{\bar\alpha}}|X>x)^{\bar\alpha n}\end{aligned}$$

where $B(x,y)=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ is the Euler beta function, and Γ is the Euler gamma function.

Plugging in

$$(1-F_X(x))\mathbb{E}(e^{\frac{tX}{\bar\alpha}}|X>x) = \int_x^\infty e^{ty/\bar\alpha}\rho(y)dy,$$

we obtain

$$= \frac{1}{B(\alpha n,\bar\alpha n+1)}\int_{\mathbb{R}}\left(F_X(x)^\alpha\left(\int_x^\infty e^{ty/\bar\alpha}\rho(y)dy\right)^{\bar\alpha}\right)^n \frac{\rho(x)}{F_X(x)}dx$$

Now, by Stirling's approximation,

$$B(\alpha n,\bar\alpha n+1) = \exp(-nH(\alpha)+O(\ln n)),$$

where

$$H(\alpha) = -\alpha\ln\alpha - \bar\alpha\ln\bar\alpha \tag{4.15}$$

is the binary entropy function.

So, by Laplace's method [BO99, Section 6.4],

$$K(t) = H(\alpha) + \max_{x \in \mathbb{R}} \left(\alpha \ln F_X(x) + \bar{\alpha} \ln \left(\int_x^\infty e^{ty/\bar{\alpha}} \rho(y) dy \right) \right) \tag{4.16}$$

provided that

$$\alpha \ln F_X(x) + \bar{\alpha} \ln \left(\int_x^\infty e^{ty/\bar{\alpha}} \rho(y) dy \right) \tag{4.17}$$

has a unique global maximum at some x_0, is C^2 in a neighborhood of x_0, and $\frac{\rho(x_0)}{F_X(x_0)} > 0$.

Taking derivative, such a maximum is a root of

$$F_X(x) = \frac{\alpha}{\bar{\alpha}} \int_0^\infty e^{ty/\bar{\alpha}} \rho(x+y) dy. \tag{4.18}$$

At $t = 0$ has solution $x = F_X^{-1}(\alpha)$, and we obtain

$$K(0) = H(\alpha) + \alpha \ln F_X(F_X^{-1}(\alpha)) + \bar{\alpha} \ln \int_{F_X^{-1}(\alpha)}^\infty \rho(y) dy = 0$$

as it should.

Near $t = 0$, x can be expanded as a power series $x = F_X^{-1}(\alpha) + x_1 t + x_2 t^2 + o(t^2)$, which can then be plugged into the equation of $K(t) = \mu(\alpha)t + \frac{1}{2}\sigma(\alpha)^2 + o(t^2)$, from which we obtain

$$\sqrt{n}(\mathrm{CVaR}_\alpha(L_n) - \mu(\alpha)) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma(\alpha)^2) \tag{4.19}$$

Example 4.9. X is uniform on $[0,1]$, then

$$K(t) = H(\alpha) + \max_{x \in (0,1)} \left(\alpha \ln x + \bar{\alpha} \ln \frac{\bar{\alpha}}{t} \left(e^{t/\bar{\alpha}} - e^{tx/\bar{\alpha}} \right) \right) \tag{4.20}$$

The maximizer x is the root of

$$x = \frac{\alpha}{t} \left(e^{t(1-x)/\bar{\alpha}} - 1 \right)$$

which has asymptotic expansion

$$x = \alpha + x_1 t + x_2 t^2 + o(t^2).$$

Plugging it in and solving up to t^2 order, we obtain

$$x = \alpha + \frac{1}{2}\alpha\bar{\alpha}t + \frac{1}{6}\alpha\bar{\alpha}(1 - 3\alpha)t^2$$

and so

$$K(t) = \frac{1}{2}(1+\alpha)t + \frac{1}{24}(1-\alpha)(1+3\alpha)t^2 + o(t^2)$$

So we obtain the CLT for X:

$$\mu(\alpha) = \frac{1}{2}(\alpha+1) = \mathrm{CVaR}_\alpha(X), \quad \sigma(\alpha)^2 = \frac{1}{12}(1-\alpha)(1+3\alpha) \tag{4.21}$$

This is illustrated in Figures 4.2 and 4.3.

Example 4.10. A discrete X with a finite discrete distribution $\sum_{i=1}^N p_i\delta_{x_i}$ can be approximated by very concentrated uniform distributions, that is,

$$\rho(x) = \begin{cases} \frac{p_i}{\epsilon} & x \in [x_i, x_i+\epsilon] \\ 0 & \text{else} \end{cases},$$

where ϵ is a positive number smaller than $\min_{1\le i\le N-1}(x_{i+1}-x_i)$.

Let $P_i = p_1 + ... + p_i$ for all $i \in [N]$, then if $\alpha \in (P_{i-1}, P_i)$ for some $i \in [N]$, then $F_X^{-1}(\alpha) = x_i + \frac{\epsilon}{p_i}(\alpha - P_{i-1})$. Then, the maximum in Equation 4.17 is the root of

$$P_{i-1} + \frac{p_i}{\epsilon}(x-x_i) = \frac{\alpha}{\epsilon t}\left(p_i\left(e^{(\epsilon-(x-x_i))t/\bar{\alpha}} - 1\right) + \sum_{j=i+1}^N \left(p_j e^{(x_j-x)t/\bar{\alpha}}\left(e^{\epsilon t/\bar{\alpha}}-1\right)\right)\right)$$

which has the power expansion

$$x = x_i + \frac{\epsilon}{p_i}(\alpha - P_{i-1}) + At + Bt^2 + o(t^2)$$

Plugging it in to solve for A, B, then expanding K, and taking the $\epsilon \to 0$ limit, we obtain

$$K(t) = \mu(\alpha)t + \frac{1}{2}\sigma(\alpha)^2t^2$$

where

$$\mu(\alpha) = \frac{1}{\bar{\alpha}}(x_i(P_i-\alpha) + \sum_{k>i}^N p_kx_k) = \mathrm{CVaR}_\alpha(X), \tag{4.22}$$

$$\sigma(\alpha)^2 = \frac{1}{\bar{\alpha}^2}\left((P_i(1-P_i)x_i^2 - 2P_ix_i\sum_{k>i}^N p_kx_k + \sum_{k>i}^N p_kx_k^2 - \left(\sum_{k>i}^N p_kx_k\right)^2\right) \tag{4.23}$$

After routine algebra, this is simplified to $\mathbb{V}\left(\frac{1}{\bar{\alpha}}(X - F_X^{-1}(\alpha))^+\right)$, where for any random variable Y, $Y_+ = \max(Y, 0)$ is the positive part of Y.

