Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea Magistrale in Matematica

Formalizing information theory in Lean 4: divergences, hypothesis testing and the data processing inequality

Tesi di Laurea Magistrale in Teoria dell'Informazione

Relatore:
Chiar.mo Prof.
ANDREA PASCUCCI

Correlatore:
Dott.
RÉMY DEGENNE

Presentata da:
LORENZO LUCCIOLI

Anno Accademico 2023-2024

A mamma e papà,
a Stella.

Sommario

Questa tesi presenta la formalizzazione di risultati chiave del progetto "Lower bounds for hypothesis testing based on information theory" utilizzando l'interactive theorem prover Lean 4. Il lavoro si concentra sulla formalizzazione di strumenti di teoria dell'informazione, con particolare enfasi sulle divergenze e le loro applicazioni all'hypothesis testing. Il contributo principale è la formalizzazione di una versione generale della disuguaglianza di data processing (DPI) per le f-divergenze, un risultato fondamentale in teoria dell'informazione. Vengono confrontate tre diverse dimostrazioni della DPI, ciascuna con diversi gradi di generalità e ipotesi; la terza dimostrazione è la più generale e mette in evidenza la connessione tra la DPI e il problema di hypothesis testing. Oltre ai contributi teorici, la tesi riflette sulle scelte d'implementazione, sulle sfide affrontate e sulle intuizioni emerse durante il processo di formalizzazione. Questo lavoro sottolinea l'utilità degli interactive theorem prover nell'affrontare problemi matematici complessi, offrendo nuove prospettive nei campi della probabilità e della teoria dell'informazione.

Abstract

This thesis presents the formalization of key results from the "Lower bounds for hypothesis testing based on information theory" project using the Lean 4 interactive theorem prover. The work focuses on formalizing tools from information theory, with a particular emphasis on information divergences and their applications to hypothesis testing. The main contribution is the formalization of a general version of the data processing inequality (DPI) for f-divergences, a fundamental result in information theory. We compare three different proofs of the DPI, each with varying levels of generality and assumptions; the third proof is the most general, and highlights the connection between the DPI and hypothesis testing. In addition to the theoretical contributions, the thesis reflects on the design decisions, challenges, and insights gained during the formalization process. This work underscores the utility of interactive theorem proving in tackling complex mathematical problems, offering new perspectives in probability and information theory.

Contents

Introduction

Although not yet a mainstream practice within the mathematical community, the formalization and verification of mathematical proofs have attracted increasing attention in recent years, particularly with the advent of interactive theorem provers. These tools serve not only to guarantee the correctness of proofs but also to facilitate the generalization and application of mathematical results across a range of domains. One such tool, Lean 4, has gained significant traction within the mathematical community, offering a robust platform for formalizing complex mathematical theories.

This thesis presents a description of a part of the formalization project titled "Lower bounds for hypothesis testing based on information theory"[1], aimed at formalizing results about information divergences and their applications to hypothesis testing in Lean 4. To achieve this, the formalization of several tools from information theory was necessary. Notably, some results from the project have already been merged in an open-source library, and further contributions are planned in the near future. The central result of this thesis is the formalization of a very general version of the data processing inequality (DPI), a fundamental result in information theory. Some space will also be dedicated to the discussion of the design choices taken during the formalization process, as well as the challenges encountered and the lessons learned. The work is presented in five chapters, each focusing on a different aspect of the project.

The initial chapter introduces the concept of interactive theorem provers, with a particular emphasis on Lean 4. We first look at how a proof in Lean looks like, and how to interact with the proof assistant, we briefly discuss the logical foundations of Lean, then we describe its mathematical

[1]This work was carried out within the Scool team at the Inria center of the University of Lille, under the supervision of Rémy Degenne (Univ. Lille, Inria, CNRS, Centrale Lille, CRIStAL). See `https://remydegenne.github.io/testing-lower-bounds/blueprint/` for a description of the project. We will frequently refer to the code, that can be found in the following repository: `https://github.com/RemyDegenne/testing-lower-bounds`.

library (Mathlib), and the reasons why proof assistants can be beneficial for mathematical research.

The second chapter serves to establish the fundamental concepts that are essential for a comprehensive understanding of the subsequent material. This includes a discussion of transition kernels, the Lebesgue decomposition, and a version of the integration by parts formula.

The third chapter is dedicated to information divergences, wherein a comprehensive overview of f-divergences, the Kullback-Leibler divergence, and other measures of dissimilarity between probability distributions is provided. These divergences play a pivotal role in numerous applications of information theory. We approach their analysis from a highly general standpoint, avoiding the limitation of focusing solely on probability measures and instead considering a broader range of measures.

The fourth chapter takes a brief detour to discuss hypothesis testing, a fundamental problem in statistics and information theory. We formalize key concepts such as risk, and within this framework we define another information divergence, called statistical information, which plays a crucial role in a proof of the DPI.

The fifth chapter represents a pivotal contribution to this thesis, wherein three distinct proofs of the data processing inequality for f-divergences are presented. The DPI is a central result in information theory, stating that the divergence between two measures cannot increase when we apply a (possibly random) transformation. The initial proof has weak hypotheses but is limited to deterministic kernels. The second proof is a generalization of the first, extending its scope to Markov kernels but necessitating stronger hypotheses. The third proof also covers Markov kernels but employs a distinct approach that circumvents some of the assumptions of the second one.

The aim of this thesis is to demonstrate the power of interactive theorem proving in formalizing complex mathematical concepts and to contribute to the existing body of knowledge by providing new insights and generalizations, particularly within the context of probability and information theory.

Notation

- Let $\mathbb{R}$ be the real numbers, we indicate with $\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{+\infty, -\infty\}$ the extended real numbers and with $\overline{\mathbb{R}}_+ \coloneqq [0, +\infty) \cup \{+\infty\}$ the extended non-negative real numbers.

- Let $\mathcal{X}$ be a set and $\mathscr{F}_\mathcal{X}$ a σ-algebra on $\mathcal{X}$, then we say that $(\mathcal{X}, \mathscr{F}_\mathcal{X})$ is a measurable space. We often avoid writing the σ-algebra explicitly when it is not necessary.

- Let $\mathcal{X}$ be a measurable space, we denote by $\mathcal{M}(\mathcal{X})$ the space of measures on $\mathcal{X}$, and by $\mathcal{P}(\mathcal{X})$ the space of probability measures on $\mathcal{X}$.

- Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, $\mu \in \mathcal{M}(\mathcal{X})$, $A \subseteq \mathcal{X}$ a measurable set, and $f \colon \mathcal{X} \to \mathcal{Y}$ a measurable function. We denote the integral of f with respect to μ over A with one of the following notations: $\int_A f \,\mathrm{d}\mu$, $\int_A f(x)\mu(dx)$, $\int_A f(x) \,\mathrm{d}\mu(x)$. If the integral is over the entire space $\mathcal{X}$ we can also write $\mu[x \mapsto f(x)]$.

- Let $\mathcal{X}$ be a measurable space, $\mu \in \mathcal{M}(\mathcal{X})$, we say that a certain property holds *almost everywhere* with respect to μ (shortened μ-a.e.) if there exists a measurable set $A \subseteq \mathcal{X}$ such that this property holds for all $x \in A$ and $\mu(\mathcal{X} \setminus A) = 0$.

- Let $\mathcal{X}$ be a measurable space, $\mu \in \mathcal{M}(\mathcal{X})$, $f \colon \mathcal{X} \to \overline{\mathbb{R}}_+$ a measurable function, we denote by $f \cdot \mu$ the measure with density f with respect to μ, defined by $(f \cdot \mu)(A) \coloneqq \int_A f \,\mathrm{d}\mu$ for all measurable sets $A \subseteq \mathcal{X}$.

- Let $p \in [0, 1]$, we denote by $\mathrm{Ber}(p)$ the Bernoulli distribution with parameter p, that is the probability measure on $\{0, 1\}$ such that $\mathrm{Ber}(p)(\{0\}) = 1 - p$ and $\mathrm{Ber}(p)(\{1\}) = p$.

- Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, $\mu \in \mathcal{M}(\mathcal{X})$, $\nu \in \mathcal{M}(\mathcal{Y})$, we denote by $\mu \otimes \nu \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ the product measure of μ and ν. Moreover, if

$n \in \mathbb{N}$, we denote by $\mu^{\otimes n} = \mu \otimes \mu \otimes \cdots \otimes \mu$ the product of μ with itself n times.

- Let $(\mathcal{X}, \mathscr{F}_{\mathcal{X}})$ be a measurable space. We say that $\mathcal{X}$ is a standard Borel space if there exists a metric on $\mathcal{X}$ that generates $\mathscr{F}_{\mathcal{X}}$ as the Borel σ-algebra and such that $\mathcal{X}$ is complete and separable.

- Let $f\colon \mathcal{X} \to \mathcal{Y}$ be a function and $A \subseteq \mathcal{X}$, we denote by $f_{|A}\colon A \to \mathcal{Y}$ the restriction of f to A. Note that we can also consider the restriction of a measure $\mu \in \mathcal{M}(\mathcal{X})$ to a sub σ-algebra $\mathcal{A} \subseteq \mathscr{F}_{\mathcal{X}}$, using the same notation $\mu_{|\mathcal{A}}$.

Chapter 1

Interactive Theorem Provers

Interactive theorem provers, also referred to as proof assistants, are tools that facilitate the formalization and verification of mathematical proofs. They integrate elements of both programming and mathematics, enabling the construction of proofs in a rigorous, formal language. We focus our attention on Lean 4, since it is the proof assistant that we are using for our project.

1.1 Lean 4

Lean 4 is a functional programming language, but, more importantly for our purposes, it is also an interactive theorem prover, i.e., a software that allows the user to write mathematical definitions, statements, and proofs in a formal language, automatically checking their logical correctness and, in some cases, even assisting in the construction of the proofs.

The Lean theorem prover was initially developed by Leonardo de Moura [Mou+15] and first released in 2013. Its latest iteration, Lean 4, was released in 2021 [MU21] and has since been growing in popularity even among professional mathematicians, leading to some noteworthy formalization projects. These formalization efforts include completed projects such as the Liquid Tensor Experiment[1] inspired by Peter Scholze, and the PFR (Polynomial

[1] `https://leanprover-community.github.io/liquid/`. This was a Lean 3 project.

Freiman-Ruzsa Conjecture) project[2] led by Terence Tao, as well as recently started ones, like the FLT (Fermat's Last Theorem) project[3] led by Kevin Buzzard. Moreover, from 2023 the development and maintenance of Lean has been taken over by the recently established Lean FRO (Focused Research Organization)[4], with the aim of further developing the Lean ecosystem, improving its performance and making it more user-friendly.

1.2 Reading Lean code

Looking at a piece of code in Lean for the first time can be intimidating, especially for someone who is not accustomed to working with programming languages. This section aims to provide a concise overview of the practical aspects of utilizing Lean, as well as an explanation of how to interpret a Lean proof.

It is first necessary to bear in mind that Lean is a programming language, and therefore the main method of interacting with it is through a code editor. The most popular code editor for Lean is Visual Studio Code[5], which has a Lean extension that provides syntax highlighting, autocompletion, and other features that facilitate the writing of Lean code. However, there are other options available, such as the Lean web editor[6].

We will now examine an example of a definition in Lean, namely the definition of a measurable function.

```
def Measurable {α β : Type*} [MeasurableSpace α] [MeasurableSpace β]
    (f : α → β) : Prop :=
  ∀ (t : Set β), MeasurableSet t → MeasurableSet (f ⁻¹' t)
```

This piece of code defines what it means for a function f between two

[2] `https://teorth.github.io/pfr/`. See also [Tao23b; Tao23a].

[3] `https://imperialcollegelondon.github.io/FLT/`. This project has only recently started and is anticipated to take years to complete.

[4] `https://lean-fro.org`.

[5] `https://code.visualstudio.com`.

[6] `https://live.lean-lang.org`.

measurable spaces α and β to be measurable. We will now break down this definition in order to elucidate the meaning of each component.

- The first word is `def`, which is a keyword in Lean that is used to define new objects, this tells Lean that a new object is about to be defined. Other keywords similar to `def` are `structure`, which can be used for definitions that bundle together multiple objects or properties, and `class`, which also allows the object to be inferred by the typeclass inference system, that will be described shortly.

- The next word is `Measurable`, which is the name of the object that we are defining. At this stage, we may choose any name we wish; however, it is good practice to choose a name that is descriptive of what we are defining, since we will use it later to refer to this object.

The next part is a list of arguments, which serve as inputs for defining the object in question. In our example, each time we want to say that some function is measurable, Lean will require the following information: what are the domain and codomain of the function, the fact that they are measurable spaces, and the function itself. There are three different types of arguments, which can be differentiated based on the type of brackets used to enclose them.

- The first argument {α β : `Type*`} indicates that α and β, which will be used later as domain and codomain of the function, are two generic types; this is an implicit argument, as it is enclosed in curly braces { }, meaning that when we use this definition we will not need to specify these types explicitly. Arguments in curly braces are usually meant to be inferred from other arguments, for example in our case α and β can be inferred from the type of the function f. Throughout this thesis, we will sometimes omit some of the implicit arguments, when they can be readily understood from the context.

- The next two arguments [`MeasurableSpace` α] [`MeasurableSpace` β] indicate that α and β are measurable spaces. These arguments are also

implicit, but they are enclosed in square brackets [ ]; this means that Lean will attempt to infer them using the typeclass inference system, which is a mechanism that enables Lean to automatically find some properties that were previously proven in special theorems called instances. This same mechanism can also deduce certain properties based on existing instances, for example we will never need to specify that $\mathbb{R}$ is a topological space, because it can be inferred from the fact that it is a pseudo-metric space. See also [MU21; SUM20] for more information about typeclasses in Lean 4.

- The final argument (`f` : $\alpha \to \beta$) is the function that is to be deemed measurable; it is enclosed in parentheses ( ), meaning that it is an explicit argument, and we will need to specify it every time we employ this definition. For example if we want to say that the exponential function `exp` is measurable, we would write `Measurable exp`.

- The portion of the code situated between the colon : and the colon equal sign := represents the type of the object that is being defined. In this case, it is `Prop`, since our object is the proposition that the function f is measurable.

- The final component is the body of the definition, which represents the actual content of the proposition that we are defining. We can read it as follows: for every t subset of β ($\forall$ (`t` : `Set` β)), if t is measurable (`MeasurableSet t`), then the preimage of t under f is also measurable (`MeasurableSet` (`f` $^{-1}$' `t`)).

We can notice how in the Lean code we can use a wide range of symbols, including the Greek letters α, β and the universal quantifier $\forall$. Indeed, any Unicode character can be employed. This enables the code to resemble mathematical notation more closely, making it easier to read for mathematicians.

Let us look at another example, this time we consider a lemma with a short proof:

```
theorem add_sub_cancel (a b : ℝ) : b + a - b = a := by
  rw [add_comm]
  rw [add_sub_assoc]
  rw [sub_self]
  exact add_zero a
```

Once more, we deconstruct the code in order to understand its meaning:

- The keyword `theorem` at the beginning is used to start a new theorem. It can be substituted with `lemma` without altering the code's meaning, with `example` if we do not need to give it a name for later reference, or with `instance`, which enables the result to be used by the typeclass inference system.

- `add_sub_cancel` is the name that we are giving to the theorem, we will be able to use it in later code to refer to this theorem.

- Before the colon : we have the arguments of the theorem, that is the hypotheses that need to be verified for the theorem to hold. The same kinds of arguments that were present in the definition can be used here, with the same meaning. In this case, we only have two real numbers a and b as explicit arguments.

- After the colon : we have the type of the theorem. This is where the thesis of the theorem is presented, typically in the form of a proposition. In this case, the thesis is that the equality $b + a - b = a$ holds (for all real numbers a and b).

- Finally, after := `by` we have the proof of the theorem. This is a sequence of proof steps, each one transforming the goal into a simpler one, until we reach a goal that can be solved directly. The first step is `rw [add_comm]`, where `add_comm` is the name of a lemma stating the commutativity of addition, and `rw` is the rewrite tactic, that allows us to replace a part of the goal with an equivalent one. In this case it replaces $b + a$ with $a + b$,

transforming our goal from $b + a - b = a$ to $a + b - b = a$. Similarly, the next two lines use the lemmas `add_sub_assoc` (associativity of addition and subtraction) and `sub_self` (subtraction of a number from itself is zero) to further simplify the goal as follows:

$$a + b - b = a \quad \rightsquigarrow \quad a + (b - b) = a \quad \rightsquigarrow \quad a + 0 = a$$

The last line, `exact add_zero a`, uses the `exact` tactic to finish the proof, it instructs Lean that the goal is precisely equal to a previously proven lemma, in this case `add_zero` applied to the number a, which states that adding zero to a number does not alter it.

It should be noted that this proof can be written in a more concise manner, for example we could have combined multiple rewrites in a single line: `rw [ add_comm, add_sub_assoc, sub_self]`. Or we could even have written the entire proof as a single command using the tactic `ring`, which automatically solves equations using the properties of a commutative ring.

Another thing to notice is that this proof employs a backward reasoning approach, whereby we start from the goal, and simplify it until it becomes trivial. It is also possible to use forward reasoning, starting from the hypotheses and building the proof step by step, but this is less common in Lean.

It is important to mention that any code written in Lean is automatically verified by a dedicated software component, known as the kernel, which ensures its logical consistency. So if we make any mistake in our proof, the kernel will not accept it, and will provide an error message that can be used to identify the source of the issue.

Note that Lean will complain also if a proof is left unfinished. Nevertheless, it is possible to write a partial proof and write the keyword `sorry` at the end, this instructs Lean to accept the proof as it is (we can even omit the proof altogether), so that we can postpone its completion to a later time, but still use the result in other proofs.

Additionally, Lean can warn us that we are doing something that is not recommended, for instance if a hypothesis is added to a lemma and never used, or if a result is left with `sorry` in the proof, then the Lean linter will show a yellow squiggle beneath the line and indicate the nature of the issue.

Moreover, Lean not only verifies the soundness of the proof but also provides assistance in its construction. For instance, it keeps track of the hypotheses and goals throughout the proof. This information is always displayed to the user in a distinct window inside the editor, called the Lean Infoview. The Infoview shows us the current state of the proof at the cursor position, in particular it tells us what is the goal, what are the variables in the context, and what are the hypotheses available. For example, if we position the cursor at the end of the first line of the proof of the above theorem, the Infoview will appear as follows:

```
a b : ℝ
⊢ a + b - b = a
```

This tells us that the goal (after the symbol $\vdash$) is to prove that $a+b-b = a$, and that we have two real numbers a and b in the context.

1.3 Logical foundations of Lean

When doing mathematics, we must assume some axioms upon which to build our theory, and the same is true when using a proof assistant. The vast majority of mathematicians typically assume, more or less explicitly, the axioms of Zermelo-Fraenkel set theory with the axiom of choice (ZFC) as a foundation for their work. Lean, on the other hand, is based on dependent type theory, in particular on the Calculus of Constructions [CH88], a logical formalism also employed by other proof assistants, such as Coq/Rocq[7].

At this point, a reasonable concern could be whether proving a statement in Lean is equivalent to proving it in set theory, that is, can we trust the

[7] `https://coq.inria.fr`

results that we obtain from Lean? The answer is affirmative. Indeed, it is possible to build the usual set theory in Lean, and it has been proven that Lean's type theory is equiconsistent with a slightly strengthened version[8] of ZFC [Car19].

It is not necessary to know the details of type theory in order to use Lean, but it is useful to keep in mind that the elementary object in Type theory is a type, and that every object is a term (i.e. an element) of some type. Roughly speaking, we can think of types as sets, but there are some key differences from the usual set theory. For example, whereas an object may be an element of multiple sets simultaneously, in type theory a term can only have a single type. Consequently, if one wishes to treat a term as belonging to a different type (as is the case when a natural number is regarded as an element of the real numbers), it is necessary to employ a coercion, that is, a function from the original type to the new one. Fortunately, Lean is designed to make this process as smooth as possible, and it is often able to infer the correct coercion automatically, sometimes without the end user even noticing it. A further peculiarity of type theory is that propositions are types, and if we have a proposition P, then a term p of type P (written $P : Prop$ and $p : P$) is a proof of P. Furthermore, a proposition P can either have zero elements (if it is not provable) or one element (if it is provable), meaning that all the possible proofs of P are considered equivalent.

1.4 Mathlib

Writing mathematics in a proof assistant is not an easy task, and even if a proof on paper appears to be relatively simple or brief, the act of formalizing can be quite time-consuming. This is due in part to the necessity of meticulously and precisely articulating each definition and proof, without leaving any gaps, as the proof assistant will not take anything for granted. For this reason, it is crucial to have a method for storing and sharing the

[8] In particular ZFC + {there are n inaccessible cardinals | $n < \omega$}.

results that have already been proven, so that they can be reused directly in other proofs, both by the same author and by other users, without having to prove them again. This is the purpose of Mathlib[9], the mathematical library of Lean.

At the time of writing, Mathlib contains more than 1.5 million lines of code[10], encompassing the majority of the mathematics taught in an undergraduate course, as well as some more advanced topics. Mathlib is an open-source project, and its development is driven by the community and orchestrated by a team of maintainers. While contributions from any individual are welcome, the code must undergo a review process conducted by a maintainer before being merged into the main branch, and a high standard of quality is required. This approach allows the library to grow rapidly, while keeping a high level of coherence and maintainability.

One of the primary objectives of Mathlib is to provide lemmas that can be reused in a multitude of contexts. Consequently, a continuous effort is made to generalize the code as much as possible, trying to avoid superfluous assumptions. This approach also has the advantage of making the true scope of a mathematical result more evident. Furthermore, for each mathematical object that is defined, there is a corresponding set of lemmas, typically referred to as the API (Application Programming Interface) borrowing the term from computer science, that are proven about it and facilitate the use of that object in practice.

The use of Mathlib is essential for every Lean user who plans to formalize any non-trivial mathematical theory, and as such it is also the basis of our work. In particular, we rely heavily on the measure theory library, which provides a solid foundation for the formalization of probability theory with tools such as integrals, Radon-Nikodym derivatives, and disintegration of measures; we also use the library's section dedicated to convex functions. Furthermore, we make extensive use of tactics, that are commands in Lean

[9] `https://github.com/leanprover-community/mathlib4`.

[10] For some statistics about Mathlib, see `https://leanprover-community.github.io/mathlib_stats.html`.

designed to facilitate the construction of proofs, automating some specific tasks and making the proof process more efficient. For example, we use the tactic `measurability` to help us prove that a function or a set is measurable, `congr` to simplify the proof of equalities when some parts of the terms are identical, and `linarith` that automatically solves linear arithmetic problems.

Throughout this thesis, we will frequently cite statements and results from Mathlib. These can be found either in the online documentation[11] or directly in the source code[12].

1.5 Why proof assistants?

There are numerous reasons why one might choose to use a proof assistant to write mathematics, a collection of which can be found in [Avi24]. In this section, I will focus on three of them: ensuring the correctness of the proofs, facilitating the generalization of results, and enabling the coordination of multiple researchers on the same project.

The trust placed in mathematics is based on the assumption that the proofs are correct. This is typically ensured through the peer review process. However, this process is heavily reliant on manual checking conducted by humans, and occasionally small errors can slip through. For example, during our formalization project, a small inconsistency was identified[13] in the proof of Theorem 1 in the paper "Rényi Divergence and Kullback-Leibler Divergence" [EH14], which is a standard reference for the Rényi divergence, with a high number of citations. Moreover, there may be instances where the intricacy of the proof makes this process exceedingly challenging, or where the significance

[11] `https://leanprover-community.github.io/mathlib4_docs/`.

[12] `https://github.com/leanprover-community/mathlib4/tree/master`

[13] The authors have since published an errata: `https://www.timvanerven.nl/assets/publications/2014/Renyi-errata-2024-4-2.pdf`.

of the result requires a higher level of certainty[14]. The use of a proof assistant can assist in this regard, as the software automatically checks the proof for logical correctness. In the case of Lean, the piece of code responsible for verification is referred to as the kernel. The kernel is a relatively concise component that has been subjected to extensive testing and is designed to be highly reliable. This makes it extremely unlikely for a proof accepted by Lean to be incorrect. Moreover, it is possible to use other independently developed kernels, further reducing the risk of bugs. In addition to its greater reliability, this process is also considerably faster than manual checking[15].

While ensuring the logical correctness of the proofs is a significant concern for logicians and computer scientists, it can be argued that the majority of errors that pass through the peer review process are minor discrepancies that can be readily addressed and do not fundamentally impact the validity of the primary result. Therefore, from a mathematical standpoint, a more important reason to use a proof assistant is that it makes it easier to generalize mathematical results. Let us illustrate this with an example.

This is our current formulation[16] of the data processing inequality for the KL divergence in Lean:

```
lemma kl_comp_right_le [StandardBorelSpace α]
    [CountableOrCountablyGenerated α β]
    [IsFiniteMeasure μ] [IsFiniteMeasure ν] [IsMarkovKernel κ] :
    kl (μ ∘ₘ κ) (ν ∘ₘ κ) ≤ kl μ ν
```

The first three lines contain the hypotheses, and the last one is the thesis. We can notice that this version of the lemma assumes μ to be a finite

[14]These were among the motivations cited by Peter Sholze while proposing the challenge that inspired the Liquid Tensor Experiment. See `https://www.ma.imperial.ac.uk/~buzzard/xena/pdfs/liquid_tensor_experiment.pdf`.

[15]Checking a single new lemma typically takes less than a second, and the entire Mathlib library can be checked in less than an hour.

[16]This statement can be generalized, in particular relaxing the hypotheses of standard Borel space and countability. However, as will be discussed in Chapter 5, there are some technical difficulties that have prevented us to do it so far.

measure. What if we suspect that the result holds for general measures as well? On paper, we would have to carefully examine the proof, to ascertain whether the hypothesis is indeed superfluous. In Lean, we can simply erase [`IsFiniteMeasure` μ] and see if the proof still compiles without throwing any errors: if it does then we are done, otherwise the theorem prover will indicate precisely where the proof fails, enabling us to either attempt to resolve the issue or gain insight into why the proof does not hold in that generality. After that, we could easily look for the lemmas that use this inequality and determine whether it is possible to propagate the generalization. Furthermore, in the event that any given lemma contains a hypothesis that is entirely unused, Lean will automatically throw a warning, suggesting the user to remove it.

Another advantage of using a proof assistant is that it allows multiple researchers to work together on the same project, coordinating their efforts on shared repositories (for example on GitHub), without the need to manually check each other's proofs or to meet to discuss in person. In particular, a tedious and time-consuming aspect of working together on a mathematical project is the necessity to verify the proofs of others; this is especially true when the mathematicians are working together for the first time, and have yet to establish mutual trust, or when the level of expertise of the researchers is different, such as in the case of a student working with a professor. The use of a proof assistant can remove this burden, as the software will automatically check the proofs, allowing for a more focused interaction on the high-level mathematical content and the general direction of the project. This is true both in the case of a small project, such as the one presented in this thesis, and in the case of larger projects, like the ones mentioned above, where the use of a proof assistant can also help in the coordination of the work. In the context of formalization efforts involving numerous people, the leading researchers can define the goals, outline the structure of the project, and provide the informal proofs. Meanwhile, other contributors can focus on the formalization of the results, without the necessity of understanding the

entire project in detail, or even just give small contributions in their spare time, which can collectively yield a significant outcome. One tool that can be particularly useful in this context is leanblueprint[17], a piece of software that enables the creation of an informal blueprint for the project, wherein informal statements can be linked to their formal counterparts. This tool automatically produces a PDF document, as well as a website and a dependency graph. It can be used to track progress, facilitate coordination, and document informal proofs before they are formalized. Additionally, it can be employed to present results after formalization is complete.

Lastly, it is worth mentioning that this is not an exhaustive list of the advantages of using a proof assistant. For instance, these tools can be used to support the teaching of mathematics, and they offer a way of enabling computers to interact with rigorous mathematics. This can be useful in the development of AI, where the formal proofs can be used to enhance the capabilities of autonomous systems, with the potential to revolutionize the way we do mathematics and to allow AI to help in mathematical research.

[17] `https://github.com/PatrickMassot/leanblueprint`.

Chapter 2

Preliminary Notions

In this chapter we will introduce a number of definitions and results that will prove useful in the sequel. In particular, we will define the concept of kernel, which generalizes the notion of measure and will be central throughout all the thesis. Then we will examine the Lebesgue decomposition for measures and kernels, which enables us to define the Radon-Nikodym derivative and singular parts; these are going to be at the heart of the definition of f-divergence in the following chapter and will frequently appear in proofs. Finally, we will present a version of the integration by parts formula that will be used to prove the central result of this thesis, the data processing inequality.

2.1 Transition kernels

A concept that appears frequently in probability is that of conditioning. For example, given two random variables X and Y, we can ask what is the probability distribution of Y given the additional information that X has assumed a specific value x. This is usually denoted by $P_{Y|X=x}$, and it is called the conditional probability of Y given X.

Transition kernels (henceforth referred to as "kernels") are a way of representing the idea of a transformation that involves some degree of randomness, i.e. a function that has stochastic outputs, and they offer a versatile approach

to handling conditional probabilities. In the context of information theory, they are used to represent channels through which information flows. Moreover, kernels can be regarded as both a generalization of measures and a generalization of measurable functions. Consequently, many results mentioning these objects can be generalized to kernels. Throughout this thesis we will use the language of kernels extensively. This section introduces the concept of kernels as well as some operations that can be performed with them, and some of their properties.

Definition 2.1.1 (Kernel)**.** Let $(\mathcal{X}, \mathscr{F}_\mathcal{X}), (\mathcal{Y}, \mathscr{F}_\mathcal{Y})$ be measurable spaces. A *kernel* from $\mathcal{X}$ to $\mathcal{Y}$ is a function $\kappa\colon \mathcal{X} \times \mathscr{F}_\mathcal{Y} \to \overline{\mathbb{R}}_+$ such that for every $x \in \mathcal{X}$ the function $\kappa(x, \cdot)$ is a measure on $\mathcal{Y}$, and for every $B \in \mathscr{F}_\mathcal{Y}$ the function $x \mapsto \kappa(x, B)$ is measurable. We write $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$.

Remark 2.1.2. Currently, in Mathlib a kernel $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ is defined as a measurable function $\kappa\colon \mathcal{X} \to \mathcal{M}(\mathcal{Y})$:

```
structure Kernel (α β : Type*)
    [MeasurableSpace α] [MeasurableSpace β] where
  toFun : α → Measure β
  measurable' : Measurable toFun
```

This definition is equivalent to the aforementioned one, and relies on the natural measurable structure that can be defined on the space of measures, which Lean finds automatically using the typeclass instance system, thanks to the following instance:

```
instance instMeasurableSpace : MeasurableSpace (Measure α) :=
  ⨆ (s : Set α) (_ : MeasurableSet s), (borel ℝ≥0∞).comap fun μ => μ s
```

That is, given a measurable structure over $\mathcal{X}$, we consider the smallest sigma algebra on $\mathcal{M}(\mathcal{X})$ such that for all measurable sets $A \subseteq \mathcal{X}$ the projection $\mu \mapsto \mu(A)$ is measurable with respect to the Borel σ-algebra on $\overline{\mathbb{R}}_+$.

Two special kinds of kernels are worth mentioning: the constant kernels and the deterministic kernels. They occur when the function is degenerate,

respectively in the first and second argument, and they are the reason why we can say that kernels are a generalization of measures and measurable functions.

Definition 2.1.3 (Constant kernel)**.** Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, $\mu \in \mathcal{M}(\mathcal{Y})$. The *constant kernel* with value μ is the kernel defined by $\kappa(x, \cdot) = \mu$ for every $x \in \mathcal{X}$.

Definition 2.1.4 (Deterministic kernel)**.** Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, $f \colon \mathcal{X} \to \mathcal{Y}$ a measurable function. The *deterministic kernel* associated with f is the kernel defined by $\kappa(x, \cdot) = \delta_{f(x)}$ for every $x \in \mathcal{X}$, where δ_y is the Dirac measure at y.

We will sometimes refer to the deterministic kernel associated with f using the same notation f.

The notation we use for kernels highlights the fact that we can think of them as stochastic functions from a space to another, with the output being not a single fixed value but rather a measure on the second space. This analogy is more evident when the output measures are probability measures, as we can interpret the output of the kernel as taking different values with certain probabilities. Kernels that only output probability measures are referred to as Markov kernels and they are the most commonly used. However, we can also consider classes of kernels that satisfy weaker properties, and many results that hold for Markov kernels can be extended to those classes.

Definition 2.1.5. Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ a kernel. We say that κ is:

- a *Markov* kernel if for every $x \in \mathcal{X}$ the measure $\kappa(x, \cdot)$ is a probability measure.
- a *finite* kernel if $\kappa(x, \cdot)$ is a finite measure uniformly in x, i.e. there exists some $M \in \mathbb{R}$ such that for every $x \in \mathcal{X}$ we have $\kappa(x, \mathcal{Y}) \leq M$.
- an *s-finite* kernel if it is a countable sum of finite kernels.

Remark 2.1.6. It is evident that every Markov kernel is a finite kernel, and that every finite kernel is s-finite. These classes of kernels are the analogues of probability measures, finite measures and s-finite measures. Indeed, in the case of a constant kernel, they are identical. One could notice, however, that we are not defining an analogue of the σ-finiteness for measures, which is widely used in measure theory. It is indeed possible to define the notion of σ-finite kernel, as a kernel that is a countable sum of finite kernels that are pairwise mutually singular[1]. This way the class of σ-finite kernels would find itself between that of finite and s-finite kernels. However, for our purposes we will only need finite and s-finite kernels and the definition of σ-finiteness is not present in Mathlib at the moment. It is conceivable that with additional effort, some of the results we have used and formalized could be extended from finite to σ-finite kernels.

Remark 2.1.7. Recall that a measure μ is said to be s-finite if it is a countable sum of finite measures. This notion is not as widespread as that of finite or σ-finite measure, but it is nevertheless sufficient for many purposes. Moreover, this class encompasses many interesting measures that are not σ-finite. For example, a measure that assigns infinite mass to a single point can still be s-finite, as we can use a countable sum of Dirac measures to represent it at that point. For a more detailed discussion on s-finiteness see [VO18].

Remark 2.1.8. Since the Dirac measure is a probability measure, all deterministic kernels are Markov kernels.

Kernels can also be combined with measures or other kernels in various ways, to obtain new measures or kernels.

Definition 2.1.9 (Composition)**.** Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be measurable spaces, $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$, $\eta\colon \mathcal{Y} \rightsquigarrow \mathcal{Z}$ and $\mu \in \mathcal{M}(\mathcal{Y})$. We define the *composition* of κ and η and

[1]Here the notion of mutual singularity of kernels is intentionally left vague, as it can be interpreted in different ways, giving rise to different definitions of σ-finiteness, all of which collapse to the classic definition in the case of measures. See [VO18, Definition 1] for additional details.

the composition of μ and η respectively as the kernel $\eta \circ \kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Z}$ and the measure $\eta \circ \mu \in \mathcal{M}(\mathcal{Z})$ such that for every $x \in \mathcal{X}$ and $B \in \mathscr{F}_{\mathcal{Z}}$ we have

$$(\eta \circ \kappa)(x, B) = \int_{\mathcal{Y}} \eta(y, B)\kappa(x, dy),$$

and

$$(\eta \circ \mu)(B) = \int_{\mathcal{Y}} \eta(y, B)\mu(dy).$$

Definition 2.1.10 (Composition product)**.** Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be measurable spaces, $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$, $\eta \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$, $\mu \in \mathcal{M}(\mathcal{X})$. We define the *composition product* of κ and η and the composition product of μ and κ respectively as the kernel $\kappa \otimes \eta \colon \mathcal{X} \rightsquigarrow \mathcal{Y} \times \mathcal{Z}$ and the measure $\mu \otimes \kappa \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ such that for every $x \in \mathcal{X}$, $A \in \mathscr{F}_{\mathcal{Y}}$ and $B \in \mathscr{F}_{\mathcal{Z}}$ we have

$$(\kappa \otimes \eta)(x, A \times B) = \int_A \eta((x, y), B)\kappa(x, dy),$$

and

$$(\mu \otimes \kappa)(A \times B) = \int_A \kappa(x, B)\mu(dx).$$

Remark 2.1.11. The composition and composition product of a measure and a kernel are just the special cases of the same operations between kernels when one of the kernels is constant.

Moreover, we defined these operations based on their behavior when applied to measurable sets or rectangles. However, these properties can be easily extended to the case of integrals, in particular when we have a kernel and a measure the following hold:

i) $\int_B f(y)(\kappa \circ \mu)(dy) = \int_{\mathcal{X}} \int_B f(y)\kappa(x, dy)\mu(dx)$,

ii) $\int_{A \times B} g(x, y)(\mu \otimes \kappa)(dx, dy) = \int_A \int_B g(x, y)\kappa(x, dy)\mu(dx)$,

where $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$, $\mu \in \mathcal{M}(\mathcal{X})$, $f \colon \mathcal{Y} \to \overline{\mathbb{R}}$ and $g \colon \mathcal{X} \times \mathcal{Y} \to \overline{\mathbb{R}}$ are integrable functions, $A \in \mathscr{F}_{\mathcal{X}}$ and $B \in \mathscr{F}_{\mathcal{Y}}$.

Definition 2.1.12 (Product)**.** Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be measurable spaces, $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$, $\eta\colon \mathcal{X} \rightsquigarrow \mathcal{Z}$. We define the *product* of κ and η as the kernel $\kappa \times \eta\colon \mathcal{X} \rightsquigarrow \mathcal{Y} \times \mathcal{Z}$ such that for every $x \in \mathcal{X}$, $A \in \mathscr{F}_{\mathcal{Y}}$ and $B \in \mathscr{F}_{\mathcal{Z}}$ we have

$$(\kappa \times \eta)(x, A \times B) = \kappa(x, A)\eta(x, B).$$

Definition 2.1.13 (Parallel product)**.** Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{W}$ be measurable spaces, $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Z}$, $\eta\colon \mathcal{Y} \rightsquigarrow \mathcal{W}$. We define the *parallel product* of κ and η as the kernel $\kappa \parallel \eta\colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z} \times \mathcal{W}$ such that for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $A \in \mathscr{F}_{\mathcal{Z}}$ and $B \in \mathscr{F}_{\mathcal{W}}$ we have

$$(\kappa \parallel \eta)((x, y), A \times B) = \kappa(x, A)\eta(y, B).$$

Remark 2.1.14. The product and parallel product are the pointwise products of the measures:

$$(\kappa \times \eta)(x, \cdot) = \kappa(x, \cdot) \otimes \eta(x, \cdot),$$

$$(\kappa \parallel \eta)((x, y), \cdot) = \kappa(x, \cdot) \otimes \eta(y, \cdot),$$

where $\otimes$ is the usual product of measures.

Remark 2.1.15. The operations that we just described actually give rise to kernels, in particular the measurability condition is satisfied. For a proof of this, see [Kal21] and the Lean formalization.

Remark 2.1.16. In Lean, in order to distinguish the symbols used for these operations from existing notation within the library, we employ the notations $\otimes_k$ and $\otimes_m$, respectively for the composition product between two kernels and the composition product between a measure and a kernel. Similarly, we use $\circ_k$ and $\circ_m$ for the composition, $\times_k$ for the product (the product between measures is written $\times_m$) and $\parallel_k$ for the parallel product.

We also define some special kernels that are useful for combining other kernels in various ways.

Definition 2.1.17 (Identity, copy, discard and swap kernels)**.** Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces and $\{*\}$ a measurable space with one element. Then we

define the *identity*, *copy*, *discard* and *swap* kernels as the deterministic kernels $id_{\mathcal{X}} \colon \mathcal{X} \rightsquigarrow \mathcal{X}$, $copy_{\mathcal{X}} \colon \mathcal{X} \rightsquigarrow \mathcal{X} \times \mathcal{X}$, $discard_{\mathcal{X}} \colon \mathcal{X} \rightsquigarrow \{*\}$ and $swap_{\mathcal{X},\mathcal{Y}} \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y} \times \mathcal{X}$ associated respectively with the functions $x \mapsto x$, $x \mapsto (x, x)$, $x \mapsto *$ and $(x, y) \mapsto (y, x)$.

In some cases, when it is evident from the context, we may omit the mention of the dependency on the spaces. Moreover, for the composition with the swap kernel we use the notation $(\cdot)_{\leftrightarrow}$, i.e. $swap_{\mathcal{X},\mathcal{Y}} \circ \kappa = (\kappa)_{\leftrightarrow}$.