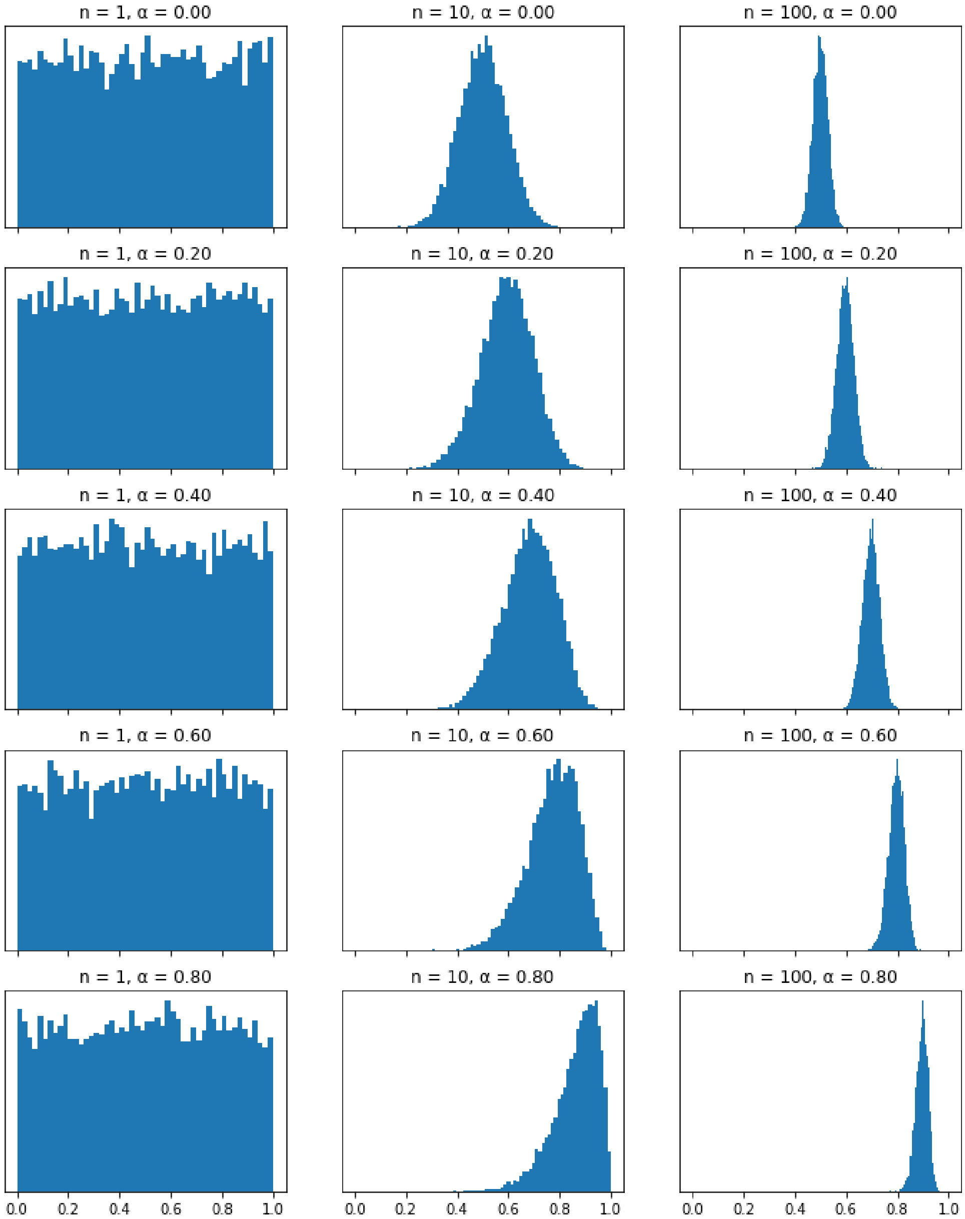
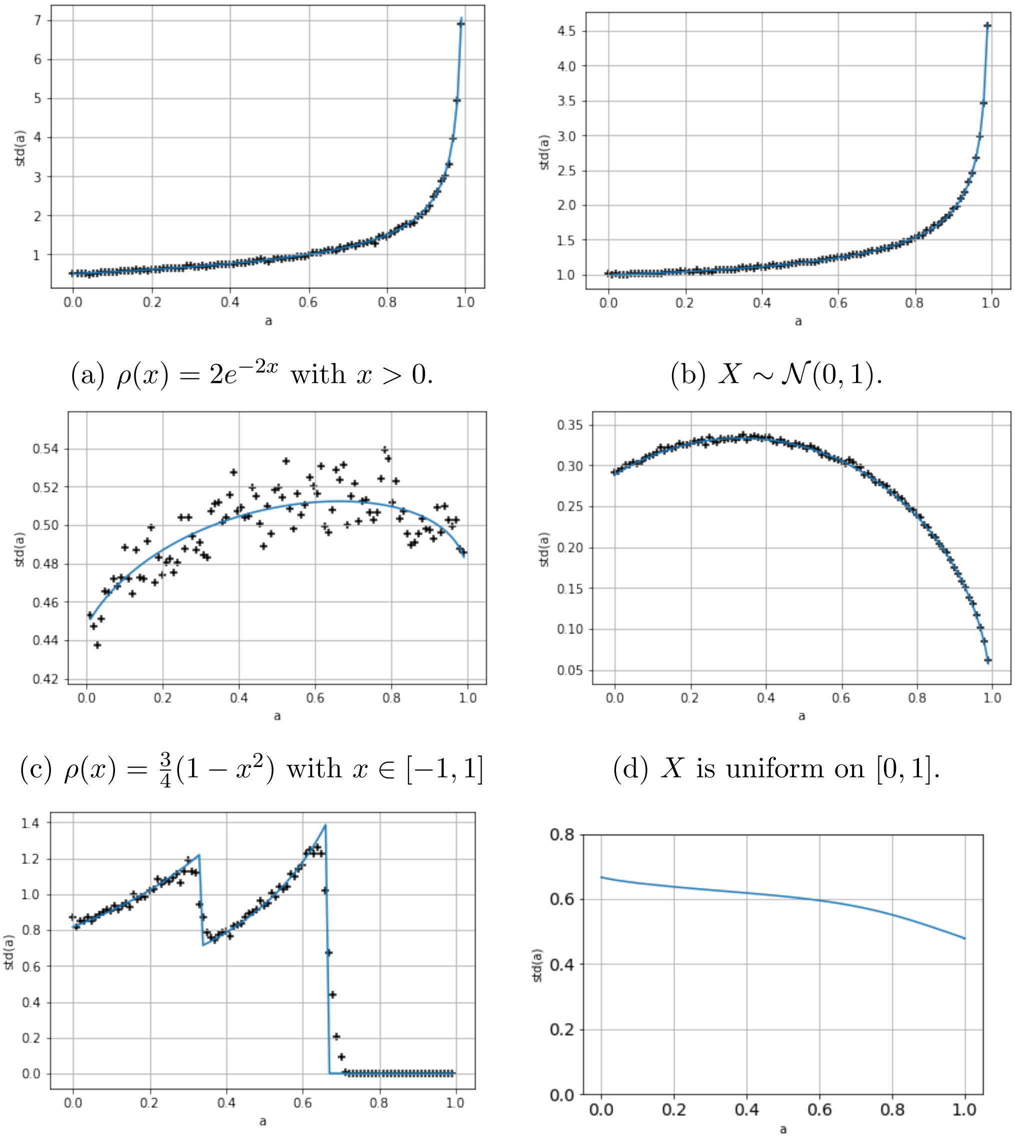


Figure 4.2: A demonstration of the CLT for CVaR. Here, X is the uniform distribution on $[0, 1]$, and the PDF of $\mathrm{CVaR}_\alpha(L_n)$ is plotted as a function of n and α. As n increases, the distributions converge to normal distributions. Increasing α both shifts the distribution to the right, and distort it away from normality. Each histogram is the result of 10^4 trials.



(a) $\rho(x) = 2e^{-2x}$ with $x > 0$.

(b) $X \sim \mathcal{N}(0, 1)$.

(c) $\rho(x) = \frac{3}{4}(1 - x^2)$ with $x \in [-1, 1]$

(d) X is uniform on $[0, 1]$.

(e) X is discrete uniform on $\{0, 1, 2\}$.

(f) $\rho(x) = -xe^{-x^2/2}$ with $x < 0$.

Figure 4.3: Numerical confirmation of Theorem 4.11 (CLT for CVaR). $\sigma(\alpha)$ plots for six different distributions of X are calculated, and five of them fits the numerical simulation. The last plot is calculated only theoretically, without numerical verification. The curves are the exact theoretical prediction of the standard deviation of $\sqrt{n}\,\mathrm{CVaR}_\alpha(L_n)$ as $n \to \infty$, given by Equation 4.25. Each point in the scatterplots is obtained by sampling $\sqrt{n}\,\mathrm{CVaR}_\alpha(L_n)$ for 1000 times, where $n = 1000$.

When α is equal to some P_i, that is, when F_X^{-1} is discontinuous at α, the result is not determined by this method, as $K(t)$ does not have continuous second-derivative in a neighborhood of $t = 0$.

For arbitrary X with finite variance, its distribution can be approximated as the limit of discrete distributions, and so we have obtained

Theorem 4.11 (CLT for CVaR)**.** *For any real random variable X with finite variance, and any $\alpha \in (0, 1)$, if F_X^{-1} is continuous at α, then the empirical process of the* CVaR_α *of X satisfies*

$$\sqrt{n}(\mathrm{CVaR}_\alpha(L_n) - \mathrm{CVaR}_\alpha(X)) \xrightarrow{d} \mathcal{N}(0, \sigma(\alpha)) \tag{4.24}$$

where

$$\begin{aligned}\sigma(\alpha)^2 &= \frac{1}{\bar{\alpha}}\mathbb{E}[(X - F_X^{-1}(\alpha))^2 | X > F_X^{-1}(\alpha)] - \mathbb{E}[(X - F_X^{-1}(\alpha))|X > F_X^{-1}(\alpha)]^2 \\ &= \mathbb{V}\left(\frac{1}{\bar{\alpha}}(X - F_X^{-1}(\alpha))^+\right)\end{aligned} \tag{4.25}$$

A more abstract and general version of the CLT for empirical CVaR, that weakens assumption of independence of the process (X_n) to merely α-mixing, is proved in [Che07, Theorem 1].