Remark 2.1.18. It is interesting to note that the operations that we have defined above are not independent of one another; rather, some can be composed to yield others. In particular, it is sufficient to define the composition and parallel composition, along with the kernels in Definition 2.1.17, in order to obtain the product and composition product as follows:

$$\kappa \times \eta = (\kappa \parallel \eta) \circ copy_{\mathcal{X}},$$

$$\mu \otimes \kappa = (id_{\mathcal{X}} \parallel \kappa) \circ copy_{\mathcal{X}} \circ \mu = (id_{\mathcal{X}} \times \kappa) \circ \mu,$$

$$\kappa \otimes \xi = (id_{\mathcal{Y}} \parallel \xi) \circ (id_{\mathcal{Y}} \parallel swap_{\mathcal{Y},\mathcal{X}}) \circ (copy_{\mathcal{Y}} \parallel id_{\mathcal{X}}) \circ (\kappa \parallel id_{\mathcal{X}}) \circ copy_{\mathcal{X}},$$

where $\mu \in \mathcal{M}(\mathcal{X})$, $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$, $\eta \colon \mathcal{X} \rightsquigarrow \mathcal{Z}$ and $\xi \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$.

Using these basic operations it is possible to interpret kernels through the lens of category theory. In particular, it is possible to define a category where the objects are the measurable spaces and the morphisms are certain classes of kernels, like Markov kernels. Categories constructed in this manner are referred to as Markov categories or copy-discard categories. This approach allows us to view probability theory from a very general point of view, and can also be used to make kernel computations easier, by representing them with string diagrams. For a more detailed discussion on this topic see [Per24; Fri+23].

An important property of Markov kernels, that we will use frequently in the following chapters, is that they preserve the total mass of a measure when combined with it through composition or composition product. For additional properties of these operations, see Appendix A and [Kal21].

Proposition 2.1.19. *Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ a Markov kernel, $\mu \in \mathcal{M}(\mathcal{X})$. Then:*

i) $(\kappa \circ \mu)(\mathcal{Y}) = \mu(\mathcal{X})$,

ii) $(\mu \otimes \kappa)(\mathcal{X} \times \mathcal{Y}) = \mu(\mathcal{X})$.

In particular, if μ is a probability measure, then $\kappa \circ \mu$ and $\mu \otimes \kappa$ are probability measures.

Proof. **(i)** By definition of composition we have

$$(\kappa \circ \mu)(\mathcal{Y}) = \int_{\mathcal{X}} \kappa(x, \mathcal{Y})\mu(dx) = \int_{\mathcal{X}} 1\mu(dx) = \mu(\mathcal{X}).$$

(ii) By definition of composition product we have

$$(\mu \otimes \kappa)(\mathcal{X} \times \mathcal{Y}) = \int_{\mathcal{X}} \kappa(x, \mathcal{Y})\mu(dx) = \int_{\mathcal{X}} 1\mu(dx) = \mu(\mathcal{X}).$$

⫽

Another important definition is that of Bayesian inverse, which is a way to invert a kernel. Just like the inverse of a function, the Bayesian inverse of a kernel is a new kernel that goes in the opposite direction. We can think of the kernel $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ as a source of observable data in $\mathcal{Y}$ given some hidden parameter in $\mathcal{X}$, then the Bayesian inverse takes the observed data and outputs a distribution representing how likely it is that the hidden parameter was a certain value, given a *prior* distribution $\mu \in \mathcal{M}(\mathcal{X})$ on the hidden parameter space. The distribution on the parameter space outputted by the Bayesian inverse is often called the *posterior*, so the Bayesian inverse can also be called the *posterior kernel*. For another point of view on the Bayesian inverse see [Cle+17; Dah+18].

In Chapter 4 we will see how the setting of hypothesis testing is similar to the one we just described, and how the Bayesian inverse can be used to solve some problems in this context. Moreover, we will use the Bayesian inverse in one of the proofs of the data processing inequality (Theorem 5.2.5).

Definition 2.1.20 (Bayesian inverse)**.** Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, $\mu \in \mathcal{M}(\mathcal{X})$, $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$. Then we say that a kernel $\eta\colon \mathcal{Y} \rightsquigarrow \mathcal{X}$ is a *Bayesian inverse* of κ with respect to μ if

$$(\eta \parallel id_{\mathcal{Y}}) \circ copy_{\mathcal{Y}} \circ \kappa \circ \mu = (id_{\mathcal{X}} \parallel \kappa) \circ copy_{\mathcal{X}} \circ \mu. \tag{2.1}$$

In this case we write $\eta = \kappa^{\dagger}_{\mu}$.

Remark 2.1.21. The condition in Equation (2.1) is equivalent to each of the following:

$$(\kappa^{\dagger}_{\mu} \times id_{\mathcal{Y}}) \circ \kappa \circ \mu = (id_{\mathcal{X}} \times \kappa) \circ \mu, \tag{2.2}$$

$$((\kappa \circ \mu) \otimes \kappa^{\dagger}_{\mu})_{\leftrightarrow} = \mu \otimes \kappa. \tag{2.3}$$

The existence of a Bayesian inverse of a general kernel $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ with respect to a general measure $\mu \in \mathcal{M}(\mathcal{X})$ is not always guaranteed. Nevertheless, if a Bayesian inverse exists, it is unique almost everywhere with respect to $\kappa \circ \mu$.

The following remark presents some sufficient conditions for the existence of the Bayesian inverse. See also Lemma 4.2.2 for an example where the Bayesian inverse exists and can even be computed explicitly.

Remark 2.1.22. It is possible to impose certain conditions on the spaces, kernel, and measure in order to guarantee the existence of the Bayesian inverse. In particular, let $\mathcal{X}$ be a standard Borel space, $\mathcal{Y}$ a measurable space, $\mu \in \mathcal{M}(\mathcal{X})$ a finite measure and $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ a finite kernel. Then the Bayesian inverse $\kappa^{\dagger}_{\mu}$ exists.

These hypotheses are those employed in the Lean definition of the Bayesian inverse:

```
def bayesInv [StandardBorelSpace α] [Nonempty α]
    (κ : Kernel α β) [IsFiniteKernel κ]
    (μ : Measure α) [IsFiniteMeasure μ] : Kernel β α :=
  ((μ ⊗ₘ κ).map Prod.swap).condKernel
```

We define the Bayesian inverse as the conditional kernel of the measure $(\mu \otimes \kappa)_{\leftrightarrow}$. The conditional kernel is one of two objects defined by the disintegration

of a measure. This process involves taking a measure $\nu \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ on a product space and splitting it as the composition product of a measure on the first space $\nu_{\mathcal{X}} \in \mathcal{M}(\mathcal{X})$ (the first marginal of ν) and a kernel $\nu_{\mathcal{Y}|\mathcal{X}} \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$, called the conditional kernel, such that $\nu = \nu_{\mathcal{X}} \otimes \nu_{\mathcal{Y}|\mathcal{X}}$. This operation is not always well-defined, but if $\mathcal{Y}$ is a standard Borel space and the measure is finite then it is. See [Kal21] for more details about the disintegration of measures.

Note also that the formalized definition requires the finiteness of the measure and kernel. However, it may be possible to prove this result for s-finite measures and kernels as well, with some additional effort.

2.2 Lebesgue Decomposition

In the context of measure theory, it is useful to consider the null sets, i.e. the sets to which the measure assigns zero mass. In particular, it is interesting, if we have two measures, to ascertain whether there are sets that are null for one measure but not for the other, or whether there is a set that is null for the first measure, while its complement is null for the second. This kind of observation gives rise to the notions of absolute continuity and singularity of measures, which are in some way analogous to the notions of parallel and orthogonal vectors in linear algebra.

Definition 2.2.1 (Absolutely continuous measure)**.** Let $\mathcal{X}$ be a measurable space and $\mu, \nu \in \mathcal{M}(\mathcal{X})$. We say that μ is *absolutely continuous* with respect to ν if for every $A \subseteq \mathcal{X}$ measurable such that $\nu(A) = 0$ we have $\mu(A) = 0$. In this case we write $\mu \ll \nu$.

Definition 2.2.2 (Mutually singular measures)**.** Let $\mathcal{X}$ be a measurable space and $\mu, \nu \in \mathcal{M}(\mathcal{X})$. We say that μ and ν are *mutually singular* if there exists $A \subseteq \mathcal{X}$ measurable such that $\mu(A) = 0$ and $\nu(A^c) = 0$. In this case we write $\mu \perp \nu$.

Furthering the linear algebra analogy, just like we can decompose a vector into the sum of its parallel and orthogonal parts with respect to another

vector, we can, under some hypotheses, decompose a measure into the sum of an absolutely continuous part and a singular part with respect to another measure. This is referred to as the Lebesgue decomposition and allows us to define the Radon-Nikodym derivative.

In the next chapter, we will see how this decomposition can be used to compare two measures, by confronting separately their absolutely continuous and singular parts. This will give rise to the notion of f-divergence.

In the sequel, given a measure $\mu \in \mathcal{M}(\mathcal{X})$ and a measurable function $f\colon \mathcal{X} \to \overline{\mathbb{R}}$, we will use the notation $f \cdot \mu$ to denote the measure defined by $(f \cdot \mu)(A) = \int_A f \,\mathrm{d}\mu$ for every measurable set A.

Theorem 2.2.3 (Lebesgue Decomposition)**.** *Let $\mathcal{X}$ be a measurable space, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ such that μ is s-finite and ν is σ-finite. Then there exists a measure $\xi \in \mathcal{M}(\mathcal{X})$ such that $\nu \perp \xi$ and a measurable function $f\colon \mathcal{X} \to \overline{\mathbb{R}}_+$ such that $\mu = f \cdot \nu + \xi$. Moreover, ξ is unique and f is ν-a.e. unique.*

Proof. See [Dud02, Theorem 5.5.3] for a proof of the case where both measures are σ-finite. For the general case see the proof in Mathlib, in particular the instance `MeasureTheory.Measure.haveLebesgueDecomposition_of_sigmaFinite` for the existence and the theorems `MeasureTheory.Measure.eq_singularPart` and `MeasureTheory.Measure.eq_rnDeriv` for the uniqueness. ⚔

Remark 2.2.4. The measure ξ is called the singular part of μ with respect to ν, and it is denoted by $\mu_{\perp\nu}$. The function f is called the Radon-Nikodym derivative of μ with respect to ν, and it is denoted by $\frac{d\mu}{d\nu}$. Moreover, $\frac{d\mu}{d\nu} \cdot \nu$ is called the absolutely continuous part of μ with respect to ν, since $\frac{d\mu}{d\nu} \cdot \nu \ll \nu$. With these notations we have that $\mu = \frac{d\mu}{d\nu} \cdot \nu + \mu_{\perp\nu}$.

Remark 2.2.5. In the case where $\mu \ll \nu$ we have that $\mu_{\perp\nu} = 0$ and $\mu = \frac{d\mu}{d\nu} \cdot \nu$. This is also called the Radon-Nikodym theorem.

Remark 2.2.6. The hypothesis of σ-finiteness for ν is necessary for the uniqueness of the Radon-Nikodym derivative. Consider the following counter example. Let $\mathcal{X} \coloneqq \{*\}$ be a measurable space with one element, $\mu \in \mathcal{M}(\mathcal{X})$ the

measure that assigns infinite mass to $*$. Observe that μ is s-finite, but not σ-finite. Let us consider the Lebesgue decomposition of μ with respect to itself. Since μ assigns positive mass to every point in $\mathcal{X}$, the only measure that is mutually singular with respect to μ is the zero measure. Therefore, the Lebesgue decomposition is of the form $\mu = f \cdot \mu$, for some function f. But it is evident that every strictly positive function will work, therefore the Radon-Nikodym derivative is not unique (not even almost everywhere).

Now that we have defined the Lebesgue decomposition for measures, a natural question is whether this concept can be extended to kernels. The naive way to do this is by considering the kernels $\kappa, \eta \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ as collections of measures indexed by the elements of the domain and then applying the Lebesgue decomposition to each of these measures:

$$\kappa(x) = \frac{d\kappa(x)}{d\eta(x)} \cdot \eta(x) + \kappa(x)_{\perp \eta(x)}.$$

The problem with this approach is that the resulting Radon-Nikodym derivative, which can be seen as a function $\mathcal{X} \times \mathcal{Y} \to \overline{\mathbb{R}}_+$, is measurable once we fix the first argument, but it is not necessarily jointly measurable. It thus becomes necessary to find an alternative means of defining the Lebesgue decomposition for kernels, one that allows for the joint measurability required for the calculation of integrals on the product space.

It turns out that it is possible to define the Radon-Nikodym derivative in a way that is jointly measurable. However, this requires certain assumptions to be made regarding the spaces, in particular we need the second space $\mathcal{Y}$ to be countably generated, i.e. there exists a countable collection of sets that generates the σ-algebra of $\mathcal{Y}$; alternatively we can require the first space $\mathcal{X}$ to be countable.

Notice that these hypotheses are satisfied if the second space is a standard Borel space, as this implies[2] that it is countably generated.

[2]Let $\mathcal{Y}$ be a standard Borel space, then there exists a countable dense set $D \subseteq \mathcal{Y}$ and the Borel σ-algebra is generated by the balls with rational radius with centers in D, which are countable.

Theorem 2.2.7 (Lebesgue decomposition theorem for kernels)**.** *Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces such that $\mathcal{X}$ is countable or $\mathcal{Y}$ is countably generated, and let $\kappa, \eta \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ be finite kernels. Then there exists a kernel $\xi \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ such that $\eta(x) \perp \xi(x)$ for every $x \in \mathcal{X}$ and a measurable function $f \colon \mathcal{X} \times \mathcal{Y} \to \overline{\mathbb{R}}_+$ such that $\kappa(x) = f(x, \cdot) \cdot \eta(x) + \xi(x)$ for every $x \in \mathcal{X}$. We denote $\kappa_{\perp\eta} = \xi$ and $\dfrac{d\kappa}{d\eta} = f$.*

Proof. In the case where $\mathcal{X}$ is countable, the naive approach works, since the joint measurability over $\mathcal{X} \times \mathcal{Y}$ is then implied by the measurability of $\dfrac{d\kappa(x)}{d\eta(x)}$ for every $x \in \mathcal{X}$. For the countably generated case, the proof involves a more complicated construction, for the details see the definition `ProbabilityTheory.Kernel.rnDeriv` in Mathlib and the relative file, in particular the lemma `ProbabilityTheory.Kernel.rnDeriv_add_singularPart`. ⁂

2.3 Generalized integration by parts

Integration by parts is one of the most powerful tools of real analysis, which, in its standard version, requires the two functions to be differentiable. However, for the purposes of our project, it is necessary to apply it in a situation where we have no differentiability guaranteed, in fact, one of the functions may even be discontinuous. In order to achieve this, we will present a version of the theorem for the Riemann-Stieltjes integral and adapt it to the Lebesgue integral.

For the definition and some properties of the Riemann-Stieltjes integral see Appendix C and [Apo74, Chapter 7].

Theorem 2.3.1 (Integration by parts)**.** *Let $a, b \in \mathbb{R}$ such that $a < b$, $f, g \colon [a, b] \to \mathbb{R}$ be bounded functions such that f is Riemann-Stieltjes integrable with respect to g. Then g is Riemann-Stieltjes integrable with respect to f, and we have*

$$\int_a^b f(x)\,\mathrm{d}g(x) + \int_a^b g(x)\,\mathrm{d}f(x) = f(b)g(b) - f(a)g(a).$$

Proof. See [Apo74, Theorem 7.6]. ⁂

The Riemann-Stieltjes integral is closely related to a particular type of Lebesgue integral, called the Lebesgue-Stieltjes integral. We will provide a brief introduction to this concept and adapt the integration by parts theorem to this setting, as this version of the theorem is the one that we have formalized in Lean.

Definition 2.3.2 (Stieltjes function)**.** Let $f\colon \mathbb{R} \to \mathbb{R}$. We say that f is a *Stieltjes function* if it is nondecreasing and right continuous.

It can be proven (see [Whe15, Theorems 11.8 and 11.10]) that a Stieltjes function f uniquely defines a measure on the Borel σ-algebra of $\mathbb{R}$ in the following way.

Definition 2.3.3 (Lebesgue-Stieltjes measure and integral)**.** Let f be a Stieltjes function. We define the *Lebesgue-Stieltjes measure* associated with f, denoted Λ_f, as the only measure on the Borel σ-algebra of $\mathbb{R}$ such that for every $a, b \in \mathbb{R}$ with $a < b$ we have

$$\Lambda_f((a, b]) = f(b) - f(a).$$

Moreover, if $g\colon \mathbb{R} \to \mathbb{R}$, we call the integral $\int g \, \mathrm{d}\Lambda_f$ the *Lebesgue-Stieltjes integral* of g with respect to f.

Remark 2.3.4. f is the CDF (cumulative distribution function) of Λ_f.

Remark 2.3.5. Definition 2.3.3 also works for nondecreasing right continuous functions defined only on intervals, in that case the Lebesgue-Stieltjes measure is defined only on the Borel σ-algebra of the interval.

The Stieltjes functions and the Lebesgue-Stieltjes measure are already defined in Mathlib, but only for functions on the whole real line, not on intervals.

Corollary 2.3.6. *Let $a, b \in \mathbb{R}$ such that $a < b$, f, g be Stieltjes functions such that f is continuous on $[a, b]$. Then we have*

$$\int_{(a,b]} f(x) \, \mathrm{d}\Lambda_g(x) = f(b)g(b) - f(a)g(a) - \int_{(a,b]} g(x) \, \mathrm{d}\Lambda_f(x).$$

Proof. See Remark C.0.5. ⚔

Remark 2.3.7. The integration by parts theorem is available in Mathlib, under the name `intervalIntegral.integral_deriv_mul_eq_sub_of_hasDeriv_right` (there are actually multiple lemmas nearby with slightly different hypotheses). However, this formulation of the theorem requires both functions to be at least continuous, while for our application one of the functions is not necessarily continuous. The following is the statement of the integration by parts theorem that is currently in the project:

```
lemma integral_stieltjes_meas_by_parts (f g : StieltjesFunction)
    (a b : ℝ) (hf : ContinuousOn f (Set.Icc a b)) :
    ∫ x in a..b, f x ∂g.measure
      = (f b) * (g b) − (f a) * (g a) − ∫ x in a..b, g x ∂f.measure := by
  sorry
```

The proof of the theorem has yet to be formalized, and it is currently left as a task for ourselves or other members of the Lean community to undertake in the future. This decision of leaving an unproven result was dictated by the will to focus our formalization efforts on the core part of the project, namely the proof of the data processing inequality and the development of information theory related objects. However, the inclusion of a non-formalized element in a formalization project leaves some room for uncertainty about the absolute solidity of the final results. To address this issue, in Remark C.0.5 we provide a comprehensive justification for the mathematical assumption that we are operating under, based on reliable mathematical sources.

Note also that this situation gives us the opportunity to illustrate the flexibility of Lean as a formalization tool. In fact, it allows us to assume a result that we trust to be true, postponing the proof of that result and letting us focus on building the piece of theory that we are currently most interested in. Be warned, however, not to abuse this possibility, as it can lead to some bugs, such as forgotten hypotheses in the result we are assuming to be true, which can potentially undermine the soundness of the whole project and turn

out to be very time-consuming to fix, as the number of results depending on the assumption becomes too large.

Chapter 3

Information divergences

Information divergences serve as a means of quantifying the dissimilarity between two probability measures. They are employed extensively in the fields of information theory, statistics, and machine learning. The majority of the applications of divergences focus on probability measures. However, divergences can also be defined for general measures, and in this work we will consider them in this more general setting, restricting to smaller classes of measures only when necessary.

This more general approach allows the application of techniques from convex optimization to the study of divergences. For example, one may consider a fixed measure and seek to identify the measure that minimizes a divergence with respect to it. In fact, convex optimization often requires the set of feasible solutions to be a convex cone. However, the set of probability measures does not fall into this category, whereas the set of finite measures does.

A potential definition of divergence (between probability measures) is given in a categorical setting in [Per24, Chapter 2]. However, due to certain technical complexities in its implementation[1], this definition will not work

[1]The definition of divergence in [Per24] requires the divergence between a measure and itself to be zero. This is not true for the current formalization of f-divergences, unless the measure is σ-finite. See also Remark 3.1.4.

for some of the divergences that we have implemented in the case of general measures. For this reason, a general definition of divergence will not be introduced in this chapter. However, a general principle in information theory is information cannot be generated merely by manipulating existing data. As divergences represent a means of quantifying information, it is reasonable to posit that they should satisfy a property that reflects this principle. This property is known as the data processing inequality, and it states that the divergence between two measures cannot increase when both measures are composed with the same Markov kernel. In this chapter we will describe a large class of divergences, known as the f-divergences, along with some specific examples of f-divergences and also some other divergences that do not belong to this class. Chapter 5 we will provide further details about the data processing inequality, and will present a proof demonstrating that this inequality is satisfied by the aforementioned divergences.

Note that the notion of divergence does not require symmetry (in particular a divergence does not need to be a distance). Indeed, many widely used divergences like the Kullback-Leibler divergence (see Definition 3.2.1) are not symmetric. A possible way to interpret this asymmetry is that divergences can be understood as a way to quantify, in a binary testing setting, how easy it is to exclude that the samples are taken from a probability distribution μ if they are actually sampled from another distribution ν. The following example is provided to illustrate this concept.

Example 3.0.1. Let $\mu \coloneqq \mathrm{Ber}(0.5)$ be the Bernoulli distribution with mean 0.5, and ν the Dirac measure at 0, that is $\nu \coloneqq \mathrm{Ber}(0)$. If the true measure is μ then we have a chance to sample 1, in this case we can directly exclude that the samples are taken from ν with complete confidence. On the other hand, if the true measure is ν then we will never be able to exclude with certainty that the samples are taken from μ. This is reflected in the Kullback-Leibler divergence between μ and ν being infinite, while the divergence between ν and μ is finite. In particular, as we will see in Example 3.2.4, $\mathrm{KL}(\mu, \nu) = +\infty$ and $\mathrm{KL}(\nu, \mu) = \log(2)$.

3.1 f-divergences

Definition 3.1.1 (f-divergence)**.** Let $\mathcal{X}$ be a measurable space, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ and $f\colon \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}$ a convex function such that $f(1) = 0$. The *f-divergence* between μ and ν is defined as

$$D_f(\mu, \nu) \coloneqq \int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\nu + f'(\infty)\mu_{\perp\nu}(\mathcal{X}).$$

To understand the intuition behind f-divergences, it is instructive to consider the Lebesgue decomposition $\mu = \dfrac{d\mu}{d\nu} \cdot \nu + \mu_{\perp\nu}$. In the case where $\mu = \nu$, this decomposition becomes $\mu = 1 \cdot \mu + 0$, therefore it is a good idea to study how much the Radon-Nikodym derivative deviates from 1 and how much the singular part differs from zero. The integral part of the f-divergence is a way of quantifying the first part, i.e. how much μ and ν differ in the regions where both have positive mass. The other part quantifies the degree to which μ assigns mass to the regions where ν has zero mass. Different choices of f correspond to different ways of weighing those elements. The reasons for requiring $f(1) = 0$ should now be more apparent. Moreover, as will be highlighted in Remark 3.1.6, this hypothesis is not overly restrictive. The convexity of f, on the other hand, is a natural requirement, both because it guarantees the existence of $f'(\infty)$ (see Remark D.0.2), it ensures some desirable properties like the one in the following remark, and is crucial in all the proofs of the DPI that we present in Chapter 5.

Remark 3.1.2. Let $f\colon \mathbb{R} \to \mathbb{R}$ be convex, and μ, ν be finite measures. Then for every line of equation $y = ax + b$ that is tangent to the graph of f, we have that the graph of f is above the line, that is $f(x) \geq ax + b$ for every $x \in \mathbb{R}$. Therefore

$$\int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\nu \geq \int_{\mathcal{X}} \left(a\frac{d\mu}{d\nu} + b\right) \mathrm{d}\nu = a\int_{\mathcal{X}} \frac{d\mu}{d\nu}\,\mathrm{d}\nu + b\nu(\mathcal{X}).$$

The left-hand side of this equation is finite, since μ and ν are finite measures, and $0 \leq \int_{\mathcal{X}} \dfrac{d\mu}{d\nu}\,\mathrm{d}\nu \leq \mu(\mathcal{X}) < +\infty$. It follows that the integral cannot be $-\infty$. Moreover, for a similar reason, the integral of the negative part of f must be

finite, hence the integral cannot be undefined. We conclude that the only case where $f\left(\frac{d\mu}{d\nu}\right)$ is not integrable is if the integral is $+\infty$.

Remark 3.1.3. The following is the definition of f-divergence formalized in Lean:

```
def fDiv (f : ℝ → ℝ) (μ ν : Measure α) : EReal :=
  if ¬ Integrable (fun x ↦ f ((∂μ/∂ν) x).toReal) ν then ⊤
  else ∫ x, f ((∂μ/∂ν) x).toReal ∂ν
    + derivAtTop f * μ.singularPart ν Set.univ
```

This definition is slightly different from the aforementioned one. First, we defined the f-divergence without any hypothesis on f. The convexity of f and the fact that $f(1) = 0$ are still needed for many properties to hold; however, we require these conditions as hypotheses of the specific lemmas, not of the definition itself. This is a common practice in formal theorem proving, as it allows us to talk about some mathematical objects without having to explicitly state many hypotheses each time we mention them.

Another difference is in the way that this definition handles the case where $f\left(\frac{d\mu}{d\nu}\right)$ is not integrable: in this case Definition 3.1.1 can lead to many different situations, since the integral could either be $+\infty$, $-\infty$ or undefined and $f'(\infty)\mu_{\perp\nu}(\mathcal{X})$ could be finite, $+\infty$ or $-\infty$; consequently, the f-divergence could be $+\infty$, $-\infty$ or undefined. In the Lean definition we chose to simplify this situation by setting the f-divergence to $+\infty$ in all these cases. This decision makes the implementation easier, since the integral part is now always finite, eliminating the need to address undefined forms such as $\infty - \infty$. Furthermore, at present, there is no definition of integral in Mathlib that allows us to talk about general integrals that can also have

$\pm\infty$ as values[2], so handling the cases where the integral is infinite would have required some tricks to circumvent the problem, potentially making the formalization much harder. Another justification for this choice is that in the case where f is convex and the measures are finite, hypotheses that are requires for many results about f-divergences, $f\left(\frac{d\mu}{d\nu}\right)$ is not integrable if and only if the f-divergence is $+\infty$ (see Remark 3.1.2).

The last difference is in the domain and codomain of the function f. In Definition 3.1.1 we defined f as a function from $\overline{\mathbb{R}}_+$ to $\overline{\mathbb{R}}$, this is the most natural choice, as we apply f to the Radon-Nikodym derivative, which can take values in $\overline{\mathbb{R}}_+$, and there is no reason to exclude the possibility of f taking infinite values, since we can still integrate it meaningfully. In the Lean definition, however, we require f to be a function from $\mathbb{R}$ to $\mathbb{R}$; this choice is mainly dictated by the limits of the tools that we have at our disposal in Mathlib.

First, the Bochner integral, which we use in the Lean definition, cannot take as input a function with infinite values. Furthermore, working with the type `EReal` in Lean is more cumbersome than working with the real numbers, since its implementation in Mathlib is still lacking in some aspects[3]. Note also that if we want f to be convex, then it can never take the value $-\infty$ unless it is constant. Furthermore, none of the functions that we will consider

[2]The two types of integrals currently implemented in Mathlib are the Bochner integral, called `MeasureTheory.integral`, and the lower Lebesgue integral, called `MeasureTheory.lintegral`. The Bochner integral takes a function with values in a normed additive commutative group and gives back values in that same group, in our case $\overline{\mathbb{R}}$ is not a group, therefore the function must have values in $\mathbb{R}$ and the integral cannot be infinite; in the case where the function is not integrable the integral takes the junk value 0. The lower Lebesgue integral, on the other hand, takes a function with values in $\overline{\mathbb{R}}_+$; it can give back $+\infty$ as a value, but it can only handle nonnegative functions. In our case, having singled out the case when the function is not integrable, we can safely use the Bochner integral.

[3]Actually, numerous results formalized during this project about `EReal` have been added to Mathlib, and we have many other results that we use locally and we are planning to port to Mathlib in the near future.

in this work will take infinite values, so this restriction is not a significant limitation.

Regarding the domain, using $\overline{\mathbb{R}}_+$ would have prevented us from leveraging most results from the convex analysis part of Mathlib like `ConvexOn.map_average_le` (Jensen's inequality), since they are stated for a real normed space, a condition that $\overline{\mathbb{R}}_+$ does not satisfy. Furthermore, in the case where μ is σ-finite, the Radon-Nikodym derivative is finite almost everywhere with respect to ν[4], so it is not really restrictive to take $\mathbb{R}$ as the domain of f.

Remark 3.1.4. Notice how both Definition 3.1.1 and the Lean definition of f-divergence rely on the Radon-Nikodym derivative in a crucial way. This can result in some issues, namely the f-divergence not being well-defined, if the Radon-Nikodym derivative is not unique. This can occur if the measures lack the σ-finiteness hypothesis, see Remark 2.2.6. For this reason we are going to assume that all the measures are σ-finite in the rest of this chapter. Nevertheless, we encourage the reader to check the Lean code for more fine-grained details about the hypotheses of the theorems and lemmas that we have formalized. Indeed, some results might hold true even without the uniqueness of the Radon-Nikodym derivative. For instance, see `ProbabilityTheory.fDiv_zero`, which states that $D_0(\mu, \nu) = 0$ for every $\mu, \nu \in \mathcal{M}(\mathcal{X})$.

Keeping the required hypotheses to a minimum during the formalization is a generally good practice, not only because it increases the generality of the results, but also because it improves the usability of the code, as fewer hypotheses need to be stated each time a result is used, and better performance, as it puts less of a burden on the Lean typeclass inference system, which is often a bottleneck in the compilation process.

We now turn to a useful property of f-divergences: if we add a constant or a linear part to the function f, the f-divergence changes by a quantity that depends only on the total mass of the measures.

[4]See the Mathlib lemma `MeasureTheory.Measure.rnDeriv_lt_top` for a proof of this fact.

Proposition 3.1.5. *Let $\mathcal{X}$ be a measurable space, $\mu, \nu \in \mathcal{M}(\mathcal{X})$, $f\colon \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}$ a convex function, $a, b \in \mathbb{R}$, $\tilde{f} \coloneqq x \mapsto f(x) + ax + b$. Then*

$$D_{\tilde{f}}(\mu, \nu) = D_f(\mu, \nu) + a\mu(\mathcal{X}) + b\nu(\mathcal{X}).$$

Proof. The proof is just a simple computation:

$$\begin{aligned} D_{\tilde{f}}(\mu, \nu) &= \int_{\mathcal{X}} \tilde{f}\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\nu + \tilde{f}'(\infty)\mu_{\perp\nu}(\mathcal{X}) \\ &= \int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\nu + a\int_{\mathcal{X}} \frac{d\mu}{d\nu}\,\mathrm{d}\nu + b\nu(\mathcal{X}) + f'(\infty)\mu_{\perp\nu}(\mathcal{X}) + a\mu_{\perp\nu}(\mathcal{X}) \\ &= D_f(\mu, \nu) + a\mu(\mathcal{X}) + b\nu(\mathcal{X}). \end{aligned}$$

⫽

Remark 3.1.6. Using Proposition 3.1.5, we can see how the hypothesis $f(1) = 0$ is not really restrictive, since we can always add a constant part to f to make it satisfy this condition: $\tilde{f}(x) = f(x) - f(1)$.

Moreover, we can further modify f with a linear part to control its (right) derivative. For example, we can consider a centered version of f by setting $\tilde{f}(x) = f(x) - f(1) - f'_+(1)(x-1)$. This results in $\tilde{f}(1) = 0$, $\tilde{f}'(1) = 0$ and $D_{\tilde{f}}(\mu, \nu) = D_f(\mu, \nu) - f(1)\nu(\mathcal{X}) - f'_+(1)(\mu(\mathcal{X}) - \nu(\mathcal{X}))$.

3.2 Kullback-Leibler divergence

The Kullback-Leibler divergence, also referred to as relative entropy, represents a foundational concept in information theory, with applications spanning diverse fields such as statistics and machine learning. For further details about the KL divergence see [PW24, Chapter 2].

Definition 3.2.1 (Kullback-Leibler divergence)**.** Let $\mathcal{X}$ be a measurable space and $\mu, \nu \in \mathcal{M}(\mathcal{X})$. The *Kullback-Leibler divergence* between μ and ν is defined as

$$\mathrm{KL}(\mu, \nu) \coloneqq \begin{cases} \displaystyle\int_{\mathcal{X}} \log\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\mu & \text{if } \mu \ll \nu, \\ +\infty & \text{otherwise.} \end{cases}$$

The following result demonstrates that the KL divergence can be seen as an example of f-divergence, with $f(x) = x \log x$.

Proposition 3.2.2. *Let $\mathcal{X}$ be a measurable space, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ and $f(x) \coloneqq x \log x$. Then $D_f(\mu, \nu) = \mathrm{KL}(\mu, \nu)$.*

Proof. First suppose that $\mu \ll \nu$, then by Proposition B.0.1 we have that $\mu_{\perp\nu} = 0$. Therefore,

$$\begin{aligned} D_f(\mu,\nu) &= \int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\nu + f'(\infty)\mu_{\perp\nu}(\mathcal{X}) \\ &= \int_{\mathcal{X}} \frac{d\mu}{d\nu} \log\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\nu \\ \text{By Proposition B.0.1} \quad &= \int_{\mathcal{X}} \log\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\mu \\ &= \mathrm{KL}(\mu,\nu). \end{aligned}$$

Now suppose that $\mu \not\ll \nu$, then by Proposition B.0.1 we have that $\mu_{\perp\nu}(\mathcal{X}) > 0$. Notice also that $f'(x) = \log x + 1$, so $f'(\infty) = +\infty$, and by Remark 3.1.2 $\int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\nu > -\infty$. Hence,

$$D_f(\mu,\nu) = \int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\nu + f'(\infty)\mu_{\perp\nu}(\mathcal{X}) = +\infty = \mathrm{KL}(\mu,\nu).$$

⩚

Remark 3.2.3. In the case where $\mathcal{X}$ is discrete, the condition $\mu \ll \nu$ is equivalent to $\mathrm{supp}(\mu) \subseteq \mathrm{supp}(\nu)$, where $\mathrm{supp}(\mu) \coloneqq \{x \in \mathcal{X} \mid \mu(\{x\}) > 0\}$ is the support of μ. In this case the KL divergence can be written as

$$\mathrm{KL}(\mu,\nu) = \begin{cases} \displaystyle\sum_{x \in \mathrm{supp}(\mu)} \mu(\{x\}) \log\left(\frac{\mu(\{x\})}{\nu(\{x\})}\right) & \text{if } \mathrm{supp}(\mu) \subseteq \mathrm{supp}(\nu), \\ +\infty & \text{otherwise.} \end{cases}$$

Notice also that we can consider the sum to be over all the elements of $\mathcal{X}$, with the understanding that $0 \log(0) = 0 \log(\frac{0}{0}) = 0$.

An analogous result also holds in the case of discrete measures.

Example 3.2.4. Let $p, q \in [0,1]$, $\mu := \mathrm{Ber}(p)$ and $\nu := \mathrm{Ber}(q)$. Let us compute the KL divergence between μ and ν.

We begin by considering the case where $q \in \{0,1\}$. If $p = q$ then $\mu = \nu$ and by Proposition E.0.1 we have that $\mathrm{KL}(\mu,\nu) = 0$. If, instead, $p \neq q$, then $\mathrm{supp}(\mu) \not\subseteq \mathrm{supp}(\nu)$, and by Remark 3.2.3 we have that $\mathrm{KL}(\mu,\nu) = +\infty$.

Now suppose that $q \neq 0, 1$, therefore $\mathrm{supp}(\mu) \subseteq \{0,1\} = \mathrm{supp}(\nu)$. To compute the Radon-Nikodym derivative, observe that it has to satisfy $\mu(\{x\}) = \dfrac{d\mu}{d\nu}(x)\nu(\{x\})$ for $x \in \{0,1\}$, so we have that $\dfrac{d\mu}{d\nu}(0) = \dfrac{1-p}{1-q}$ and $\dfrac{d\mu}{d\nu}(1) = \dfrac{p}{q}$. Hence, by Remark 3.2.3 we have that

$$\begin{aligned}\mathrm{KL}(\mu,\nu) &= \sum_{x\in\mathcal{X}} \mu(\{x\}) \log\left(\frac{\mu(\{x\})}{\nu(\{x\})}\right) \\ &= \mu(\{0\}) \log\left(\frac{\mu(\{0\})}{\nu(\{0\})}\right) + \mu(\{1\}) \log\left(\frac{\mu(\{1\})}{\nu(\{1\})}\right) \\ &= p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right).\end{aligned}$$

We can summarize these computations with the following formula

$$\mathrm{KL}(\mathrm{Ber}(p), \mathrm{Ber}(q)) = p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right),$$

with the understanding that $0\log(0) = 0\log(\frac{0}{0}) = 0$ and $a \log\left(\dfrac{a}{0}\right) = +\infty$ for every $a > 0$.

As we mentioned at the beginning of this chapter, a property that we expect a good divergence to have is the data processing inequality. The Kullback-Leibler divergence actually satisfies a stronger property, known as the chain rule, which can be used to prove the DPI, but is not true for general f-divergences.

Theorem 3.2.5 (Chain rule)**.** *Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces such that $\mathcal{Y}$ is a standard Borel space, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ finite measures and $\kappa, \eta \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ Markov kernels. Then*

$$\mathrm{KL}(\mu \otimes \kappa, \nu \otimes \eta) = \mathrm{KL}(\mu,\nu) + \mathrm{KL}(\mu \otimes \kappa, \mu \otimes \eta).$$

Proof. See the Lean code, in particular `ProbabilityTheory.kl_compProd`. ⁂

An important consequence of the chain rule is the tensorization property of the KL divergence, which allows for the straightforward computation of the KL divergence between two product measures.

Corollary 3.2.6 (Tensorization)**.** *Let $n \in \mathbb{N}$, $(\mathcal{X}_i)_{i\in\{1,\dots,n\}}$ be a finite collection of countably generated measurable spaces and $\mu_i, \nu_i \in \mathcal{P}(\mathcal{X}_i)$ probability measures for $i \in \{1, \dots, n\}$. Then*

$$\mathrm{KL}\left(\prod_{i=1}^{n} \mu_i, \prod_{i=1}^{n} \nu_i\right) = \sum_{i=1}^{n} \mathrm{KL}(\mu_i, \nu_i).$$

Proof. See the Lean code, in particular `ProbabilityTheory.kl_prod_two`. ⁂

Remark 3.2.7. In particular, if we have multiple copies of the same measure, i.e. $\mu_i = \mu$ and $\nu_i = \nu$ for every i, then the tensorization property tells us that

$$\mathrm{KL}\left(\mu^{\otimes n}, \nu^{\otimes n}\right) = n\,\mathrm{KL}(\mu, \nu).$$

This is particularly useful when we have a number of independent samples from the same distribution, which can be thought of as a single sample from the product measure. This can happen in the context of hypothesis testing, as seen in Example 4.0.4, and can be used, for example, to compute lower bounds on the sample complexity of an estimation problem, i.e. the number of samples needed to achieve a certain level of accuracy.

3.3 Hellinger and Rényi divergences

The Hellinger α-divergences and the Rényi divergences are two important parametric families of divergences that are closely related to each other. Both families can be viewed as a generalization of the KL divergence, and some slightly different definitions for them can be found in the literature [Rén65; CA10; EH14]. The definitions presented in this section are those used in our formalization. In particular, we use the definition of Rényi divergence

from [EH14], while the Hellinger α-divergence is defined in such a way that the Rényi divergence can be expressed in terms of it, using a transformation similar to the one described in [CA10].

Many results about the Hellinger and Rényi divergences require the measures to be finite, at least with our proofs and current Mathlib tools. Consequently, we are going to assume this hypothesis in this section. However, we refer the reader to the Lean code for the versions of these results with more precise and weak hypotheses.