In [Bra+08, Theorem 3.1], an alternative formula for $\sigma(\alpha)$ is given:

$$\sigma(\alpha)^2 = \frac{1}{(1-\alpha)^2}\int_{F_X^{-1}(\alpha)}^{\infty}\int_{F_X^{-1}(\alpha)}^{\infty} (F_X(\min(x, y)) - F_X(x)F_X(y))\, dxdy \tag{4.26}$$

As for α where F_X^{-1} is discontinuous, we conjecture that the CLT simply fails, and instead, the limit distribution is a "mixed" normal distribution.

Definition 4.12. Given $\sigma_1, \sigma_2 > 0$, the mixed normal distribution $\mathcal{N}_{mixed}(\mu, \sigma_1, \sigma_2)$ is the distribution with PDF

$$\rho(x) = \begin{cases} \rho_1(x)\frac{2\sigma_1}{\sigma_1+\sigma_2} & x \le \mu \\ \rho_2(x)\frac{2\sigma_2}{\sigma_1+\sigma_2} & x \ge \mu \end{cases} \tag{4.27}$$

where ρ_i is the PDF of $\mathcal{N}(\mu, \sigma_i)$, with $i = 1, 2$.

Conjecture 4.13. *Given X with finite variance, for any $\alpha \in (0, 1)$ such that F_X^{-1} is discontinuous at α,*

$$\sqrt{n}(\mathrm{CVaR}_\alpha(L_n) - \mathrm{CVaR}_\alpha(X)) \xrightarrow{d} \mathcal{N}_{mixed}(\mu, \sigma_1, \sigma_2) \tag{4.28}$$

where

$$\mu = \mathrm{CVaR}_\alpha(X), \tag{4.29}$$

and

$$\sigma_1 = \lim_{z \nearrow \alpha} \sigma(z), \quad \sigma_2 = \lim_{z \searrow \alpha} \sigma(z). \tag{4.30}$$

are the two one-sided limits of $\sigma(\alpha)$*.*

This conjecture cannot be proved by Theorem 4.8, since when α is at those critical values, $K(t)$ has no second-derivative at 0.

We tested this conjecture by numerically calculating the distribution of $\mathrm{CVaR}_\alpha(L_n)$, for a big $n = 10^5$, X being uniformly distributed on $\{0, 1, 2\}$, and for $\alpha \approx 1/3$, around a discontinuous point of F_X^{-1}.

As shown in Figure 4.4, as α is about the discontinuous point, the right side of the bell curve suddenly shrinks in width from σ_l down to σ_u. Then, just after α crosses the discontinuity, left side shrinks too.

The conjecture predicts that it should have a mixed normal distribution defined by

$$\mu = 1.5, \sigma_l = \sqrt{1.5/n} = 0.003873, \sigma_u = \sqrt{0.5/n} = 0.002236.$$

As shown in Figure 4.5, this is close to the numerical best fit

$$\mu = 1.5 + 2.1 \times 10^{-4}, \sigma_l = 0.003729, \sigma_u = 0.002217.$$

4.2.2 SLLN for CVaR

Now we present the SLLN for the empirical process of CVaRassuming X has finite variance.

In [AT02, Proposition 4.1], this is proved assuming only that

$$\mathbb{E}((-X)^+) < \infty,$$

but the proof is more involved.

Theorem 4.14 (SLLN for CVaR)**.** *For any real random variable X with finite variance, and any $\alpha \in [0, 1]$, then the empirical process of the* CVaR_α *of X satisfies*

$$\mathrm{CVaR}_\alpha(L_n) \overset{a.s.}{\to} \mathrm{CVaR}_\alpha(X). \tag{4.31}$$

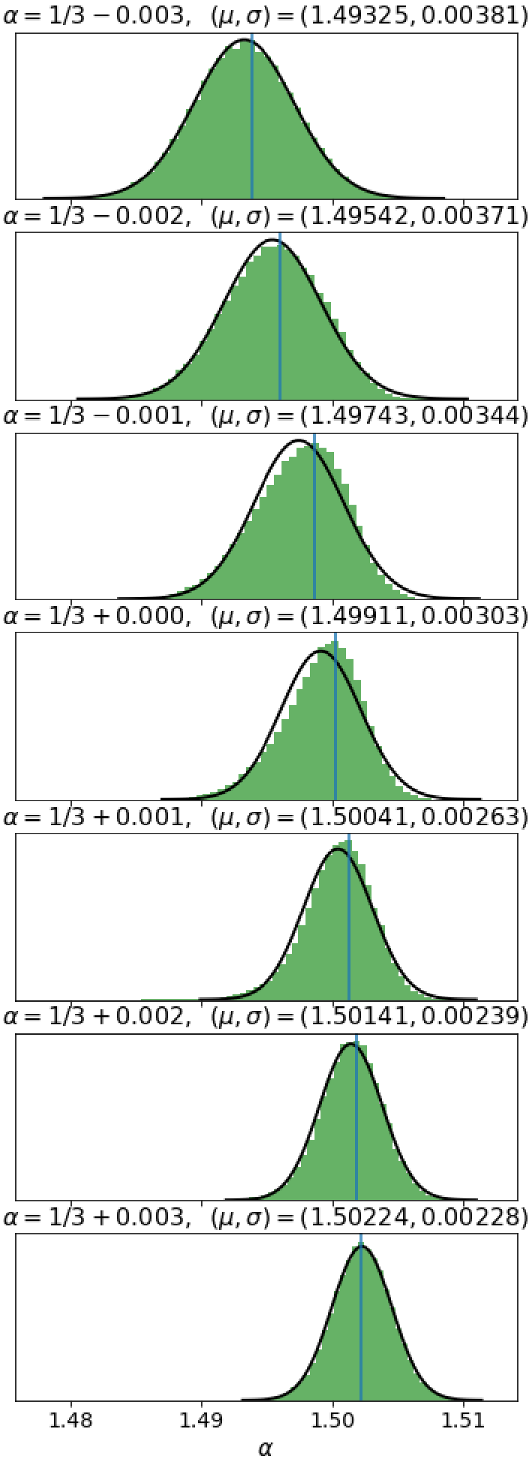


Figure 4.4: Distribution of $\mathrm{CVaR}_\alpha(L_n)$, with $n = 10^5$, X being the uniform distribution on $\{0, 1, 2\}$, as α crosses the $1/3$ boundary. Each histogram results from 10^5 trials. Each black curve is the best fit normal distribution $\mathcal{N}(\mu, \sigma)$, with parameters μ, σ written above. Each vertical line denotes the estimated maximum of the distribution of $\mathrm{CVaR}_\alpha(L_n)$. Notice how nearing $\alpha = 1/3$, the fit to normal distribution degrades , and the estimated maximum shifts ahead.

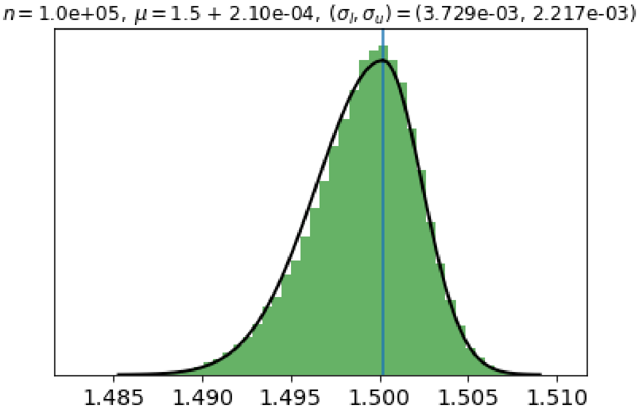


Figure 4.5: Distribution of $\mathrm{CVaR}_{1/3}(L_n)$, with $n = 10^5$, X being the uniform distribution on $\{0, 1, 2\}$. The histogram results from 10^5 trials. The black curve is the best fit mixed normal distribution $\mathcal{N}(\mu, \sigma_l, \sigma_u)$, with $\mu = 1.5 + 2.1 \times 10^{-4}, \sigma_l = 0.003729, \sigma_u = 0.002217$, and the vertical line denotes the location of μ.

Proof. There are four cases to consider.

1. If $\alpha = 0$, it is simply the SLLN for expectations.

2. If $\alpha = 1$, it is proved in Theorem 4.5.

3. If $\alpha \in (0, 1)$, and F_X^{-1} is continuous at α, then the power series $K(t) = \mu(\alpha)t + \frac{1}{2}\sigma(\alpha)^2 t^2 + o(t^2)$ in a neighborhood of 0 means that K is stricly convex and C^2 in that neighborhood. Now apply part (a) of Theorem 4.8.

4. If $\alpha_0 \in (0, 1)$, and F_X^{-1} is not continuous at α_0, then we prove by "squeezing with nearby points of continuity".