Definition 3.3.1 (Hellinger function)**.** Let $\alpha \in [0, +\infty)$. We call *Hellinger function* of order α the function $f_\alpha \colon \mathbb{R} \to \mathbb{R}$ defined as follows:

$$f_\alpha(x) \coloneqq \begin{cases} \mathbb{1}_{\{0\}}(x) & \text{if } \alpha = 0, \\ x \log(x) & \text{if } \alpha = 1, \\ \frac{x^\alpha - 1}{\alpha - 1} & \text{otherwise.} \end{cases}$$

Definition 3.3.2 (Hellinger α-divergence)**.** Let $\alpha \in [0, +\infty)$, $\mathcal{X}$ be a measurable space and $\mu, \nu \in \mathcal{M}(\mathcal{X})$. We define the *Hellinger α-divergence* between μ and ν as the f-divergence with the Hellinger function of order α:

$$\mathrm{H}_\alpha(\mu, \nu) \coloneqq D_{f_\alpha}(\mu, \nu).$$

Proposition 3.3.3. *Let $\alpha \in [0, +\infty)$, $\mathcal{X}$ be a measurable space and $\mu, \nu \in \mathcal{M}(\mathcal{X})$. Then the Hellinger divergence takes the following form:*

$$\mathrm{H}_\alpha(\mu, \nu) = \begin{cases} \nu\left(\left\{x \mid \frac{d\mu}{d\nu}(x) = 0\right\}\right) & \text{if } \alpha = 0, \\ \mathrm{KL}(\mu, \nu) & \text{if } \alpha = 1, \\ \displaystyle\int_{\mathcal{X}} f_\alpha\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\nu & \text{if } \alpha \in (0,1) \text{ or } \alpha > 1 \text{ and } \mu \ll \nu, \\ +\infty & \text{if } \alpha > 1 \text{ and } \mu \not\ll \nu. \end{cases}$$

Proof. First, notice that for all $\alpha \neq 0, 1$, $f'_\alpha(x) = \frac{\alpha}{\alpha - 1} x^{\alpha - 1}$, therefore

$$f'_\alpha(\infty) = \begin{cases} 0 & \text{if } \alpha < 1, \\ +\infty & \text{if } \alpha \geq 1. \end{cases}$$

($\alpha = 0$) This is a simple computation:

$$\mathrm{H}_0(\mu,\nu) = \int_{\mathcal{X}} \mathbb{1}_{\{0\}}\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\nu = \nu\left(\left\{x \in \mathcal{X} \mid \frac{d\mu}{d\nu}(x) = 0\right\}\right).$$

($\alpha = 1$) This follows directly from Proposition 3.2.2.

($\alpha \neq 0, 1$) This is a simple application of the definition of f-divergence, using the value of $f'_\alpha(\infty)$ that we computed above and Proposition B.0.1. ⁂

Definition 3.3.4 (Rényi divergence)**.** Let $\alpha \in [0, +\infty)$, $\mathcal{X}$ be a measurable space and $\mu, \nu \in \mathcal{M}(\mathcal{X})$. We define the *Rényi divergence* of order α between μ and ν as

$$\mathrm{R}_\alpha(\mu,\nu) := \begin{cases} \frac{1}{\alpha-1}\log\left(\nu(\mathcal{X}) + (\alpha-1)\,\mathrm{H}_\alpha(\mu,\nu)\right) & \text{if } \alpha \neq 1, \\ \mathrm{KL}(\mu,\nu) & \text{if } \alpha = 1. \end{cases}$$

The following proposition ensures that our definition is consistent with the one given in [EH14, Sumary table and Remark 1]. The values of the Rényi divergence for the orders 0 and 1, where the other definition would clearly become degenerate, are set as such to ensure continuity when considering $\mathrm{R}_\alpha(\mu,\nu)$ as a function of α, see [EH14] for further details.

Proposition 3.3.5. *Let $\alpha \in [0, +\infty)$, $\mathcal{X}$ be a measurable space and $\mu, \nu \in \mathcal{M}(\mathcal{X})$. Then the Rényi divergence takes the following form:*

$$\mathrm{R}_a(\mu,\nu) = \begin{cases} -\log\left(\nu\left\{x \mid 0 < \frac{d\mu}{d\nu}(x)\right\}\right) & \textit{if } \alpha = 0, \\ \mathrm{KL}(\mu,\nu) & \textit{if } \alpha = 1, \\ \frac{1}{\alpha-1}\log\left(\int \left(\frac{d\mu}{d\nu}\right)^\alpha \mathrm{d}\nu\right) & \textit{if } \alpha \in (0,1) \textit{ or } \alpha > 1 \textit{ and } \mu \ll \nu, \\ +\infty & \textit{if } \alpha > 1 \textit{ and } \mu \not\ll \nu. \end{cases}$$

Proof. **($\alpha = 0$)** By Proposition 3.3.3 we have

$$\begin{aligned} \mathrm{R}_0(\mu,\nu) &= \frac{1}{0-1}\log\left(\nu(\mathcal{X}) + (0-1)\,\mathrm{H}_0(\mu,\nu)\right) \\ &= -\log\left(\nu(\mathcal{X}) - \nu\left(\left\{x \mid \frac{d\mu}{d\nu}(x) = 0\right\}\right)\right) \end{aligned}$$

$$= -\log\left(\nu\left(\mathcal{X} \setminus \left\{x \mid \frac{d\mu}{d\nu}(x) = 0\right\}\right)\right)$$
$$= -\log\left(\nu\left\{x \mid 0 < \frac{d\mu}{d\nu}(x)\right\}\right).$$

($\alpha = 1$) True by definition.

($\alpha \in (0,1)$ or $\alpha > 1$ and $\mu \ll \nu$) By Proposition 3.3.3 we have

$$\begin{aligned}
\mathrm{R}_\alpha(\mu,\nu) &= \frac{1}{\alpha-1}\log\left(\nu(\mathcal{X}) + (\alpha-1)\,\mathrm{H}_\alpha(\mu,\nu)\right) \\
&= \frac{1}{\alpha-1}\log\left(\nu(\mathcal{X}) + (\alpha-1)\int_{\mathcal{X}} f_\alpha\left(\frac{d\mu}{d\nu}\right)\,\mathrm{d}\nu\right) \\
&= \frac{1}{\alpha-1}\log\left(\nu(\mathcal{X}) + (\alpha-1)\int_{\mathcal{X}} \frac{1}{\alpha-1}\left(\left(\frac{d\mu}{d\nu}\right)^\alpha - 1\right)\,\mathrm{d}\nu\right) \\
&= \frac{1}{\alpha-1}\log\left(\nu(\mathcal{X}) + \int_{\mathcal{X}} \left(\frac{d\mu}{d\nu}\right)^\alpha \mathrm{d}\nu - \nu(\mathcal{X})\right) \\
&= \frac{1}{\alpha-1}\log\left(\int_{\mathcal{X}} \left(\frac{d\mu}{d\nu}\right)^\alpha \mathrm{d}\nu\right).
\end{aligned}$$

($\alpha > 1$ and $\mu \not\ll \nu$) By Proposition 3.3.3 we have

$$\begin{aligned}
\mathrm{R}_\alpha(\mu,\nu) &= \frac{1}{\alpha-1}\log\left(\nu(\mathcal{X}) + (\alpha-1)\,\mathrm{H}_\alpha(\mu,\nu)\right) \\
&= \frac{1}{\alpha-1}\log\left(\nu(\mathcal{X}) + (\alpha-1)(+\infty)\right) \\
&= \frac{1}{\alpha-1}\log\left(+\infty\right) = +\infty.
\end{aligned}$$

⋔

Remark 3.3.6. As previously stated, the definition provided for the Hellinger α-divergence is such that the Rényi divergence can be expressed in terms of it, through a specific transformation. Let us consider the case where the measures are normalized, i.e. they are probability measures. Then the transformation that we want to use is the following one:

$$y \mapsto \frac{1}{\alpha-1}\log\left(1 + (\alpha-1)y\right).$$

Notice how this is a nondecreasing function; this feature makes it possible to transfer some properties, like the data processing inequality, from the Hellinger divergence to the Rényi divergence. We can see that when $\alpha = 1$, this transformation becomes degenerate due to the division by zero, so we have to treat this case separately anyway. Then, why do we define H_1 as the KL divergence instead of assigning it some other junk value? This choice can be readily justified by observing that $\lim_{\alpha \to 1} \frac{1}{\alpha - 1} \log(1 + (\alpha - 1)y) = y$ for every $y \in \mathbb{R}$, so it is natural for the Hellinger divergence and the Rényi divergence to coincide for $\alpha = 1$. In the case of more general (finite) measures, the aforementioned transformation would yield a term of the form $1 - \nu(\mathcal{X})$ inside the logarithm, which is not what we want for the Rényi divergence. To address this issue we have to make a slight modification to the transformation, which now depends on the second measure:

$$y \mapsto \frac{1}{\alpha - 1} \log(\nu(\mathcal{X}) + (\alpha - 1)y) .$$

Let us look examine the definitions of the Hellinger and Rényi divergences that we have formalized in Lean.

```
def hellingerFun (a : ℝ) : ℝ → ℝ :=
  if a = 0 then fun x ↦ if x = 0 then 1 else 0
  else if a = 1 then fun x ↦ x * log x
  else fun x ↦ (a - 1)⁻¹ * (x ^ a - 1)

def hellingerDiv (a : ℝ) (μ ν : Measure α) : EReal :=
  fDiv (hellingerFun a) μ ν

def renyiDiv (a : ℝ) (μ ν : Measure α) : EReal :=
  if a = 1 then kl μ ν
  else (a - 1)⁻¹ * ENNReal.log ((↑(ν Set.univ)
    + (a - 1) * (hellingerDiv a μ ν)).toENNReal)
```

The design choices made for the implementation of these mathematical objects are worthy of further observation and analysis. This will be addressed in the following remarks.

Remark 3.3.7. We can notice that some if-else statement are used in the Lean

definitions of the Hellinger function and the Rényi divergence. Managing this kind of branching can be cumbersome, especially when we have to prove results about these objects, since we frequently have to consider each case separately in the proofs. Therefore, choosing the right place to put the branching can make the formalization process easier and the code more maintainable and elegant. As we already mentioned in Remark 3.3.6, with the transformation that we are using for the Rényi divergence, it is impossible to avoid separating the case $\alpha = 1$. In particular, in Lean, the division by zero always yields 0. Therefore, regardless of the value assigned to H_1, the resulting Rényi divergence would have been 0. Thus, in this instance, it was unavoidable to separate this case with an if statement within the definition of the Rényi divergence. However, the situation is different with $\alpha = 0$, and with the Hellinger divergence and function. Indeed, we could have defined R_0 as in Proposition 3.3.5, leaving H_0 as a junk value. An alternative approach would have been to define the Hellinger function without any branching; in this case both $\alpha = 0$ and $\alpha = 1$ would have yielded the zero function[5] and the branching could have been incorporated either in the definition of the Hellinger divergence or in the one of the Rényi divergence.

In our case, the choice has been to push the if statements to the definitions that appear earlier in the code. This approach enables us to take care of the different cases at an early stage, within the API of the Hellinger function, so that many results about it can be used without having to deal with the branching. For instance, the fact that the Hellinger function is convex and continuous is true for all values of α. Moreover, we can directly employ the properties of general f-divergences on the Hellinger divergence. A further reason to put the if statements in the definition of the Hellinger function in our case is that it is the object that likely has the most limited API, given that we only required certain lemmas pertaining to its continuity, measurability, convexity, and integrability.

[5]Lean interprets division by zero as zero, and `x ^ 0` as 1 for every x, in particular for $x = 0$. So $\frac{x^0-1}{0-1} = -(1-1) = 0$ and $\frac{x^1-1}{1-1} = \frac{x-1}{0} = 0$ for every x.

Another possible choice could have been to avoid defining the Hellinger function entirely and instead incorporate the if-else statements in the definition of the Hellinger divergence, setting it as the f-divergence with $x \mapsto \frac{x^\alpha - 1}{\alpha - 1}$ for $\alpha \neq 0, 1$. Nevertheless, this approach would likely have resulted in greater code duplication. In general, dividing a complex definition into smaller parts facilitates the work, and improves code readability and maintainability. Furthermore, as the function has been given a name, it is now easier to locate relevant lemmas and references pertaining to it in the proofs.

Remark 3.3.8. In the definition of the Rényi divergence we use as logarithm the function `ENNReal.log`, which is defined as follows:

```
def ENNReal.log (x : ℝ≥0∞) : EReal :=
  if x = 0 then ⊥
    else if x = ⊤ then ⊤
    else Real.log x.toReal
```

where `Real.log` is the standard natural logarithm function in Mathlib. This function and its API have been developed in part during the course of this project, along with its inverse function, `EReal.exp`, and are currently included into Mathlib. The main reason for using this function instead of `Real.log` is due to the latter's anomalous behavior at 0, in particular `Real.log` 0 is defined[6] as 0, which renders it not monotone on $[0, +\infty)$. Moreover, `ENNReal.log` can take infinite values as arguments, which can happen in our use case; attempting to handle these cases with `Real.log` would have involved more if statements. Lastly, the codomain of `ENNReal.log`, `EReal`, is more suitable for our purposes, since it is the same as the codomain of the Rényi divergence, and allows us to avoid some unnecessary coercions.

[6] See the Mathlib documentation for more details about the motivation of such choice.

Chapter 4

Hypothesis Testing

Hypothesis testing represent a special case of the estimation problem, which can be defined as the task of estimating a particular feature of an underlying phenomenon, based on observed outcomes and a set of potential models for that phenomenon. In particular, one can devise a (possibly stochastic) function, referred to as an estimator, that takes the observed data as input and produces an estimate of the feature of interest. An interesting question then arises as to the means of assessing the performance of the test and the optimal performance that can be attained on a given testing problem.

For our purposes, we will focus on a specific type of estimation problem, known as simple binary hypothesis testing. In this context, the space of potential models is limited to two elements, each corresponding to a measure from which the observed data is sampled. The objective is to determine which of the two is the most probable source of the data.

In addition, we will define another information divergence, called statistical divergence, which will play a crucial role in a proof of the DPI in Chapter 5.

Definition 4.0.1 (Estimation problem)**.** Let $\Theta, \mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be measurable spaces, $P\colon \Theta \rightsquigarrow \mathcal{X}$ the data generating kernel from the parameter space to the sample space and two measurable functions $y\colon \Theta \to \mathcal{Y}$ and $\ell\colon \mathcal{Y} \times \mathcal{Z} \to \overline{\mathbb{R}}_+$, called respectively the *objective function* and *loss function*. Then we say that (P, y, ℓ) is an *estimation problem*, and a Markov kernel $\hat{y}\colon \mathcal{X} \rightsquigarrow \mathcal{Z}$ is said to

be *estimator*. Moreover, we say that the estimation problem is:

- a *parametric* problem if Θ is low dimensional, a *nonparametric* problem otherwise.
- a *testing* problem if $\mathcal{Y} = \mathcal{Z}$ is discrete and ℓ only takes the values 0 and 1.
- *binary* if it is a testing problem and $\mathcal{Y}$ has exactly 2 elements.
- *simple* if it is a testing problem and y is a bijection.

In conclusion, in simple binary hypothesis testing we can assume $\Theta = \mathcal{Y} = \mathcal{Z} = \{0, 1\}$, y to be the identity and $\ell(y_0, y_1) = \mathbb{1}_{\{y_0 \neq y_1\}}$ for $y_0, y_1 \in \mathcal{Y}$.

Remark 4.0.2. In principle, to formalize statements about an estimation problem, it is sufficient to introduce the objects that define the problem, namely P, y and ℓ. However, since they are frequently used in conjunction, it is advisable to bundle them into a single object, so we can introduce them all at once when needed, thereby reducing the length and complexity of the statements. In Lean, this can be achieved using a structure, as illustrated in the following example:

```
structure estimationProblem (Θ 𝒳 𝒴 𝒵 : Type*) [MeasurableSpace Θ]
    [MeasurableSpace 𝒳] [MeasurableSpace 𝒴] [MeasurableSpace 𝒵] :=
  P : kernel Θ 𝒳
  y : Θ → 𝒴
  y_meas : Measurable y
  ℓ : 𝒴 × 𝒵 → ℝ≥0∞
  ℓ_meas : Measurable ℓ
```

Notice how the structure includes not only the data of the problem, i.e. P, y and ℓ, but also the measurability properties.

This structure was initially used in the project, however, after working with it for some time, it became evident that it made the formalization very cumbersome every time we had to perform an operation (for example

composition with some kernel κ) on P, as the entire structure had to be redefined with the new kernel, even though the rest of the structure remained unchanged. For this reason, we decided to remove P from the structure:

```
structure estimationProblem (Θ 𝒴 𝒵 : Type*) [MeasurableSpace Θ]
  [MeasurableSpace 𝒴] [MeasurableSpace 𝒵] :=
 y : Θ → 𝒴
 y_meas : Measurable y
 ℓ : 𝒴 × 𝒵 → ℝ≥0∞
 ℓ_meas : Measurable ℓ
```

Remark 4.0.3. For the case of simple binary hypothesis testing, we defined an element of type `estimationProblem`, with the specific data that characterize this particular problem, as explained above:

```
def simpleBinaryHypTest : estimationProblem Bool Bool Bool where
 y := id
 y_meas := measurable_id
 ℓ := fun (y, z) ↦ if y = z then 0 else 1
 ℓ_meas := Measurable.of_discrete
```

Where `Bool` is the type of boolean values, containing the elements `true` and `false`, which is used to represent the space $\{0, 1\}$. Notice how in the fields corresponding to the measurability of the functions, we provide a proof that such property holds true.

Before introducing the concept of risk, let us give an example of estimation problem.

Example 4.0.4. Let $n \in \mathbb{N}$, $\Theta = \mathcal{Y} = \mathcal{Z} = [0, 1]$, $\mathcal{X} := \{0, 1\}^n$, $y \colon \Theta \to \mathcal{Y}$ the identity function, $\ell(p, q) = \mathbb{1}_{\{p \neq q\}}$ for $p, q \in [0, 1]$ and $P \colon \Theta \rightsquigarrow \mathcal{X}$ a kernel defined by $P(p) = Ber(p)^{\otimes n}$, where $Ber(p)$ is the Bernoulli distribution with mean p.

We can see how this problem is a parametric simple testing problem; however, it is not binary, as there are more than two potential measures for

the data (we could make it binary by fixing only two possible parameters p and q).

This problem can be interpreted as follows: we are observing a phenomenon that repeats periodically and can always have one of two outcomes (for example the fact that an object produced by a machine is defective or not); we want to estimate what is the probability of each outcome, having observed the phenomenon occur n times.

A potential estimator is the empirical mean of the observed data:

$$\hat{y}(x) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Thus, if we observe the data $x = (1, 0, 1, 1, 1)$, the estimator would produce the output $\hat{y}(x) = \frac{4}{5}$.

It follows from the Law of Large Numbers that, as the value of n increases, the performance of the estimator $\hat{y}$ will improve, and in general the task of accurately estimating the parameter p will become easier. Therefore, an interesting question is to determine the smallest value of n such that the estimator performs well enough, also known as the sample complexity of the problem. Information theoretic tools and inequalities, such as the DPI, can be employed to provide lower bounds for the sample complexity of a problem.

4.1 Risk

An important problem in estimation is to assess the performance of an estimator, based on its capacity to accurately predict the value of the objective function. One method for evaluating an estimator is to fix a parameter $\theta \in \Theta$ and consider the average loss of the estimator when the data is generated according to the distribution $P(\theta)$, which is referred to as the risk of the estimator. But we may be interested in different kinds or risk. For instance, instead of selecting a fixed parameter, we may want to fix a *prior* $\pi \in \mathcal{M}(\Theta)$, which represents our beliefs about the distribution of the parameter before

having observed the data, and consider the average risk over all possible values of the parameter. In this section we introduce different types of risk.

Definition 4.1.1 (Risk)**.** Let (P, y, ℓ) be an estimation problem, $\hat{y}$ an estimator and $\theta \in \Theta$. The *risk* of $\hat{y}$ at θ is defined as

$$r_\theta^P(\hat{y}) = (\hat{y} \circ P)(\theta) \left[z \mapsto \ell(y(\theta), z)\right].$$

Definition 4.1.2 (Bayesian risk)**.** Let (P, y, ℓ) be an estimation problem, $\hat{y}$ an estimator and $\pi \in \mathcal{M}(\Theta)$. The *Bayesian risk* of $\hat{y}$ for the prior π is defined as

$$R_\pi^P(\hat{y}) = \pi\left[\theta \mapsto r_\theta^P(\hat{y})\right] = (\pi \otimes (\hat{y} \circ P))\left[(\theta, z) \mapsto \ell(y(\theta), z)\right].$$

Remark 4.1.3. Through simple kernel calculations, it is possible to prove that $R_\pi^P(\hat{y})$ is equal to the mean of the measure $\ell \circ (y \parallel \hat{y}) \circ (\pi \otimes P)$.