Since F_X^{-1} is monotone, by Lebesgue's differentiation theorem [RFR10, Section 6.2], it is almost everywhere differentiable. That is, let

$$D = \{x \in (0, 1) : F_X^{-1} \text{ is differentiable at } x\}$$

then D has Lebesgue measure 1.

For any $\epsilon > 0$, since $\mathrm{CVaR}_\alpha(X)$ is a continuous function of α, there exists $\delta > 0$ such that

$$\forall \alpha \in (\alpha_0 - \epsilon, \alpha_0 + \epsilon), |\,\mathrm{CVaR}_\alpha(X) - \mathrm{CVaR}_{\alpha_0}(X)| < \delta$$

Now take $\alpha_1, \alpha_2 \in D \cap (\alpha_0 - \epsilon, \alpha_0 + \epsilon)$ such that $\alpha_1 < \alpha_0 < \alpha_2$. Note that α_1, α_2 exists because D has measure 1.

Then, by the SLLN for $\mathrm{CVaR}_{\alpha_1}, \mathrm{CVaR}_{\alpha_2}$, and the monotonicity of CVaR_α as a function of α, we have

$$\limsup_n \mathrm{CVaR}_{\alpha_0}(L_n) \leq \limsup_n \mathrm{CVaR}_{\alpha_2}(L_n) \overset{\text{a.s.}}{\to} \mathrm{CVaR}_{\alpha_2}(X) \leq \mathrm{CVaR}_{\alpha_0}(X) + \epsilon$$

and similarly,

$$\liminf_n \mathrm{CVaR}_{\alpha_0}(L_n) \geq \mathrm{CVaR}_{\alpha_0}(X) - \epsilon \text{ almost surely}$$

Since for all $\epsilon > 0$, these two inequalities hold almost surely, we have, almost surely,

$$\lim_n \mathrm{CVaR}_{\alpha_0}(L_n) = \mathrm{CVaR}_{\alpha_0}(X).$$

□

In fact, with a little manipulation, we immediately strengthen it to

Theorem 4.15 (Uniform SLLN for CVaR)**.** *For any real random variable X with finite variance, almost surely, for any $\alpha \in [0,1]$, the empirical process of the* CVaR_α *of X satisfies*

$$\mathrm{CVaR}_\alpha(L_n) \to \mathrm{CVaR}_\alpha(X). \tag{4.32}$$

that is,

$$Pr\left(\forall \alpha \in [0,1],\ \mathrm{CVaR}_\alpha(L_n) \to \mathrm{CVaR}_\alpha(X)\right) = 1 \tag{4.33}$$

Proof. Since for any particular $\alpha \in [0,1]$,

$$\mathrm{CVaR}_\alpha(L_n) \overset{\text{a.s.}}{\to} \mathrm{CVaR}_\alpha(X)$$

so with probability one, it holds simultaneously for the countably many rational $\alpha \in [0,1]$. Then by a "squeezing" argument like in the previous proof, it holds simultaneously for all $\alpha \in [0,1]$:

For any $\alpha \in (0,1)$, and any $\epsilon > 0$, by continuity of CVaR_α with respect to α, there exists rational $\alpha_1, \alpha_2 \in (0,1)$, such that $\alpha_1 < \alpha < \alpha_2$, and

$$\mathrm{CVaR}_\alpha(X) - \epsilon < \mathrm{CVaR}_{\alpha_1}(X) \leq \mathrm{CVaR}_\alpha(X) \leq \mathrm{CVaR}_{\alpha_2}(X) < \mathrm{CVaR}_\alpha(X) + \epsilon$$

noting that $\mathrm{CVaR}_\alpha(X)$ must be finite, since X has finite expectation, and $\alpha < 1$.

Now, if $\mathrm{CVaR}_{\alpha_i}(L_n) \to \mathrm{CVaR}_{\alpha_i}(X)$ for $i = 1, 2$, then

$$\mathrm{CVaR}_\alpha(X) - \epsilon \leq \liminf_n \mathrm{CVaR}_\alpha(L_n)$$

$$\mathrm{CVaR}_\alpha(X) + \epsilon \geq \limsup_n \mathrm{CVaR}_\alpha(L_n)$$

Since this holds for all $\epsilon > 0$, we have

$$\mathrm{CVaR}_\alpha(L_n) \to \mathrm{CVaR}_\alpha(X).$$

□

4.2.3 CLT of CVaR_α in the $\alpha \to 1$ limit

Now we study the qualitative behavior of

$$\sigma(\alpha)^2 = \mathbb{V}\left(\frac{1}{\bar{\alpha}}(X - F_X^{-1}(\alpha))^+\right)$$

at the $\alpha \to 1$ limit.

Let

$$f(\alpha) = \bar{\alpha}\sigma(\alpha) = \sqrt{\mathbb{V}\left((X - F_X^{-1}(\alpha))^+\right)} \tag{4.34}$$

then f is a monotonically decreasing function of $\alpha \in (0, 1)$.

At the $\alpha \to 0$ limit,

$$f(\alpha) \to \sigma(0)^2 = \mathbb{V}(X).$$

The behavior of f at the $\alpha \to 1$ limit depends on the right tail of the distribution of X. Qualitatively speaking, the thinner its right tail, the faster it approaches zero. Figure 4.6 is a schematic plot of the behavior of f, for X with tails of various thicknesses.

The following examples are summarized by Table 4.1.

Table 4.1: The right tail of $\bar{\alpha}\sigma(\alpha)$ in the $\alpha \to 1$ limit. The right tail of ρ is either the $x \to \infty$ or the $x \to 0$ limit. As the right tail of ρ becomes thinner, the right tail of $\bar{\alpha}\sigma(\alpha)$ converges to 0 quicker.

right tail of ρ	right tail of $\bar{\alpha}\sigma(\alpha)$
$\frac{1}{x^{3+\epsilon}}$	$\bar{\alpha}^{\frac{\epsilon}{4}}$
$\frac{1}{x^n}, (n > 3)$	$\bar{\alpha}^{\frac{n-3}{2n-2}}$
$\frac{1}{x^{1001}}$	$\bar{\alpha}^{0.499}$
e^{-x}	$\bar{\alpha}^{0.5}$
$e^{-x^2/2}$	$\frac{\sqrt{\bar{\alpha}}}{\mathrm{erfc}^{-1}(2\bar{\alpha})}$
$(-x)^{999}$	$\bar{\alpha}^{0.5005}$
$(-x)^n, (n > -1)$	$\bar{\alpha}^{\frac{n+3}{2n+2}}$
$(-x)^1$	$\bar{\alpha}$
$(-x)^0$	$\bar{\alpha}^{1.5}$
$(-x)^{-1+\epsilon}$	$\bar{\alpha}^{\frac{1}{\epsilon}}$

Now we list some examples, some of which are shown in Figure 4.3

Example 4.16. Suppose X is uniform on $\{0,1,2\}$, then

$$\sigma(\alpha)^2 = \begin{cases} \frac{2}{3\bar{\alpha}^2} \text{ if } \alpha \in [0,1/3) \\ \frac{2}{9\bar{\alpha}^2} \text{ if } \alpha \in (1/3,2/3) \\ 0 \text{ if } \alpha \in (2/3,1) \end{cases} \tag{4.35}$$

as shown in Figure 4.3e.

Example 4.17. Suppose X has a right tail of the form

$$\rho(x) \approx \frac{A}{x^n}$$

where $n > 3$ to ensure that X has finite variance, and $A > 0$ is an unspecified constant, then at the $\alpha \to 1$ limit, approximately

$$\sigma(\alpha)^2 \propto \frac{1}{\bar{\alpha}^2 \left(-\ln \bar{\alpha}\right)^{n-3}} \tag{4.36}$$

giving $f(\alpha) \propto \left(-\ln \bar{\alpha}\right)^{-(n-3)/2}$ at $\alpha \to 1$ limit.

Example 4.18. The exponential distribution with PDF

$$\rho(x) = \lambda e^{-\lambda x}, x > 0, \lambda > 0$$

has

$$\sigma(\alpha)^2 = \frac{1}{\lambda^2} \frac{1+\alpha}{1-\alpha} \tag{4.37}$$

giving $f(\alpha) \propto \bar{\alpha}^{1/2}$ at the $\alpha \to 1$ limit. See Figure 4.3a.

Example 4.19. The Gaussian distribution with PDF

$$\rho(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

has

$$\sigma(\alpha)^2 = \frac{1}{\bar{\alpha}^2} \left(1 + \alpha\phi(a)^2 + \frac{\rho(\phi(\alpha))}{\bar{\alpha}} ((1-2\alpha)\phi(\alpha) - \rho(\phi(\alpha))) \right) \tag{4.38}$$

where

$$\phi(\alpha) = F_X^{-1}(\alpha) = \sqrt{2}\,\mathrm{erf}^{-1}(2\alpha - 1)$$

is its quantile function, where erf is the error function defined by

$$\mathrm{erf}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt \tag{4.39}$$

At the $\alpha \to 1$ limit, this gives

$$f(\alpha) \approx \frac{\sqrt{\bar{\alpha}}}{\operatorname{erfc}^{-1}(2\bar{\alpha})}. \tag{4.40}$$

Here, erfc is the complementary error function, defined by

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) \tag{4.41}$$

See Figure 4.3b.

Example 4.20. If X is bounded above, shift it so that $\operatorname{ess\,sup}(X) = 0$. Then suppose it has a PDF ρ,

$$\rho(x) \approx A(-x)^n, x \in (-\epsilon, 0)$$

where $n > -1$, and A being an unspecified positive constant, then at the $\alpha \to 1$ limit, approximately

$$f(\alpha) \propto \bar{\alpha}^{\frac{n+3}{2n+2}} \tag{4.42}$$

Notably, when $n = 1$, $f(\alpha) \propto \bar{\alpha}$, so $\sigma(\alpha) \propto 1$ in fact converges to a positive constant. This is shown in Figure 4.3c.

When $n > 1$, $\sigma(\alpha)$ diverges to infinity, and when $-1 < n < 1$, it converges to 0. For example, when X is uniform, it has $n = 1$, and indeed by Equation 4.21, $\lim_{\alpha \to 1} \sigma(\alpha) = 0$.

Example 4.21. A particularly flat $\sigma(\alpha)$ that we found is shown in Figure 4.3f, defined by a real random variable with PDF $\rho(x) = -xe^{x^2/2}$, with $x < 0$. It gives

$$f(\alpha) = 2(\bar{\alpha} + \alpha^2 F(\sqrt{\ln \alpha})^2) \tag{4.43}$$

where F is the Dawson F function defined by

$$F(x) = e^{-x^2} \int_0^x e^{t^2} dt. \tag{4.44}$$

4.3 Limit theorems for other risk measures

It is natural to ask whether there are limit theorems for more general risk measures than CVaR. However, a literature search turned up nothing. As such, we believe the following results are new.

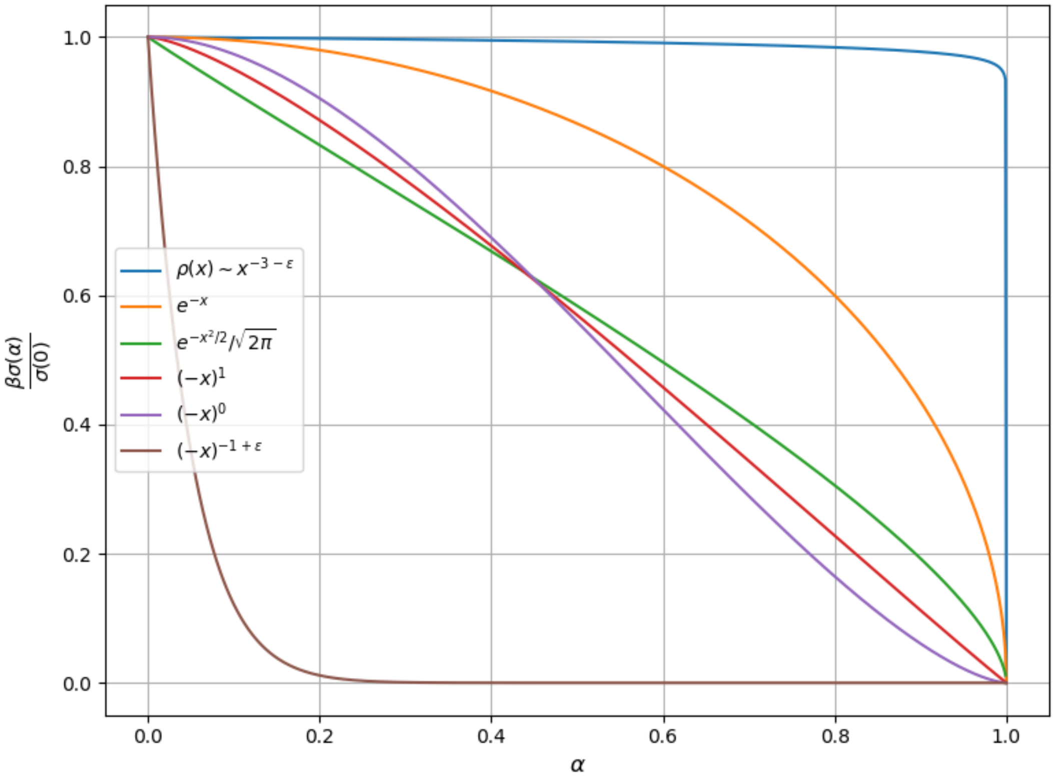


Figure 4.6: Schematic drawing of the tail behavior of $\bar{\alpha}\sigma(\alpha)$ for various distributions of X. In general, the thinner the tail, the faster it converges to 0.
Each legend lists the right tail of the density the X corresponding to each curve. From top to bottom, the right tails of X and of $\bar{\alpha}\sigma(\alpha)$ grow thinner together.

4.3.1 SLLN for spectral risk measures

When X is bounded above and below, the uniform SLLN for CVaRcan be extended to all spectral risk measures:

Theorem 4.22 (uniform SLLN for spectral risk measures)**.** *Let $X \in \mathscr{L}^\infty$, that is, it is a random variable with bounded range. Then almost surely, for any probability measure m on $[0,1]$, the spectral risk measure $\mathcal{F} = \int_0^1 \mathrm{CVaR}_\alpha \, dm(\alpha)$ satisfies a SLLN:*

$$\mathcal{F}(L_n) \to \mathcal{F}(X) \tag{4.45}$$

Proof. Let $M > |X|$ be an upper bound of X, then by monotonicity of CVaR,

$$-M < \mathrm{CVaR}_\alpha(L_n) < M, \ |\mathrm{CVaR}_\alpha(L_n)| \leq M$$

By theorem 4.15, almost surely, $\mathrm{CVaR}_\alpha(L_n)$ converges pointwise (with respect to α) to $\mathrm{CVaR}_\alpha(X)$. Then since $\int M dm(\alpha) = M$ is finite, by Lebesgue's dominated convergence theorem,

$$\lim_n \mathcal{F}(L_n) = \lim_n \int \mathrm{CVaR}_\alpha(L_n) dm(\alpha) = \int \mathrm{CVaR}_\alpha(X) dm(\alpha) = \mathcal{F}(X)$$

□

4.3.2 CLT for spectral risk measures?

In the proof of SLLN for spectral risk measures, the crucial step is using a SLLN of CVaR_α that holds uniformly over all $\alpha \in [0,1]$. Analogously, we suspect that there is a CLT for spectral risk measures that depends on a uniform CLT, similar to results collected in [Dud99].

In particular, we suspect that a proof can be found through the generalized Berry–Esseen Theorem for CVaR, as [GW11, Theorem 1.1]:

Theorem 4.23 (Berry–Esseen theorem for CVaR)**.** *Given real random variable X with finite third moment, $\alpha \in (0,1)$, such that $\sigma(\alpha) > 0$, and X has a strictly positive, continuous PDF in a neighborhood of $F^{-1}(\alpha)$, then there exists some $C_\alpha > 0$ such that*

$$\|G_n - \Phi\|_\infty \leq \frac{C_\alpha}{\sqrt{n}} \tag{4.46}$$

for all $n \in \mathbb{N}$, where G_n is the CDF of the random variable

$$\frac{\sqrt{n}}{\sigma(\alpha)}(\mathrm{CVaR}_\alpha(L_n) - \mathrm{CVaR}_\alpha(X))$$

However, using the Berry–Esseen theorem, we were unable to prove the suspected generalization to the CLT of CVaR, so we leave it as a conjecture:

Conjecture 4.24 (CLT for spectral risk measures). *For any probability measure m on $[0,1]$, define a spectral risk measure $\mathcal{F} = \int_0^1 \mathrm{CVaR}_\alpha \, dm(\alpha)$, then for any real random variable X with finite variance,*

$$\sqrt{n}(\mathcal{F}(L_n) - \mathcal{F}(X)) \xrightarrow{d} \mathcal{N}(0,\sigma) \tag{4.47}$$

for some $\sigma \geq 0$, provided that there does not exist some $\alpha_0 \in [0,1]$, such that $m(\alpha_0) > 0$, and $\sigma(\alpha)$ is discontinuous at α_0, where $\sigma(\alpha)$ is the standard deviation of the limit distribution in Equation 4.24:

$$\sqrt{n}(\mathrm{CVaR}_\alpha(L_n) - \mathrm{CVaR}_\alpha(X)) \xrightarrow{d} \mathcal{N}\left(0, \sigma(\alpha)\right)$$

Otherwise, there exists $\sigma_1 > \sigma_2 \geq 0$ such that

$$\sqrt{n}(\mathcal{F}(L_n) - \mathcal{F}(X)) \xrightarrow{d} \mathcal{N}_{mixed}(0, \sigma_1, \sigma_2) \tag{4.48}$$

4.3.3 CLT for entropic value at risk?

Other than CVaR, another example of CRM is the entropic value at risk (EVaR), proposed in [Ahm12].