Additionally, we present an alternative formula and a lower bound for the Bayesian risk in terms of the Bayesian inverse of P. Proposition 4.1.10 will show how this lower bound can in fact be achieved in certain cases.

Proposition 4.1.4. *Let (P, y, ℓ) be an estimation problem, $\hat{y}$ an estimator and $\pi \in \mathcal{M}(\Theta)$ such that the Bayesian inverse $P_\pi^\dagger$ exists. Then:*

i) $R_\pi^P(\hat{y}) = ((P_\pi^\dagger \times \hat{y}) \circ P \circ \pi)\left[(\theta, z) \mapsto \ell(y(\theta), z)\right]$,

ii) $R_\pi^P(\hat{y}) = (P \circ \pi)\left[x \mapsto \hat{y}(x)\left[z \mapsto P_\pi^\dagger(x)\left[\theta \mapsto \ell(y(\theta), z)\right]\right]\right]$,

iii) $R_\pi^P(\hat{y}) \geq (P \circ \pi)\left[x \mapsto \inf_{z \in \mathcal{Z}} P_\pi^\dagger(x)\left[\theta \mapsto \ell(y(\theta), z)\right]\right]$.

Proof. **(i)** To prove the first equality, we recall the definition of the Bayesian risk $R_\pi^P(\hat{y}) = (\pi \otimes (\hat{y} \circ P))\left[(\theta, z) \mapsto \ell(y(\theta), z)\right]$, and then perform some kernel operations using Proposition A.0.2:

$$\begin{aligned}
(P_\pi^\dagger \times \hat{y}) \circ P \circ \pi &= (\mathrm{id} \parallel \hat{y}) \circ (P_\pi^\dagger \times \mathrm{id}) \circ (P \circ \pi) \\
\text{By Equation (2.2)} \qquad &= (\mathrm{id} \parallel \hat{y}) \circ (\mathrm{id} \times P) \circ \pi \\
\text{By Proposition A.0.2} \qquad &= (\mathrm{id} \times (\hat{y} \circ P)) \circ \pi \\
\text{By Remark 2.1.18} \qquad &= \pi \otimes (\hat{y} \circ P).
\end{aligned}$$

(ii) We continue from the equality in (i):

$$\begin{aligned} R_\pi^P(\hat{y}) &= ((P_\pi^\dagger \times \hat{y}) \circ P \circ \pi)\,[(\theta, z) \mapsto \ell(y(\theta), z)] \\ &= (P \circ \pi)\left[x \mapsto (P_\pi^\dagger(x) \times \hat{y}(x))\,[(\theta, z) \mapsto \ell(y(\theta), z)]\right] \\ \text{By Fubini's theorem} \quad &= (P \circ \pi)\left[x \mapsto \hat{y}(x)\left[z \mapsto P_\pi^\dagger(x)\,[\theta \mapsto \ell(y(\theta), z)]\right]\right]. \end{aligned}$$

(iii) To prove the last inequality it is sufficient to use (ii) and the fact that, since $\hat{y}$ is a Markov kernel, we have

$$\begin{aligned} \hat{y}(x)\left[z \mapsto P_\pi^\dagger(x)\,[\theta \mapsto \ell(y(\theta), z)]\right] &\geq \inf_{z \in \mathcal{Z}} P_\pi^\dagger(x)\,[\theta \mapsto \ell(y(\theta), z)]\,\hat{y}(\mathcal{X}) \\ &= \inf_{z \in \mathcal{Z}} P_\pi^\dagger(x)\,[\theta \mapsto \ell(y(\theta), z)]\,. \end{aligned}$$

⋔

Definition 4.1.5 (Bayes risk)**.** Let (P, y, ℓ) be an estimation problem and $\pi \in \mathcal{M}(\Theta)$. The *Bayes risk* of (P, y, ℓ) for the prior π is defined as

$$\mathcal{R}_\pi^P = \inf_{\hat{y}: \mathcal{X} \rightsquigarrow \mathcal{Z}} R_\pi^P(\hat{y}),$$

where the infimum is taken over Markov kernels.

Moreover, an estimator $\hat{y}$ is said to be a *Bayes estimator* for the prior π if it achieves the infimum of the Bayesian risk, that is, if $R_\pi^P(\hat{y}) = \mathcal{R}_\pi^P$.

There are alternative approaches to quantifying the risk of an estimation problem that do not require the specification of a prior, such as the minimax risk. However, these will not be considered in this project.

The following proposition shows us that the composition of a Markov kernel with the data generating kernel cannot reduce the risk, i.e. it can only make the estimation task harder. This can be regarded as a form of data processing inequality, which will be discussed in detail in Chapter 5. Indeed, one of the proofs of the data processing inequality for the f-divergences relies on this result.

Proposition 4.1.6. *Let (P, y, ℓ) be an estimation problem, $\pi \in \mathcal{M}(\Theta)$ and $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{X}'$ a Markov kernel. Then we have*

$$\mathcal{R}_\pi^P \leq \mathcal{R}_\pi^{\kappa \circ P}.$$

Proof. It is sufficient to observe that composing a kernel with P produces the same Bayesian risk as composing the estimator with the same kernel:

$$R_\pi^{\kappa \circ P}(\hat{y}) = (\pi \otimes (\hat{y} \circ \kappa \circ P))\,[(\theta, z) \mapsto \ell(y(\theta), z)] = R_\pi^P(\hat{y} \circ \kappa).$$

Then, we can conclude using the fact that an infimum is greater when taken over a smaller set; in this case the sets are the set of all Markov kernels $\mathcal{X} \rightsquigarrow \mathcal{Z}$ and the set of all Markov kernels of the form $\hat{y}' \circ \kappa : \mathcal{X} \rightsquigarrow \mathcal{Z}$ with $\hat{y}' : \mathcal{X}' \rightsquigarrow \mathcal{Z}$:

$$\mathcal{R}_\pi^P = \inf_{\hat{y}:\mathcal{X}\rightsquigarrow\mathcal{Z}} R_\pi^P(\hat{y}) \leq \inf_{\hat{y}':\mathcal{X}'\rightsquigarrow\mathcal{Z}} R_\pi^P(\hat{y}' \circ \kappa) = \inf_{\hat{y}':\mathcal{X}'\rightsquigarrow\mathcal{Z}} R_\pi^{\kappa\circ P}(\hat{y}') = \mathcal{R}_\pi^{\kappa\circ P}.$$

⩛

Definition 4.1.7 (Generalized Bayes estimator)**.** Let (P, y, ℓ) be an estimation problem, $\pi \in \mathcal{M}(\Theta)$ and $f\colon \mathcal{X} \to \mathcal{Z}$ a measurable function. We say that f is a *generalized Bayes estimator* for (P, y, ℓ) with respect to the prior π if it is of the form $x \mapsto \operatorname{argmin}_z P_\pi^\dagger(x)[\theta \mapsto \ell(y(\theta), z)]$ for almost every x with respect to $P \circ \pi$.

Remark 4.1.8. Note that the generalized Bayes estimator is a deterministic kernel. This highlights the fact that, even though we allow the estimator to include some randomness, the optimal performance can be attained by a deterministic function, at least in the case where a generalized Bayes estimator exists.

Remark 4.1.9. Just like the estimation problem, the generalized Bayes estimator can be defined using a structure in Lean, as it carries more than one piece of information: the measurability of the function and the fact that it minimizes some quantity.

```
structure IsGenBayesEstimator [StandardBorelSpace Θ] [Nonempty Θ]
  (E : estimationProblem Θ 𝒴 𝒵) (P : Kernel Θ 𝒳) [IsFiniteKernel P]
  (f : 𝒳 → 𝒵) (π : Measure Θ) [IsFiniteMeasure π] : Prop where
 measurable : Measurable f
 property : ∀ᵐ x ∂(P ∘ₘ π), ∫⁻ θ, E.ℓ (E.y θ, f x) ∂(P†π) x
  = ⨅ z, ∫⁻ θ, E.ℓ (E.y θ, z) ∂(P†π) x
```

We also defined a class for the property of an estimation problem to admit a generalized Bayes estimator:

```
class HasGenBayesEstimator [StandardBorelSpace Θ] [Nonempty Θ]
  (E : estimationProblem Θ 𝒴 𝒵) (P : Kernel Θ 𝒳) [IsFiniteKernel P]
  (π : Measure Θ) [IsFiniteMeasure π] where
 estimator : 𝒳 → 𝒵
 property : IsGenBayesEstimator E P estimator π
```

We could have avoided the definition this class by simply introducing the generalized Bayes estimator whenever we need it, with hypotheses of the form (`f` : $\mathcal{X} \to \mathcal{Z}$) (`hf` : `IsGenBayesEstimator E P f` π). Nevertheless, there are certain results, such as Proposition 4.1.10, that require the existence of a generalized Bayes estimator but do not make use of it in the statement. In such instances, it is preferable to have the property of admitting a generalized Bayes estimator, thus obviating the need to introduce a function that will not be used. Moreover, using a class rather than a structure allows us to use it as an argument inside the square brackets, enabling Lean to infer it automatically through the typeclass inference system whenever an instance of the property is already available. This is the case, for example, of the simple binary hypothesis testing problem, as outlined in the proof of Proposition 4.2.3.

The next result shows that if a generalized Bayes estimator exists, then it is a Bayes estimator, and it achieves the lower bound for the Bayesian risk given in Proposition 4.1.4.

Proposition 4.1.10. *Let (P, y, ℓ) be an estimation problem, $\pi \in \mathcal{M}(\Theta)$ and $\hat{y}_B$ a generalized Bayes estimator for (P, y, ℓ) with respect to π. Then we have*

$$R_\pi^P(\hat{y}_B) = (P \circ \pi)\left[x \mapsto \inf_{z \in \mathcal{Z}} P_\pi^\dagger(x)\left[\theta \mapsto \ell(y(\theta), z)\right]\right].$$

In particular, $\hat{y}_B$ is a Bayes estimator and

$$\mathcal{R}_\pi^P = R_\pi^P(\hat{y}_B) = (P \circ \pi)\left[x \mapsto \inf_{z \in \mathcal{Z}} P_\pi^\dagger(x)\left[\theta \mapsto \ell(y(\theta), z)\right]\right].$$

Proof. Let us start by using Proposition 4.1.4:

$$\begin{aligned} R_\pi^P(\hat{y}_B) &= (P \circ \pi)\left[x \mapsto \hat{y}_B(x)\left[z \mapsto P_\pi^\dagger(x)\left[\theta \mapsto \ell(y(\theta), z)\right]\right]\right] \\ \text{\scriptsize $\hat{y}_B(x)$ is deterministic} &= (P \circ \pi)\left[x \mapsto \delta_{\hat{y}_B(x)}\left[z \mapsto P_\pi^\dagger(x)\left[\theta \mapsto \ell(y(\theta), z)\right]\right]\right] \\ &= (P \circ \pi)\left[x \mapsto P_\pi^\dagger(x)\left[\theta \mapsto \ell(y(\theta), \hat{y}_B(x))\right]\right] \\ \text{\scriptsize By definition of $\hat{y}_B$} &= (P \circ \pi)\left[x \mapsto \min_z P_\pi^\dagger(x)[\theta \mapsto \ell(y(\theta), z)]\right]. \end{aligned}$$

So $\hat{y}_B$ achieves the lower bound from Proposition 4.1.4, therefore it is a Bayes estimator and the Bayes risk is equal to its Bayesian risk. ⫲

4.2 Binary risk and statistical information

In the context of simple binary hypothesis testing, the data generating kernel is of the form $P\colon \{0,1\} \rightsquigarrow \mathcal{X}$, so we can view it as a pair of measures $\mu \coloneqq P(0)$ and $\nu \coloneqq P(1)$ and will sometimes write $P = (\mu, \nu)$. Moreover, a prior $\pi \in \mathcal{M}(\{0,1\})$ can be viewed as a pair of numbers in $[0, +\infty]$, representing the mass that π assigns to 0 and 1 respectively; we will write, with a slight abuse of notation, $\pi = (\pi(0), \pi(1))$. This framework allows us to define certain quantities that can be used to compare the two measures.

First of all, let us see how the definitions given in the previous section can be simplified in the case of simple binary hypothesis testing.

Proposition 4.2.1. *Let (P, y, ℓ) be a simple binary hypothesis testing problem, $\hat{y}\colon \mathcal{X} \rightsquigarrow \{0,1\}$ an estimator and $\pi = (\pi_0, \pi_1) \in \mathcal{M}(\{0,1\})$. Then:*

i) $P \circ \pi = \pi_0\mu + \pi_1\nu$,

ii) $r_0^P(\hat{y}) = (\hat{y} \circ \mu)(\{1\})$,

iii) $r_1^P(\hat{y}) = (\hat{y} \circ \nu)(\{0\})$,

iv) $R_\pi^P(\hat{y}) = \pi_0(\hat{y} \circ \mu)(\{1\}) + \pi_1(\hat{y} \circ \nu)(\{0\})$.

Proof. Simple computations. ⫲

To perform a similar calculation for the Bayes risk, it is first helpful to compute the Bayesian inverse of P.

Lemma 4.2.2. *Let $P\colon \{0,1\} \rightsquigarrow \mathcal{X}$ be a kernel, $\mu \coloneqq P(0)$, $\nu \coloneqq P(1)$ and $\pi = (\pi_0, \pi_1) \in \mathcal{M}(\{0,1\})$. Then the Bayesian inverse of P with respect to π is given by*

$$P_\pi^\dagger(x) = \left(\pi_0 \frac{d\mu}{dP \circ \pi}(x), \pi_1 \frac{d\nu}{dP \circ \pi}(x)\right)$$

for almost every $x \in \mathcal{X}$ with respect to $P \circ \pi$.

Proof. Let us denote with $\eta\colon \mathcal{X} \rightsquigarrow \{0,1\}$ the kernel on the right-hand side of the equation. Since the Bayesian inverse is unique up to a set of measure zero, it is enough to show that η satisfies the defining property of the Bayesian inverse:

$$(P \circ \pi) \otimes \eta \overset{?}{=} (\pi \otimes P)_\leftrightarrow.$$

This is an equality between measures over $\mathcal{X} \times \{0,1\}$, therefore it is sufficient to prove it on sets of the form $A \times \{0\}$ and $A \times \{1\}$ for $A \subseteq \mathcal{X}$ measurable.

$$\begin{aligned}
(\pi \otimes P)_\leftrightarrow(A \times \{0\}) &= \pi \otimes P(\{0\} \times A) \\
\text{By definition of } \otimes,\ 2.1.10 \quad &= \int_{\{0\}} P(\theta, A)\pi(d\theta) \\
&= \pi(\{0\})P(0, A) \\
&= \pi_0 \mu(A).
\end{aligned}$$

On the other hand, using the definition of the composition product once more, we have

$$\begin{aligned}
(P \circ \pi) \otimes \eta(A \times \{0\}) &= \int_A \eta(x, \{0\})(P \circ \pi)(dx) \\
\text{By definition of } \eta \quad &= \int_A \pi_0 \frac{d\mu}{dP \circ \pi}(x)(P \circ \pi)(dx) \\
\text{By Proposition B.0.1} \quad &= \int_A \pi_0 \mu(dx) \\
&= \pi_0 \mu(A) = (\pi \otimes P)_\leftrightarrow(A \times \{0\}),
\end{aligned}$$

where Proposition B.0.1 can be used since $\mu \ll P \circ \pi = \pi_0\mu + \pi_1\nu$ if $\pi_0 \neq 0$; in the case where $\pi_0 = 0$ the equality holds trivially.

Similar computations show that the equality also holds for $A \times \{1\}$. ⩛

Proposition 4.2.3. *Let (P, y, ℓ) be a simple binary hypothesis testing problem and $\pi = (\pi_0, \pi_1) \in \mathcal{M}(\{0,1\})$. Then the Bayes risk of the problem for the prior π is given by*

$$\mathcal{R}_\pi^P = (P \circ \pi)\left[x \mapsto \min\left\{\pi_0 \frac{d\mu}{dP \circ \pi}(x), \pi_1 \frac{d\nu}{dP \circ \pi}(x)\right\}\right].$$

Proof. First, we demonstrate that the simple binary hypothesis testing problem admits a generalized Bayes estimator. Let us consider as a candidate the function $\hat{y}_B \coloneqq \mathbb{1}_E$, with $E \coloneqq \left\{x \in \mathcal{X} \mid \pi_0 \frac{d\mu}{dP \circ \pi}(x) \leq \pi_1 \frac{d\nu}{dP \circ \pi}(x)\right\}$. We have that $\hat{y}_B$ is measurable, since E is a measurable set, so we only need to show that for every $z \in \{0,1\}$ the following holds:

$$P_\pi^\dagger(x)[\theta \mapsto \ell(y(\theta), \hat{y}_B(x))] \overset{?}{\leq} P_\pi^\dagger(x)[\theta \mapsto \ell(y(\theta), z)].$$

By the definitions of y and ℓ, and Lemma 4.2.2 we have

$$\begin{aligned} P_\pi^\dagger(x)[\theta \mapsto \ell(y(\theta), z)] &= \pi_0 \frac{d\mu}{dP \circ \pi}(x)\ell(y(0), z) + \pi_1 \frac{d\nu}{dP \circ \pi}(x)\ell(y(1), z) \\ &= \pi_0 \frac{d\mu}{dP \circ \pi}(x)\mathbb{1}_{\{0 \neq z\}} + \pi_1 \frac{d\nu}{dP \circ \pi}(x)\mathbb{1}_{\{1 \neq z\}} \\ &= \pi_0 \frac{d\mu}{dP \circ \pi}(x)\mathbb{1}_{\{z=1\}} + \pi_1 \frac{d\nu}{dP \circ \pi}(x)\mathbb{1}_{\{z=0\}}. \end{aligned}$$

Hence if $\hat{y}_B(x) = 0$ it follows that $\pi_0 \frac{d\mu}{dP \circ \pi}(x) \geq \pi_1 \frac{d\nu}{dP \circ \pi}(x)$ and

$$\begin{aligned} P_\pi^\dagger(x)[\theta \mapsto \ell(y(\theta), \hat{y}_B(x))] &= \pi_0 \frac{d\mu}{dP \circ \pi}(x)\mathbb{1}_{\{\hat{y}_B(x)=1\}} + \pi_1 \frac{d\nu}{dP \circ \pi}(x)\mathbb{1}_{\{\hat{y}_B(x)=0\}} \\ &= \pi_1 \frac{d\nu}{dP \circ \pi}(x) \\ &\leq \pi_0 \frac{d\mu}{dP \circ \pi}(x)\mathbb{1}_{\{0 \neq z\}} + \pi_1 \frac{d\nu}{dP \circ \pi}(x)\mathbb{1}_{\{1 \neq z\}}. \end{aligned}$$

On the other hand, if $\hat{y}_B(x) = 1$ we have $\pi_0 \frac{d\mu}{dP \circ \pi}(x) \leq \pi_1 \frac{d\nu}{dP \circ \pi}(x)$ and

$$P_\pi^\dagger(x)[\theta \mapsto \ell(y(\theta), \hat{y}_B(x))] = \pi_0 \frac{d\mu}{dP \circ \pi}(x)\mathbb{1}_{\{\hat{y}_B(x)=1\}} + \pi_1 \frac{d\nu}{dP \circ \pi}(x)\mathbb{1}_{\{\hat{y}_B(x)=0\}}$$

$$
\begin{aligned}
&= \pi_0 \frac{d\mu}{dP \circ \pi}(x) \\
&\leq \pi_0 \frac{d\mu}{dP \circ \pi}(x) \mathbb{1}_{\{0 \neq z\}} + \pi_1 \frac{d\nu}{dP \circ \pi}(x) \mathbb{1}_{\{1 \neq z\}}.
\end{aligned}
$$

In both cases it is straightforward to check that the last inequality holds whether $z = 0$ or $z = 1$.