Definition 4.25. For any $\alpha \in [0,1]$, and real random variable X, its α-level EVaRis

$$\mathrm{EVaR}_\alpha(X) = \inf_{t>0} \frac{1}{t} \ln \mathbb{E}\left(\frac{1}{1-\alpha} e^{tX}\right) \tag{4.49}$$

The entropic value at risk (EVaR) is similar to CVaR, and as such we have reasons to suspect it to have a similar CLT. We attempted to calculate the cumulant generating function for the empirical process of EVaR, but without success.

While having no definitive proof, numerical simulation strongly suggests that there exists a similar CLT for EVaR:

Conjecture 4.26 (CLT for EVaR). *Given any X with finite variance, and any $\alpha \in [0,1)$, its empirical process L_n satisfies*

$$\sqrt{n}(\mathrm{EVaR}_\alpha(L_n) - \mathrm{EVaR}_\alpha(X)) \xrightarrow{d} \mathcal{N}(0, \sigma(\alpha)) \tag{4.50}$$

where $\sigma(\alpha)$ is a continuous function satisfying $\sigma(0)^2 = \mathbb{V}(X)$.

Figure 4.7 plots the mean and standard deviation (normalized by $\sqrt{n}$) of the $\mathrm{EVaR}_\alpha(L_n)$, while Figure 4.8 shows that as $n \to \infty$, the distribution of $\mathrm{EVaR}_\alpha(L_n)$ becomes closer to normal distribution.

As shown in Figure 4.7, numerical calculation does not reveal any discontinuity of $\sigma(\alpha)$, akin to that of CVaR, suggesting that for EVaR, the CLT holds for all $\alpha \in [0, 1)$.

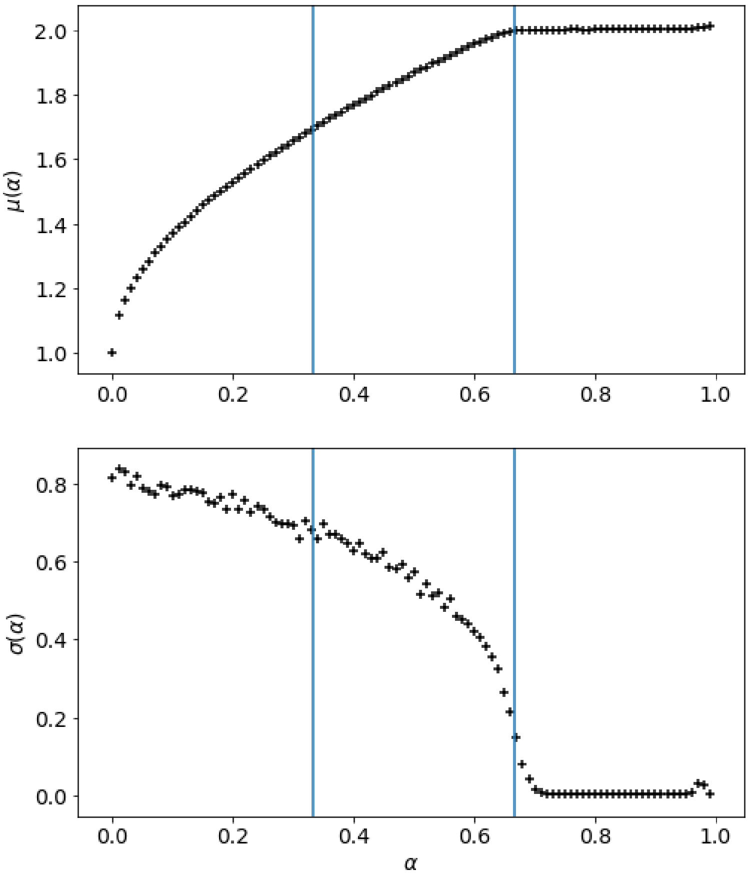


Figure 4.7: Mean and standard deviation (normalized by $\sqrt{n}$) of $\mathrm{EVaR}_\alpha(L_n)$ plotted as functions of α, with $n = 1000$, and X being the uniform distribution on $\{0, 1, 2\}$. Each point is calculated from 1000 trials. The two vertical lines denote $\alpha = 1/3, 2/3$ respectively. As apparent from the graph, σ is a continuous function over $[0, 1]$, and equals zero for $\alpha > 2/3$. The blip at the right end of $\sigma(\alpha)$ is due to numerical instability of the root-finding algorithm, which is required in the calculation of EVaR, as it involves searching for the infimum of a transcendental function, an infimum with no closed form. (see Equation 4.49).

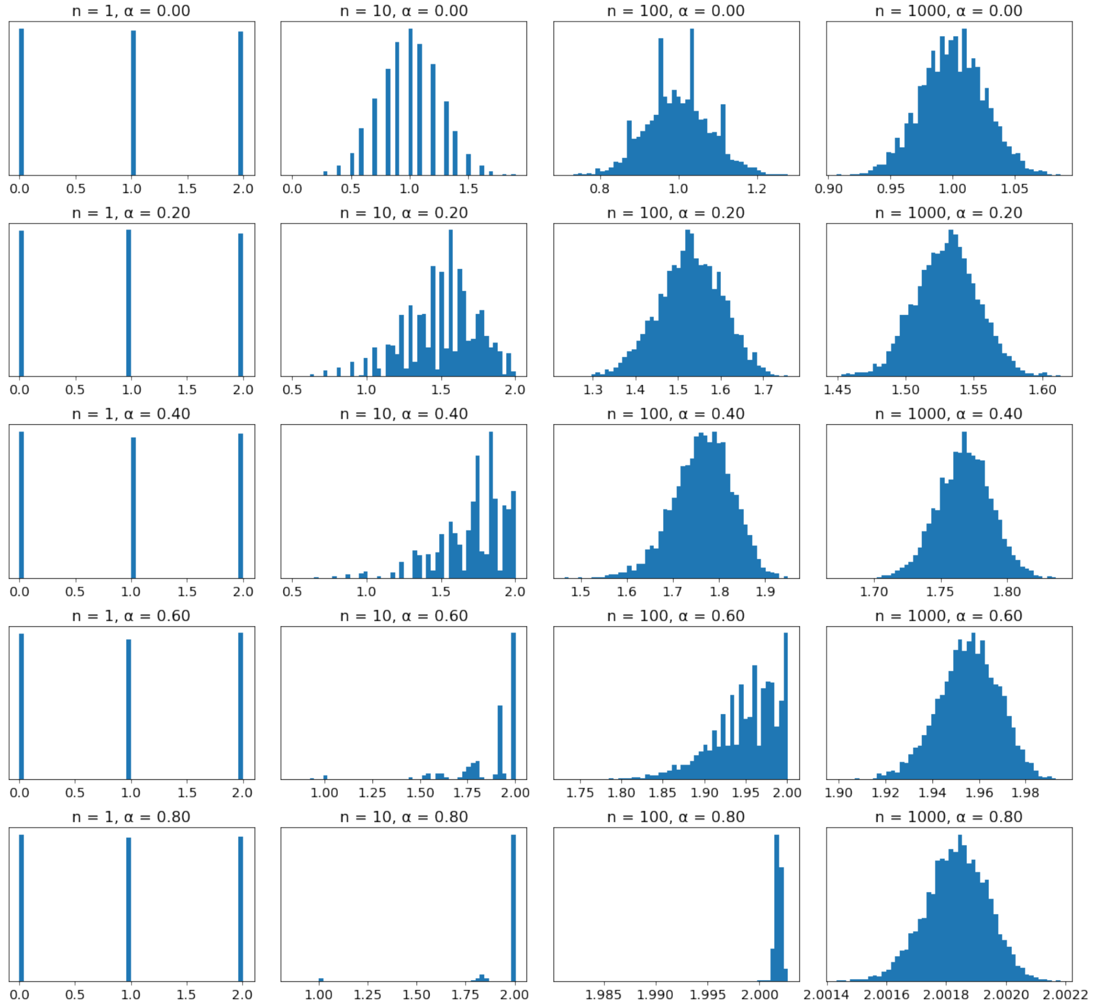


Figure 4.8: A demonstration of the CLT for EVaR. Here, X is the uniform distribution on $\{0, 1, 2\}$, and the PDF of $\text{EVaR}_\alpha(L_n)$ plotted as a function of n and α. As n increases, the distributions converge to normal distributions. Note that when $\alpha = 0.8$, the distribution becomes close to degenerate, as expected if $\sigma(\alpha) = 0$ when $\alpha > 2/3$. Each histogram is the result of 5000 trials. The x-axis is shifted and scaled in each plot to make the bell-shape apparent.

Chapter 5

Conclusion

In this chapter, we enumerate our main results, conjectures, sketch out applications to machine learning, and further research directions.

5.1 Summary of results and conjectures

Unless otherwise noted, all generalizations that follow are from expectation to coherent risk measures.

In Chapter 2, we showed:

1. The geometric-analytic correspondence of risk measures with their envelope representations (Proposition 2.36).

2. The formula (without proof) for envelope representation of CVaR (Equation 2.23).

3. Kusuoka representation theorem on finite uniform sample spaces (Section 2.4.1).

4. Counterexamples to Kusuoka representation theorem on finite nonuniform sample spaces (Section 2.4.2).

In Chapter 3, we showed:

1. Generalizations of (Section 3.1).

2. Generalizations of concentration inequalities (Section 3.2).

3. Generalizations, from expectation to spectral risk measures, of basic concepts (Section 3.3.3) and the fundamental theorem (Theorem 3.31) in statistical learning theory.

We conjectured the CVaR law of total expectations (Conjecture 3.15).

In Chapter 4, we showed:

1. The uniform strong law of large numbers (SLLN) for spectral risk measures (Theorem 4.22), whic subsumes the SLLN for spectral risk measures and the SLLN for CVaR (Theorem 4.15).
2. The central limit theorem (CLT) for CVaR (Theorem 4.11).
3. Examples of the CLT for CVaR of exemplar random variables (Section 4.2.3).

We conjectured:

1. A “mixed” CLT for CVaR (Conjecture 4.13).
2. A CLT for spectral risk measures (Conjecture 4.24).
3. A CLT for entropic value at risk (Conjecture 4.26).

5.2 Machine learning applications

The stated goal of the thesis is to investigate consequences of generalizing probability theory by replacing expectation with coherent risk measures, especially CVaR, but such general investigations are not the original motivation of the authors. We were drawn to this topic from considering the use of probability in machine learning.

As in the very first page of the thesis, many machine learning problems can be cast into the form of risk minimization. Risk is often defined as the expectation of loss. By replacing expectation with CVaR or other coherent risk measures, we can obtain a risk-management tuning knob on machine learning algorithms.

In [TGM15], a stochastic gradient descent algorithm for $\mathrm{CVaR}_{0.05}$ is used to train a Tetris-playing program that, instead of minimizing the expectation of loss (the negative of score), minimizes the $\mathrm{CVaR}_{0.05}$ of loss. It was found that, compared to a loss-minimizing agent, this agent was less likely to attempt high-risk high-reward Tetris maneuvers.

In [Cho+15], the authors described algorithms for solving Markov Decision Problems, optimized to minimize the CVaR risk measure, instead of the expectation. They derived theoretical error bounds to their algorithms, and tested the algorithm in a car navigating through a field with obstacles. As expected, a CVaR-optimizing agent drove more cautiously than the original expectation-optimizing agent.

In [CG14], the authors proposed the actor-critic algorithm for reinforcement learning, so that it minimizes the CVaR risk measure, instead of the expectation. They proved its convergence properties, and applied it to an investment problem. Its behavior was found to be more risk-averse than the original expectation-optimizing algorithm.

In [MP09], the empirical risk minimization (ERM) learning algorithm is generalized by adding to the expectation of loss with a variance term, essentially generalizing ERM by replacing expectation with a non-convex risk measure. [ND17] continues the work by proposing a similarly non-convex risk measure that is more computationally efficient, has faster rates of convergence than empirical risk minimization. The authors also demonstrated its performance on two classification problems.

In [Maj+17], CRM is applied to the problem of inverse reinforcement learning, wherein a learning agent observed humans playing a driving game, and inferred what preferences humans have from their behavior. It was found that each human's behavior was well-modeled by minimization of their own CRM, which differs between humans. Some humans have CRM that are very sensitive to variations, which corresponds to their highly risk-averse driving behavior. Other humans have CRM close to expectation, which corresponds to their highly risk-neutral driving behavior.

In [WM19b], the problem of fair machine learning is discussed. The authors proposed to define *fairness risk measures*, which are special cases of CRM, and demonstrated a tradeoff between fairness and accuracy, that is, a supervised-learning algorithm minimizing fairness risk, instead of loss expectation, gradually became less accurate as the fairness risk increasingly weights fairness over loss minimization. See also Section 5.3 of the same paper for more technical applications of CRM in machine learning.

5.3 Further research directions

We propose that future research in general risk measures can include:

1. The resolution of conjectures proposed in the paper.
2. Generalization of more advanced probabilistic and statistical inequalities, such as McDiarmid's inequality, to general risk measures.

3. Demonstration of benefit of using general risk measures in a practical, large-scale machine-learning application.

4. Collaboration between experts on the many aspects of risk, such as financial mathematicians, engineers with expertise in reliability engineering, legal theorists who deal with risks in law (such as tort law), machine learning practitioners, and psychologists who study human perceptions of risk.

Bibliography

[09] *The Risks of Financial Modeling: VaR and the Economic Meltdown, Testimony before the Subcommittee on Investigations and Oversight Committee on Science and Technology US House of Representatives.* In collab. with J. G. Rickards and N. N. Taleb. Sept. 10, 2009.

[15] *PS21 Survey: Experts See Increased Risk of Nuclear War.* PS21 - Project for the Study of the 21st Century, Nov. 12, 2015.

[Ace02] Carlo Acerbi. “Spectral Measures of Risk: A Coherent Representation of Subjective Risk Aversion”. In: 2002. DOI: 10/bzjfsg.

[AF19] Amir Ahmadi-Javid and Malihe Fallah-Tafti. “Portfolio Optimization with Entropic Value-at-Risk”. In: *European Journal of Operational Research* (2019). DOI: 10/gf66sn.

[Ahm12] Amir Ahmadi-Javid. “Entropic Value-at-Risk: A New Coherent Risk Measure”. In: *Journal of Optimization Theory and Applications* 155.3 (2012), pp. 1105–1123.

[Art+99] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. “Coherent Measures of Risk”. In: *Mathematical finance* 9.3 (1999), pp. 203–228. DOI: 10/bzjp2p.

[Art15] Emil Artin. *The Gamma Function.* Dover edition. Dover Books on Mathematics. Mineola, New York: Dover Publications, Inc, 2015.

[AT02] Carlo Acerbi and Dirk Tasche. “On the Coherence of Expected Shortfall”. In: *Journal of Banking & Finance* 26.7 (2002), pp. 1487–1503. DOI: 10/ftptzh.

[BBL03] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. “Introduction to Statistical Learning Theory”. In: *Summer School on Machine Learning*. Springer, 2003, pp. 169–207.

[Bil12] Patrick Billingsley. *Probability and Measure*. Anniversary ed. Wiley Series in Probability and Statistics. Hoboken, N.J: Wiley, 2012.

[BO99] Carl M Bender and Steven A Orszag. *Advanced Mathematical Methods for Scientists and Engineers. 1, 1,* New York: Springer, 1999.

[Bra+08] Vytaras Brazauskas, Bruce L. Jones, Madan L. Puri, and Ričardas Zitikis. "Estimating Conditional Tail Expectation with Actuarial Applications in View". In: *Journal of Statistical Planning and Inference*. Special Issue in Honor of Junjiro Ogawa (1915 - 2000): Design of Experiments, Multivariate Analysis and Statistical Inference 138.11 (Nov. 1, 2008), pp. 3590–3604. DOI: 10/cbwtgb.

[Bri19] R. A. Briggs. "Normative Theories of Rational Choice: Expected Utility". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2019. Metaphysics Research Lab, Stanford University, 2019.