Therefore, $\hat{y}_B$ is a generalized Bayes estimator, and we can conclude the proof by applying Proposition 4.1.10 as follows:

$$
\begin{aligned}
\mathcal{R}_\pi^P &= (P \circ \pi) \left[x \mapsto \inf_{z \in \{0,1\}} P_\pi^\dagger(x) \left[\theta \mapsto \ell(y(\theta), z) \right] \right] \\
\text{As above} &= (P \circ \pi) \left[x \mapsto \inf_{z \in \{0,1\}} \left(\pi_0 \frac{d\mu}{dP \circ \pi}(x) \mathbb{1}_{\{z=1\}} + \pi_1 \frac{d\nu}{dP \circ \pi}(x) \mathbb{1}_{\{z=0\}} \right) \right] \\
&= (P \circ \pi) \left[x \mapsto \min \left\{ \pi_0 \frac{d\mu}{dP \circ \pi}(x), \pi_1 \frac{d\nu}{dP \circ \pi}(x) \right\} \right].
\end{aligned}
$$

⋔

Remark 4.2.4. Lemma 4.2.2 is, in fact, just a special case of a more general result, stating that for almost every x and θ we have

$$
\frac{dP_\pi^\dagger(x)}{d\pi}(\theta) = \frac{dP(\theta)}{d(P \circ \pi)}(x).
$$

Moreover, Proposition 4.2.3 can be extended to more general cases of estimation problems, where the only requirement is for the parameter space to be equal to $\{0, 1\}$.

Definition 4.2.5 (Bayes binary risk)**.** Let $\mu, \nu \in \mathcal{M}(\mathcal{X})$ and $\pi \in \mathcal{M}(\{0, 1\})$. The *Bayes binary risk* of μ and ν with respect to the prior π, denoted as $\mathcal{B}_\pi(\mu, \nu)$, is defined as the Bayes risk of the simple binary hypothesis testing problem with data generating kernel $P = (\mu, \nu)$ and prior π.

Remark 4.2.6. Two expressions for the Bayes binary risk can be readily derived, one using the definition of the Bayes risk and Proposition 4.2.1:

$$
\mathcal{B}_\pi(\mu, \nu) = \inf_{\hat{y}: \mathcal{X} \rightsquigarrow \{0,1\}} \pi_0(\hat{y} \circ \mu)(\{1\}) + \pi_1(\hat{y} \circ \nu)(\{0\}),
$$

and the other using Proposition 4.2.3:

$$\mathcal{B}_\pi(\mu,\nu) = (P\circ\pi)\left[x \mapsto \min\left\{\pi_0 \frac{d\mu}{dP\circ\pi}(x), \pi_1 \frac{d\nu}{dP\circ\pi}(x)\right\}\right]. \tag{4.1}$$

The binary risk inherits a kind of data processing inequality from the Bayes risk, as shown in the following proposition.

Proposition 4.2.7. *Let $\mu,\nu \in \mathcal{M}(\mathcal{X})$, $\pi \in \mathcal{M}(\{0,1\})$ and $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ a Markov kernel. Then*

$$\mathcal{B}_\pi(\mu,\nu) \leq \mathcal{B}_\pi(\kappa\circ\mu, \kappa\circ\nu).$$

Proof. This is an immediate application of Proposition 4.1.6, after noticing that $\kappa\circ P = (\kappa\circ\mu, \kappa\circ\nu)$. ⁂

We can now define a new information divergence, the statistical information, which will prove to be crucial in the following chapter.

Definition 4.2.8 (Statistical information)**.** Let $\mu,\nu \in \mathcal{M}(\mathcal{X})$ and $\pi \in \mathcal{M}(\{0,1\})$. We define the *statistical information* between μ and ν with respect to the prior π as

$$\mathcal{I}_\pi(\mu,\nu) = \min\{\pi_0\mu(\mathcal{X}), \pi_1\nu(\mathcal{X})\} - \mathcal{B}_\pi(\mu,\nu).$$

Remark 4.2.9. Consider an estimator $\hat{y}$ for the simple binary hypothesis testing problem that is not allowed to look at the data. In other words, $\hat{y}$ is a constant kernel, that is, a measure. We can use Proposition 4.2.1 and Proposition A.0.3 to compute the Bayesian risk of $\hat{y}$:

$$R^P_\pi(\hat{y}) = \pi_0(\hat{y}\circ\mu)(\{1\}) + \pi_1(\hat{y}\circ\nu)(\{0\}) = \pi_0\mu(\mathcal{X})\hat{y}(\{1\}) + \pi_1\nu(\mathcal{X})\hat{y}(\{0\}).$$

Since $\hat{y}$ is a Markov kernel, the quantity on the right is a convex combination of $\pi_0\mu(\mathcal{X})$ and $\pi_1\nu(\mathcal{X})$, so the infimum over all possible $\hat{y}$ is achieved by taking $\hat{y}$ as a constant function[1], so that the convex combination collapses over the smallest of the two values. In other words, the minimum is $\min\{\pi_0\mu(\mathcal{X}), \pi_1\nu(\mathcal{X})\}$.

[1]Such an estimator always bets on the outcome that is more likely according to the prior. We will refer to this estimator as "blind," as it is unable to examine the data.

In light of this, the statistical information can be interpreted as representing the difference between the risk of the best possible estimator and the risk of the best possible blind estimator. It is reasonable to consider this quantity a divergence, as when the two measures are similar, observing the data does not significantly enhance the estimator's performance. Conversely, when the measures are very different, employing a blind estimator can markedly impact outcomes.

Remark 4.2.10. Considering an estimation problem and then optimizing over the blind estimators is equivalent to considering the same estimation problem with the data erased and then optimizing over all possible estimators. The appropriate tool to erase some data in the context of information theory is the discard kernel, which we will denote here by $d_{\mathcal{X}}$. Using the discard kernel we can write

$$\min\{\pi_0\mu(\mathcal{X}), \pi_1\nu(\mathcal{X})\} = \mathcal{B}_\pi(d_{\mathcal{X}} \circ \mu, d_{\mathcal{X}} \circ \nu).$$

Thus, we can see the statistical information as a difference of two Bayes risks

$$\mathcal{I}_\pi(\mu, \nu) = \mathcal{B}_\pi(d_{\mathcal{X}} \circ \mu, d_{\mathcal{X}} \circ \nu) - \mathcal{B}_\pi(\mu, \nu). \tag{4.2}$$

In the Lean code, Equation (4.2) is used as a definition for the statistical information:

```
def statInfo (μ ν : Measure 𝒳) (π : Measure Bool) : ℝ≥0∞ :=
  bayesBinaryRisk (Kernel.discard 𝒳 ∘ₘ μ) (Kernel.discard 𝒳 ∘ₘ ν) π
    - bayesBinaryRisk μ ν π
```

Furthermore, Equation (4.2) allows us to easily prove the non-negativity of the statistical information, leveraging the data processing inequality for the Bayes risk.

Proposition 4.2.11. *Let* $\mu, \nu \in \mathcal{M}(\mathcal{X})$ *and* $\pi \in \mathcal{M}(\{0, 1\})$. *Then*

$$\mathcal{I}_\pi(\mu, \nu) \geq 0.$$

Proof. The discard kernel is a Markov kernel (see Remark 2.1.8), so we can apply Proposition 4.2.7 to get

$$\mathcal{B}_\pi(\mu,\nu) \leq \mathcal{B}_\pi(d_{\mathcal{X}} \circ \mu, d_{\mathcal{X}} \circ \nu).$$

Then we can use Equation (4.2) to conclude the proof. ⫲

The data processing inequality for the Bayes risk can also be extended to the statistical information, as shown in the following proposition.

Proposition 4.2.12 (DPI for the statistical information)**.** *Let $\mu, \nu \in \mathcal{M}(\mathcal{X})$, $\pi \in \mathcal{M}(\{0,1\})$ and $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ a Markov kernel. Then*

$$\mathcal{I}_\pi(\kappa \circ \mu, \kappa \circ \nu) \leq \mathcal{I}_\pi(\mu,\nu).$$

Proof. By Proposition 2.1.19, the composition with a Markov kernel does not modify the total mass of a measure, so we have

$$\begin{aligned}
\mathcal{I}_\pi(\kappa \circ \mu, \kappa \circ \nu) &= \min\{\pi_0(\kappa \circ \mu)(\mathcal{Y}), \pi_1(\kappa \circ \nu)(\mathcal{Y})\} - \mathcal{B}_\pi(\kappa \circ \mu, \kappa \circ \nu) \\
\text{By Proposition 2.1.19} \quad &= \min\{\pi_0\mu(\mathcal{X}), \pi_1\nu(\mathcal{X})\} - \mathcal{B}_\pi(\kappa \circ \mu, \kappa \circ \nu) \\
\text{By Proposition 4.2.7} \quad &\leq \min\{\pi_0\mu(\mathcal{X}), \pi_1\nu(\mathcal{X})\} - \mathcal{B}_\pi(\mu,\nu) \\
&= \mathcal{I}_\pi(\mu,\nu).
\end{aligned}$$

⫲

Lastly, we will examine some alternative expressions for the statistical information that will prove useful in the following chapter to relate it to a particular f-divergence.

Proposition 4.2.13. *Let $\mu, \nu \in \mathcal{M}(\mathcal{X})$ be finite measures such that $\mu \ll \nu$, $\pi \in \mathcal{M}(\{0,1\})$ and $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ a Markov kernel. Then we have:*

$$i)\ \mathcal{I}_\pi(\mu,\nu) = \begin{cases} \nu\left[x \mapsto \max\{0, \pi_0 \dfrac{d\mu}{d\nu}(x) - \pi_1\}\right] & \text{if } \pi_0\mu(\mathcal{X}) \leq \pi_1\nu(\mathcal{X}), \\ \nu\left[x \mapsto \max\{0, \pi_1 - \pi_0 \dfrac{d\mu}{d\nu}(x)\}\right] & \text{if } \pi_0\mu(\mathcal{X}) \geq \pi_1\nu(\mathcal{X}), \end{cases}$$

ii) $\mathcal{I}_\pi(\mu,\nu) = -\frac{1}{2}|\pi_0\mu(\mathcal{X}) - \pi_1\nu(\mathcal{X})| + \frac{1}{2}\nu\left[x \mapsto \left|\pi_0\frac{d\mu}{d\nu}(x) - \pi_1\right|\right].$

Proof. First, let us observe that, since $\mu \ll \nu$, we have that $\mu, \nu \ll P \circ \pi = \pi_0\mu + \pi_1\nu$. We now consider the case where $\pi_0\mu(\mathcal{X}) \leq \pi_1\nu(\mathcal{X})$.

(i) We begin with the definition of the statistical information and use the properties of the max and min, along with those of the Radon-Nikodym derivative, to obtain the following:

$$
\begin{aligned}
\mathcal{I}_\pi(\mu,\nu) &= \min\{\pi_0\mu(\mathcal{X}), \pi_1\nu(\mathcal{X})\} - \mathcal{B}_\pi(\mu,\nu) \\
\text{By Equation (4.1)} \quad &= \pi_0\mu(\mathcal{X}) - (P\circ\pi)\left[x \mapsto \min\left\{\pi_0\frac{d\mu}{dP\circ\pi}(x), \pi_1\frac{d\nu}{dP\circ\pi}(x)\right\}\right] \\
\min\{a,b\} = a + \min\{0, b-a\} \quad &= \pi_0\mu(\mathcal{X}) - (P\circ\pi)\left[x \mapsto \pi_0\frac{d\mu}{dP\circ\pi}(x)\right] \\
&\quad -(P\circ\pi)\left[x \mapsto \min\left\{0, \pi_1\frac{d\nu}{dP\circ\pi}(x) - \pi_0\frac{d\mu}{dP\circ\pi}(x)\right\}\right] \\
\text{By Proposition B.0.1} \quad &= \pi_0\mu(\mathcal{X}) - \pi_0\mu(\mathcal{X}) \\
&\quad -(P\circ\pi)\left[x \mapsto \min\left\{0, \pi_1\frac{d\nu}{dP\circ\pi}(x) - \pi_0\frac{d\mu}{dP\circ\pi}(x)\right\}\right] \\
-\min\{a,b\} = \max\{-a,-b\} \quad &= (P\circ\pi)\left[x \mapsto \max\left\{0, \pi_0\frac{d\mu}{dP\circ\pi}(x) - \pi_1\frac{d\nu}{dP\circ\pi}(x)\right\}\right] \\
\text{By Proposition B.0.1} \quad &= (P\circ\pi)\left[x \mapsto \max\left\{0, \pi_0\frac{d\mu}{d\nu}(x)\frac{d\nu}{dP\circ\pi}(x) - \pi_1\frac{d\nu}{dP\circ\pi}(x)\right\}\right] \\
\text{By Proposition B.0.1} \quad &= \nu\left[x \mapsto \max\left\{0, \pi_0\frac{d\mu}{d\nu}(x) - \pi_1\right\}\right].
\end{aligned}
$$

(ii) We proceed with the calculations from the previous case, employing the fact that the max can be expressed using the absolute value in the following way: $\max\{a,b\} = \frac{1}{2}(a+b+|a-b|)$.

$$
\begin{aligned}
\mathcal{I}_\pi(\mu,\nu) &= \nu\left[x \mapsto \max\left\{0, \pi_0\frac{d\mu}{d\nu}(x) - \pi_1\right\}\right] \\
\text{By the formula for the max} \quad &= \frac{1}{2}\nu\left[x \mapsto \pi_0\frac{d\mu}{d\nu}(x) - \pi_1 + \left|\pi_0\frac{d\mu}{d\nu}(x) - \pi_1\right|\right] \\
\text{By Proposition B.0.1} \quad &= \frac{1}{2}\left(\pi_0\mu(\mathcal{X}) - \pi_1\nu(\mathcal{X}) + \nu\left[x \mapsto \left|\pi_0\frac{d\mu}{d\nu}(x) - \pi_1\right|\right]\right)
\end{aligned}
$$

$$\text{Since } \pi_0\mu(\mathcal{X}) \leq \pi_1\nu(\mathcal{X}) \qquad = -\frac{1}{2}|\pi_0\mu(\mathcal{X}) - \pi_1\nu(\mathcal{X})| + \frac{1}{2}\nu\left[x \mapsto \left|\pi_0\frac{d\mu}{d\nu}(x) - \pi_1\right|\right].$$

In the case where $\pi_0\mu(\mathcal{X}) \geq \pi_1\nu(\mathcal{X})$ both equations can be derived with similar computations. ⁂

Chapter 5

Data Processing Inequality

The *Data Processing Inequality* (DPI) is a fundamental result in the field of information theory. It summarizes a pivotal idea about information: it is not possible to generate new information by processing existing data. In other words, raw data contain the most information, and the act of processing it can only reduce the amount of information. In particular, the version of the DPI that we are going to consider is a statement about f-divergences and says that the f-divergence between two measures cannot increase when both measures are composed with the same Markov kernel, in formulas:

$$D_f(\kappa \circ \mu, \kappa \circ \nu) \leq D_f(\mu, \nu).$$

If we interpret the measures as sources of data, then we can see the Markov kernel as a channel that processes the data and the DPI tells us that this channel cannot make it easier for us to distinguish between the two sources. For further reading, see [PW24].

This chapter presents three different proofs of the DPI, resulting in three versions of the inequality with slightly differing hypotheses. The initial version is the most classical one, it applies only to deterministic kernels, that is, measurable functions, and it does not require any assumptions regarding the spaces. The second proof generalizes the result to Markov kernels; however, it relies on some manipulations of the Radon-Nikodym derivative that require relatively strong assumptions on the measurable spaces, in particular we

need them to be standard Borel spaces. The final proof takes a different approach, inspired by [LV06; Lie12]. It is based on a representation of the f-divergences as the integral of the f-divergence of a particular parametric function; the f-divergence of this function can in turn be seen as the statistical information plus a term that only depends on the total masses of the measures. By employing this representation, the DPI for general f-divergences can be derived as a consequence of the DPI for the statistical information, which is a very natural result implied by the definition of the Bayesian risk. The third proof is the most general, as it applies to any Markov kernel and does not require any assumptions regarding the spaces. It is surprising that an apparently unrelated concept, such as the hypothesis testing framework, which gives rise to a very specific divergence, can be used to prove such a general result about f-divergences.

We begin with a remark that will be used to simplify some of the subsequent proofs of the DPI.

Remark 5.0.1. In order to prove the DPI for finite measures and a fixed Markov kernel $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$, it is sufficient to prove it for the case where the first measure is absolutely continuous with respect to the second.

Proof. Let $\mu, \nu \in \mathcal{M}(\mathcal{X})$ be finite measures and let us assume that for every pair of finite measures $\mu', \nu' \in \mathcal{M}(\mathcal{X})$ such that $\mu' \ll \nu'$ we have $D_f(\kappa \circ \mu', \kappa \circ \nu') \leq D_f(\mu', \nu')$.

We can use the Lebesgue decomposition (see Theorem 2.2.3) to write $\mu = \dfrac{d\mu}{d\nu} \cdot \nu + \mu_{\perp\nu}$, and hence $\kappa \circ \mu = \kappa \circ \left(\dfrac{d\mu}{d\nu} \cdot \nu\right) + \kappa \circ \mu_{\perp\nu}$.

Therefore, we have

$$
\begin{aligned}
D_f(\kappa \circ \mu, \kappa \circ \nu) &= D_f\left(\kappa \circ \left(\frac{d\mu}{d\nu} \cdot \nu\right) + \kappa \circ \mu_{\perp\nu}, \kappa \circ \nu\right) \\
&\leq D_f\left(\kappa \circ \left(\frac{d\mu}{d\nu} \cdot \nu\right), \kappa \circ \nu\right) + (\kappa \circ \mu_{\perp\nu})(\mathcal{Y}) f'(\infty) && \text{By Proposition E.0.2} \\
&\leq D_f\left(\left(\frac{d\mu}{d\nu} \cdot \nu\right), \nu\right) + (\kappa \circ \mu_{\perp\nu})(\mathcal{Y}) f'(\infty) && \text{Since } \kappa \circ \left(\tfrac{d\mu}{d\nu} \cdot \nu\right) \ll \kappa \circ \nu \\
&= D_f\left(\left(\frac{d\mu}{d\nu} \cdot \nu\right), \nu\right) + \mu_{\perp\nu}(\mathcal{X}) f'(\infty) && \text{By Proposition 2.1.19}
\end{aligned}
$$

By Proposition E.0.2 $= D_f(\mu, \nu)$.

⅄

5.1 DPI for measurable functions

In order to demonstrate the DPI for measurable functions, we will go through a proof of the DPI for the restriction to sub σ-algebras, which can be regarded as a consequence of Jensen's inequality for the conditional expectation. Then we will show that the f-divergence of two measures composed with the same function is equal to the f-divergence of the restrictions of the measures to an appropriate sub σ-algebra.

First, let us establish some results regarding the Radon-Nikodym derivative of restricted measures.