[BS08] Nick Bostrom and Anders Sandberg. *Global Catastrophic Risks Survey*. Technical Report #2008-1. Future of Humanity Institute, Oxford University, 2008, pp. 17–20.

[CFS13] Louis H. Y. Chen, Xiao Fang, and Qi-Man Shao. "From Stein Identities to Moderate Deviations". In: *The Annals of Probability* 41.1 (Jan. 2013), pp. 262–293. DOI: 10/f4nxrg.

[CG14] Yinlam Chow and Mohammad Ghavamzadeh. "Algorithms for CVaR Optimization in MDPs". In: *Advances in Neural Information Processing Systems*. 2014, pp. 3509–3517.

[CG84] J. Theodore Cox and David Griffeath. "Large Deviations for Poisson Systems of Independent Random Walks". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 66.4 (Sept. 1984), pp. 543–558. DOI: 10/bv6b2w.

[Che07] Song Xi Chen. "Nonparametric Estimation of Expected Shortfall". In: *Journal of financial econometrics* 6.1 (2007), pp. 87–107. DOI: 10/frxvzq.

[Che14] James Ming Chen. "Measuring Market Risk under the Basel Accords: VaR, Stressed VaR, and Expected Shortfall". In: *Stressed VaR, and Expected Shortfall (March 19, 2014)* 8 (2014), pp. 184–201.

[CHK13] Zengjing Chen, Kun He, and Reg Kulperger. “Risk Measures and Nonlinear Expectations”. In: *Journal of Mathematical Finance* 03.03 (2013), pp. 383–391. DOI: 10.4236/jmf.2013.33039.

[Cho+15] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. “Risk-Sensitive and Robust Decision-Making: A Cvar Optimization Approach”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 1522–1530.

[Chu01] Kai Lai Chung. *A Course in Probability Theory*. 3rd ed. San Diego: Academic Press, 2001.

[Dan+01] Jon Danielsson, Paul Embrechts, Charles Goodhart, Con Keating, Felix Muennich, Olivier Renault, and Hyun Song Shin. “An Academic Response to Basel II”. In: *Special Paper-LSE Financial Markets Group* (2001).

[Den08] Frank Den Hollander. *Large Deviations*. American Mathematical Society, 2008.

[Dia13] Jared M. Diamond. *The World until Yesterday: What Can We Learn from Traditional Societies?* New York: Penguin Books, 2013.

[Dud99] R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics 63. New York: Cambridge University Press, 1999.

[Dur10] Richard Durrett. *Probability: Theory and Examples*. 4th ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, New York: Cambridge University Press, 2010.

[DZ09] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. 2nd ed. 1998. 2nd printing 2009 edition. Berlin Heidelberg: Springer, Nov. 17, 2009.

[Fis89] Peter C. Fishburn. “Retrospective on the Utility Theory of von Neumann and Morgenstern”. In: *Journal of Risk and Uncertainty* 2.2 (1989), pp. 127–157.

[GB09] Gerd Gigerenzer and Henry Brighton. “Homo Heuristicus: Why Biased Minds Make Better Inferences”. In: *Topics in cognitive science* 1.1 (2009), pp. 107–143. DOI: 10/ch6cnf.

[Gia06] Emanuela Rosazza Gianin. “Risk Measures via G-Expectations”. In: *Insurance: Mathematics and Economics* 39.1 (2006), pp. 19–34. DOI: 10/cvgbq3.

[Gig07] Gerd Gigerenzer. *Gut Feelings: The Intelligence of the Unconscious.* Penguin Books, 2007.

[GW11] Fuqing Gao and Shaochen Wang. “Asymptotic Behavior of the Empirical Conditional Value-at-Risk”. In: *Insurance: Mathematics and Economics* 49.3 (2011), pp. 345–352. DOI: 10/d7x7sd.

[HL01] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Grundlehren Text Editions. Berlin ; New York: Springer, 2001.

[Kah11] Daniel Kahneman. *Thinking, Fast and Slow.* Macmillan, 2011.

[KS06] Levente Kocsis and Csaba Szepesvári. “Bandit Based Monte-Carlo Planning”. In: *In: ECML-06. Number 4212 in LNCS*. Springer, 2006, pp. 282–293.

[Kus01] Shigeo Kusuoka. “On Law Invariant Coherent Risk Measures”. In: *Advances in Mathematical Economics*. Springer, 2001, pp. 83–95.

[Maj+17] Anirudha Majumdar, Sumeet Singh, Ajay Mandlekar, and Marco Pavone. “Risk-Sensitive Inverse Reinforcement Learning via Coherent Risk Models.” In: *Robotics: Science and Systems*. 2017. DOI: 10/gf6rzt.

[Mar52] Harry Markowitz. “Portfolio Selection”. In: *The Journal of Finance* 7.1 (1952), pp. 77–91.

[MP09] Andreas Maurer and Massimiliano Pontil. “Empirical Bernstein Bounds and Sample Variance Penalization”. In: *arXiv preprint arXiv:0907.3740* (2009).

[ND17] Hongseok Namkoong and John C. Duchi. “Variance-Based Regularization with Convex Objectives”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 2971–2980.

[NR15] Nilay Noyan and Gábor Rudolf. “Kusuoka Representations of Coherent Risk Measures in General Probability Spaces”. In: *Annals of Operations Research* 229.1 (June 1, 2015), pp. 591–605. DOI: 10/f7chjp.

[PR08] Georg Ch Pflug and Werner Romisch. *Modeling, Measuring and Managing Risk*. Hackensack, N.J: World Scientific Pub Co Inc, 2008.

[RFR10] H. L Royden, Patrick Fitzpatrick, and H. L Royden. *Real Analysis*. Boston: Prentice Hall, 2010.

[Rob70] Herbert Robbins. “Statistical Methods Related to the Law of the Iterated Logarithm”. In: *The Annals of Mathematical Statistics* 41.5 (Oct. 1970), pp. 1397–1409. DOI: 10/dg9726.

[RS65] Herman Rubin and Jayaram Sethuraman. “Probabilities of Moderate Deviations”. In: *Sankhyā: The Indian Journal of Statistics, Series A* (1965), pp. 325–346.

[RU02] R. Tyrrell Rockafellar and Stanislav Uryasev. “Conditional Value-at-Risk for General Loss Distributions”. In: *Journal of banking & finance* 26.7 (2002), pp. 1443–1471. DOI: 10/c8fxjb.

[SB14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.

[Sun07] Cass R. Sunstein. “The Catastrophic Harm Precautionary Principle”. In: *Issues in Legal Scholarship* 6.3 (2007). DOI: 10/dmnrqh.

[Tal12] Nassim Nicholas Taleb. *Antifragile: Things That Gain from Disorder*. Random House Incorporated, 2012.

[Tao10] Terence Tao. “254A - Random Matrices, Notes 0: A Review of Probability Theory”. Lecture notes. UCLA, Jan. 2, 2010.

[TGM15] Aviv Tamar, Yonatan Glassner, and Shie Mannor. “Optimizing the CVaR via Sampling”. In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015.

[TK74] Amos Tversky and Daniel Kahneman. “Judgment under Uncertainty: Heuristics and Biases”. In: *Science* 185.4157 (1974), pp. 1124–1131. DOI: 10/gwh.

[Val09] Leslie G. Valiant. “Evolvability”. In: *Journal of the ACM (JACM)* 56.1 (2009), p. 3. DOI: 10/bqtzrn.

[Val84] Leslie G. Valiant. “A Theory of the Learnable”. In: *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*. ACM, 1984, pp. 436–445.

[Van00] Aad W. Van der Vaart. *Asymptotic Statistics*. Cambridge university press, 2000.

[Vap00] Vladimir Naumovich Vapnik. *The Nature of Statistical Learning Theory*. 2nd ed. New York: Springer, 2000.

[VC71] V. N. Vapnik and A. Ya. Chervonenkis. "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities". In: *Theory of Probability & Its Applications* 16.2 (1971), pp. 264–280. DOI: 10/bkrnds.

[Wei09] Martin L. Weitzman. "On Modeling and Interpreting the Economics of Catastrophic Climate Change". In: *The Review of Economics and Statistics* 91.1 (2009), pp. 1–19. DOI: 10/d5jkrf.

[WM19a] Robert C. Williamson and Aditya Krishna Menon. "Fairness Risk Measures". Working Draft. ANU, Feb. 2019.

[WM19b] Robert C. Williamson and Aditya Krishna Menon. "Fairness Risk Measures". In: *arXiv preprint arXiv:1901.08665* (2019).

[ZU16] Michael Zabarankin and Stan Uryasev. *Statistical Decision Problems*. Springer, 2016.