Lemma 5.1.1. *Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ such that $\mu \ll \nu$, $\mathcal{A} \subseteq \mathscr{F}_\mathcal{X}$ a sub-σ-algebra and $f\colon \mathcal{X} \to \mathcal{Y}$ a measurable function. Then:*

i) $\dfrac{d\mu_{|\mathcal{A}}}{d\nu_{|\mathcal{A}}} = \nu\left[\frac{d\mu}{d\nu} \mid \mathcal{A}\right]$ *almost everywhere with respect to ν,*

ii) $\dfrac{d(g \circ \mu)}{d(g \circ \nu)}(g(x)) = \dfrac{d\mu_{|g^*\mathcal{Y}}}{d\nu_{|g^*\mathcal{Y}}}(x)$ *for almost every $x \in \mathcal{X}$ with respect to $\nu_{|g^*\mathcal{Y}}$ (therefore also with respect to ν),*

where $\nu[\cdot \mid \mathcal{A}]$ is the conditional ν-expectation with respect to $\mathcal{A}$ and $g^*\mathcal{Y}$ is the comap of $\mathscr{F}_\mathcal{Y}$ under g (i.e. the σ-algebra generated by the sets of the form $g^{-1}(B)$ for $B \in \mathscr{F}_\mathcal{Y}$)

Proof. First, it can be observed that, since $\mu \ll \nu$, we have $\mu_{|\mathcal{A}} \ll \nu_{|\mathcal{A}}$, $\mu_{|g^*\mathcal{Y}} \ll \nu_{|g^*\mathcal{Y}}$ and $g \circ \mu \ll g \circ \nu$. Moreover, for every sub σ-algebra $\mathcal{G} \subseteq \mathscr{F}_\mathcal{Y}$, every $\mathcal{G}$-measurable function f and every $G \in \mathcal{G}$ the following equality holds:

$$\int_G f d\nu_{|\mathcal{G}} = \int_G f\nu.$$

(i) By the properties of the conditional expectation, it is enough to show that for every $A \in \mathcal{A}$ we have

$$\int_A \frac{d\mu_{|\mathcal{A}}}{d\nu_{|\mathcal{A}}} d\nu \overset{?}{=} \int_A \frac{d\mu}{d\nu} d\nu.$$

But we can easily see that this is indeed the case, since $\dfrac{d\mu_{|\mathcal{A}}}{d\nu_{|\mathcal{A}}}$ is $\mathcal{A}$-measurable, $A \in \mathcal{A}$ and using Proposition B.0.1:

$$\int_A \frac{d\mu_{|\mathcal{A}}}{d\nu_{|\mathcal{A}}} d\nu = \int_A \frac{d\mu_{|\mathcal{A}}}{d\nu_{|\mathcal{A}}} d\nu_{|\mathcal{A}} = \mu_{|\mathcal{A}}(A) = \mu(A) = \int_A \frac{d\mu}{d\nu} d\nu.$$

(ii) It is enough to show that the integrals of the two functions over every $g^*\mathcal{Y}$-measurable set coincide, but $g^*\mathcal{Y}$-measurable sets are of the form $g^{-1}(B)$ for some $B \in \mathscr{F}_{\mathcal{Y}}$. Let $B \in \mathscr{F}_{\mathcal{Y}}$, then by Proposition B.0.1 we have

$$\begin{aligned}
\int_{g^{-1}(B)} \frac{d\mu_{|g^*\mathcal{Y}}}{d\nu_{|g^*\mathcal{Y}}}(x)\, \mathrm{d}(\nu_{|g^*\mathcal{Y}})(x) &= \int_{g^{-1}(B)} 1\, \mathrm{d}(\mu_{|g^*\mathcal{Y}})(x) \\
&= \int_{g^{-1}(B)} 1\, \mathrm{d}\mu(x) \\
&= \int_B 1\, \mathrm{d}(g \circ \mu)(y) \\
\text{By Proposition B.0.1} \qquad &= \int_B \frac{d(g \circ \mu)}{d(g \circ \nu)}(y)\, \mathrm{d}(g \circ \nu)(y) \\
&= \int_{g^{-1}(B)} \frac{d(g \circ \mu)}{d(g \circ \nu)}(g(x))\, \mathrm{d}\nu(x) \\
&= \int_{g^{-1}(B)} \frac{d(g \circ \mu)}{d(g \circ \nu)}(g(x))\, \mathrm{d}(\nu_{|g^*\mathcal{Y}})(x).
\end{aligned}$$

⋇

Theorem 5.1.2. *Let $(\mathcal{X}, \mathscr{F}_{\mathcal{X}})$ be a measurable space, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ finite measures, $\mathcal{A} \subseteq \mathscr{F}_{\mathcal{X}}$ a sub-σ-algebra and $f\colon \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}$ a convex function. Then*

$$D_f(\mu_{|\mathcal{A}}, \nu_{|\mathcal{A}}) \leq D_f(\mu, \nu).$$

Proof. Since $\left(\frac{d\mu}{d\nu} \cdot \nu\right)_{|\mathcal{A}} \ll \nu_{|\mathcal{A}}$ and $(\mu_{\perp\nu})_{|\mathcal{A}}(\mathcal{X}) = \mu_{\perp\nu}(\mathcal{X})$, we can use the same reasoning as in Remark 5.0.1 to assume without loss of generality that $\mu \ll \nu$, hence $\mu_{|\mathcal{A}} \ll \nu_{|\mathcal{A}}$.

Thus, the proof becomes a computation using Jensen's inequality:

$$\begin{aligned}
D_f(\mu_{|\mathcal{A}}, \nu_{|\mathcal{A}}) &= \int_{\mathcal{X}} f\left(\frac{d\mu_{|\mathcal{A}}}{d\nu_{|\mathcal{A}}}\right) \mathrm{d}\nu_{|\mathcal{A}} \\
&= \int_{\mathcal{X}} f\left(\frac{d\mu_{|\mathcal{A}}}{d\nu_{|\mathcal{A}}}\right) \mathrm{d}\nu \\
\text{By Lemma 5.1.1} \qquad &= \int_{\mathcal{X}} f\left(\nu\left[\frac{d\mu}{d\nu} \mid \mathcal{A}\right]\right) \mathrm{d}\nu \\
\text{By Jensen's inequality}^1 \quad &\leq \int_{\mathcal{X}} \nu\left[f\left(\frac{d\mu}{d\nu}\right) \mid \mathcal{A}\right] \mathrm{d}\nu \\
\text{Property of cond. exp.} \quad &= \int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\nu \\
&= D_f(\mu, \nu).
\end{aligned}$$

⋕

We are now ready to prove the first version of the DPI.

Theorem 5.1.3 (DPI, measurable functions)**.** *Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ finite measures, $g\colon \mathcal{X} \to \mathcal{Y}$ a measurable function and $f\colon \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}$ a convex function. Then*

$$D_f(g \circ \mu, g \circ \nu) \leq D_f(\mu, \nu).$$

Proof. By Remark 5.0.1 we can assume without loss of generality that $\mu \ll \nu$, so $\mu_{|g^*\mathcal{Y}} \ll \nu_{|g^*\mathcal{Y}}$. Now we can conclude the proof using the DPI for the restriction to sub σ-algebras:

$$\begin{aligned}
D_f(g \circ \mu, g \circ \nu) &= \int_{\mathcal{Y}} f\left(\frac{d(g \circ \mu)}{d(g \circ \nu)}\right)(y)\, \mathrm{d}(g \circ \nu)(y) \\
&= \int_{\mathcal{X}} f\left(\frac{d(g \circ \mu)}{d(g \circ \nu)}\right)(g(x))\, \mathrm{d}\nu(x) \\
\text{By Lemma 5.1.1} \qquad &= \int_{\mathcal{X}} f\left(\frac{d\mu_{|g^*\mathcal{Y}}}{d\nu_{|g^*\mathcal{Y}}}\right)(x)\, \mathrm{d}\nu(x) \\
&= \int_{\mathcal{X}} f\left(\frac{d\mu_{|g^*\mathcal{Y}}}{d\nu_{|g^*\mathcal{Y}}}\right)(x)\, \mathrm{d}\nu_{|g^*\mathcal{Y}}(x)
\end{aligned}$$

[1] For a proof of the conditional Jensen's inequality see [Hyt+16, Proposition 2.6.29].

$$= D_f(\mu_{|g^*\mathcal{Y}}, \nu_{|g^*\mathcal{Y}})$$

By Theorem 5.1.2 $$\leq D_f(\mu, \nu).$$

⫽

Remark 5.1.4. Currently, in Lean, we have `ProbabilityTheory.fDiv_trim_le`, which is a formalized version of the DPI for σ-algebras (Theorem 5.1.2), but the proof still has a small gap, since Jensen's inequality for the conditional expectation is not yet in Mathlib. Moreover, we still lack a formalization of the DPI for measurable functions (Theorem 5.1.3). However, this proof is unlikely to be worked on in the near future, as a stronger version of the DPI has already been developed (see following sections).

5.2 DPI for kernels in standard Borel spaces

In this section, we generalize the proof of the DPI to Markov kernels. The proof will proceed by establishing another inequality for the f-divergence, that features the composition product instead of the composition:

$$D_f(\mu, \nu) \leq D_f(\mu \otimes \kappa, \nu \otimes \eta).$$

Then we will use the equation defining the Bayesian inverse

$$((\kappa \circ \mu) \otimes \kappa^\dagger_\mu)_\leftrightarrow = \mu \otimes \kappa$$

to make the composition appear, and finally we will use the fact that if we perform the composition product with the same kernel on both measures, the f-divergence remains the same:

$$D_f(\mu \otimes \kappa, \nu \otimes \kappa) = D_f(\mu, \nu).$$

The limit of this proof is that it does not work for general measurable spaces. In particular, we need the Bayesian inverse of the kernel with respect to the measures to exist, which is not true in general. However, as previously noted in Remark 2.1.22, it is sufficient that the first space is standard Borel.

Moreover, in order to establish the first inequality, it is necessary to work with the Radon-Nikodym derivative of the kernels, and for it to be well-behaved the second space needs[2] to be countably generated (or, alternatively, the first space needs to be countable), as seen in Theorem 2.2.7 and Proposition B.0.2.

Proposition 5.2.1. *Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces such that $\mathcal{X}$ is countable or $\mathcal{Y}$ is countably generated, let $\mu, \nu \in \mathcal{M}(\mathcal{X})$ be finite measures, $\kappa, \eta \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ Markov kernels and $f \colon \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}$ a convex function. Then*

$$D_f(\mu, \nu) \leq D_f(\mu \otimes \kappa, \nu \otimes \eta).$$

Proof. This is the definition of the f-divergence for the composition products:

$$D_f(\mu\otimes\kappa, \nu\otimes\eta) = \int_{\mathcal{X}\times\mathcal{Y}} f\left(\frac{d(\mu \otimes \kappa)}{d(\nu \otimes \eta)}\right) \mathrm{d}(\nu\otimes\eta) + (\mu\otimes\kappa)_{\perp(\nu\otimes\eta)}(\mathcal{X}\times\mathcal{Y}) f'(\infty).$$

We will handle the two terms separately.

Let us begin with the absolutely continuous part. First, notice that a property holds for almost every $(x, y) \in \mathcal{X} \times \mathcal{Y}$ with respect to $\nu \otimes \eta$ if and only if for ν-a.e. x the property holds for $\eta(x)$-a.e. y. Hence, using Proposition B.0.2, we can establish that for almost every (x, y) with respect to $\nu \otimes \eta$ we have

$$\frac{d(\mu \otimes \kappa)}{d(\nu \otimes \eta)}(x, y) = \frac{d\mu}{d\nu}(x)\frac{d\kappa}{d\eta}(x, y) = \frac{d\mu}{d\nu}(x)\frac{d\kappa(x)}{d\eta(x)}(y).$$

Therefore, we can write

$$\begin{aligned}
\int_{\mathcal{X}\times\mathcal{Y}} f\left(\frac{d(\mu \otimes \kappa)}{d(\nu \otimes \eta)}\right)(\nu \otimes \eta) &= \int_{\mathcal{X}\times\mathcal{Y}} f\left(\frac{d\mu}{d\nu}(x)\frac{d\kappa(x)}{d\eta(x)}(y)\right)(\nu \otimes \eta)(dx, dy) \\
\text{By Fubini's theorem} \qquad &= \int_{\mathcal{X}}\int_{\mathcal{Y}} f\left(\frac{d\mu}{d\nu}(x)\frac{d\kappa(x)}{d\eta(x)}(y)\right)\eta(x, dy)\,\mathrm{d}\nu(x) \\
\text{By Jensen's inequality}^3 \qquad &\geq \int_{\mathcal{X}} f\left(\int_{\mathcal{Y}} \frac{d\mu}{d\nu}(x)\frac{d\kappa(x)}{d\eta(x)}(y)\eta(x, dy)\right)\mathrm{d}\nu(x)
\end{aligned}$$

[2]This is true at least for the proof that we currently have in Lean; however, it is possible that this assumption can be removed with some additional work.

[3]We can use Jensen's inequality here because f is convex and η is a Markov kernel, so $\eta(y)$ is a probability measure.

$$= \int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}(x)\left(\frac{d\kappa(x)}{d\eta(x)}\cdot\eta\right)(x,\mathcal{Y})\right)\mathrm{d}\nu(x). \tag{5.1}$$

Regarding the singular part, we can use Proposition B.0.2 and Proposition 2.1.19 to write

$$\begin{aligned}(\mu\otimes\kappa)_{\perp(\nu\otimes\eta)}(\mathcal{X}\times\mathcal{Y}) &= (\mu_{\perp\nu}\otimes\kappa)(\mathcal{X}\times\mathcal{Y}) + \left(\left(\frac{d\mu}{d\nu}\cdot\nu\right)\otimes\kappa_{\perp\eta}\right)(\mathcal{X}\times\mathcal{Y})\\ &= \mu_{\perp\nu}(\mathcal{X}) + \int_{\mathcal{X}}\kappa_{\perp\eta}(x,\mathcal{Y})\,\mathrm{d}\left(\frac{d\mu}{d\nu}\cdot\nu\right)(x)\\ &= \mu_{\perp\nu}(\mathcal{X}) + \int_{\mathcal{X}}\frac{d\mu}{d\nu}(x)\kappa_{\perp\eta}(x,\mathcal{Y})\,\mathrm{d}\nu(x). \end{aligned}\tag{5.2}$$

Moreover, since κ is a Markov kernel, we have, using the Lebesgue decomposition (Theorem 2.2.7), that

$$\frac{d\mu}{d\nu}(x) = \frac{d\mu}{d\nu}(x)\kappa(x,\mathcal{Y}) = \frac{d\mu}{d\nu}(x)\left(\frac{d\kappa(x)}{d\eta(x)}\cdot\eta\right)(x,\mathcal{Y}) + \frac{d\mu}{d\nu}(x)\kappa_{\perp\eta}(x,\mathcal{Y}),$$

so we can use Lemma D.0.3 to obtain

$$f\left(\frac{d\mu}{d\nu}(x)\right) \le f\left(\frac{d\mu}{d\nu}(x)\left(\frac{d\kappa(x)}{d\eta(x)}\cdot\eta\right)(x,\mathcal{Y})\right) + f'(\infty)\frac{d\mu}{d\nu}(x)\kappa_{\perp\eta}(x,\mathcal{Y}),$$

and integrating both sides with respect to ν we get

$$\begin{aligned}\int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}(x)\right)\mathrm{d}\nu(x) \le \int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}(x)\left(\frac{d\kappa(x)}{d\eta(x)}\cdot\eta\right)(x,\mathcal{Y})\right)\mathrm{d}\nu(x)\\ + f'(\infty)\int_{\mathcal{X}}\frac{d\mu}{d\nu}(x)\kappa_{\perp\eta}(x,\mathcal{Y})\,\mathrm{d}\nu(x).\end{aligned}\tag{5.3}$$

To conclude the proof, we can put together the previous results in the following way:

$$\begin{aligned} D_f(\mu,\nu) &= \int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}\right)\mathrm{d}\nu + \mu_{\perp\nu}(\mathcal{X})f'(\infty)\\ \text{By Equation (5.3)} &\le \int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}(x)\left(\frac{d\kappa(x)}{d\eta(x)}\cdot\eta\right)(x,\mathcal{Y})\right)\mathrm{d}\nu(x)\\ &\quad + f'(\infty)\int_{\mathcal{X}}\frac{d\mu}{d\nu}(x)\kappa_{\perp\eta}(x,\mathcal{Y})\,\mathrm{d}\nu(x) + \mu_{\perp\nu}(\mathcal{X})f'(\infty)\\ \text{By Equation (5.1)} &\le \int_{\mathcal{X}\times\mathcal{Y}} f\left(\frac{d(\mu\otimes\kappa)}{d(\nu\otimes\eta)}\right)\mathrm{d}(\nu\otimes\eta)\end{aligned}$$

$$
\begin{aligned}
&\quad + f'(\infty)\left(\int_{\mathcal{X}} \frac{d\mu}{d\nu}(x)\kappa_{\perp\eta}(x,\mathcal{Y})\,\mathrm{d}\nu(x) + \mu_{\perp\nu}(\mathcal{X})\right) \\
\text{By Equation (5.2)} \quad &= \int_{\mathcal{X}\times\mathcal{Y}} f\left(\frac{d(\mu\otimes\kappa)}{d(\nu\otimes\eta)}\right)\,\mathrm{d}(\nu\otimes\eta) + f'(\infty)(\mu\otimes\kappa)_{\perp(\nu\otimes\eta)}(\mathcal{X}\times\mathcal{Y}) \\
&= D_f(\mu\otimes\kappa, \nu\otimes\eta).
\end{aligned}
$$

Remark 5.2.2. We can notice that the part of the proof concerning the absolutely continuous part of the f-divergence is similar to the proof of the DPI for measurable functions. Indeed, both proofs are based on Jensen's inequality.

We will now examine a special case of the previous inequality, in which the two kernels coincide. In this instance the inequality becomes an equality.

Proposition 5.2.3. *Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces such that $\mathcal{X}$ is countable or $\mathcal{Y}$ is countably generated, let $\mu, \nu \in \mathcal{M}(\mathcal{X})$ be finite measures and $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ a Markov kernel. Then*

$$D_f(\mu\otimes\kappa, \nu\otimes\kappa) = D_f(\mu,\nu).$$

Proof. The proof is a simple computation:

$$
\begin{aligned}
D_f(\mu\otimes\kappa, \nu\otimes\kappa) &= \int_{\mathcal{X}\times\mathcal{Y}} f\left(\frac{d(\mu\otimes\kappa)}{d(\nu\otimes\kappa)}\right)\,\mathrm{d}(\nu\otimes\kappa) + f'(\infty)(\mu\otimes\kappa)_{\perp(\nu\otimes\kappa)}(\mathcal{X}\times\mathcal{Y}) \\
\text{By Proposition B.0.2} \quad &= \int_{\mathcal{X}\times\mathcal{Y}} f\left(\frac{d\mu}{d\nu}(x)\right)(\nu\otimes\kappa)(dx,dy) \\
&\quad + f'(\infty)\left((\mu_{\perp\nu}\otimes\kappa)(\mathcal{X}\times\mathcal{Y}) + \left(\frac{d\mu}{d\nu}\cdot\nu\right)\otimes\kappa_{\perp\kappa}(\mathcal{X}\times\mathcal{Y})\right) \\
\text{Since } \kappa_{\perp\kappa} = 0 \quad &= \int_{\mathcal{X}}\int_{\mathcal{Y}} f\left(\frac{d\mu}{d\nu}(x)\right)\kappa(x,dy)\,\mathrm{d}\nu(x) + f'(\infty)(\mu_{\perp\nu}\otimes\kappa)(\mathcal{X}\times\mathcal{Y}) \\
\text{By Proposition 2.1.19} \quad &= \int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}(x)\right)\kappa(x,\mathcal{Y})\,\mathrm{d}\nu(x) + f'(\infty)\mu_{\perp\nu}(\mathcal{X}) \\
&= \int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}(x)\right)\,\mathrm{d}\nu(x) + f'(\infty)\mu_{\perp\nu}(\mathcal{X}) \\
&= D_f(\mu,\nu).
\end{aligned}
$$

Lemma 5.2.4. *Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces and $\mu, \nu \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ finite measures. Then*

$$D_f(\mu_{\leftrightarrow}, \nu_{\leftrightarrow}) = D_f(\mu, \nu).$$

Proof. Let $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$ be measurable sets, then we can use the Lebesgue decomposition (Theorem 2.2.3) to write

$$\mu_{\leftrightarrow}(B \times A) = \frac{d\mu_{\leftrightarrow}}{d\nu_{\leftrightarrow}} \cdot \nu_{\leftrightarrow}(B \times A) + \mu_{\leftrightarrow \perp \nu_{\leftrightarrow}}(B \times A),$$

on the other hand we have

$$\begin{aligned}
\mu_{\leftrightarrow}(B \times A) &= \mu(A \times B) \\
&= \left(\frac{d\mu}{d\nu} \cdot \nu\right)(A \times B) + \mu_{\perp\nu}(A \times B) \\
&= \left(\frac{d\mu}{d\nu} \cdot \nu\right)_{\leftrightarrow}(B \times A) + (\mu_{\perp\nu})_{\leftrightarrow}(B \times A) \\
&= \left(\frac{d\mu}{d\nu}\right)_{\leftrightarrow} \cdot \nu_{\leftrightarrow}(B \times A) + (\mu_{\perp\nu})_{\leftrightarrow}(B \times A).
\end{aligned}$$

Therefore, the uniqueness of the Lebesgue decomposition (Theorem 2.2.3) implies that

$$\frac{d\mu_{\leftrightarrow}}{d\nu_{\leftrightarrow}} = \left(\frac{d\mu}{d\nu}\right)_{\leftrightarrow} \quad \text{and} \quad \mu_{\leftrightarrow \perp \nu_{\leftrightarrow}} = (\mu_{\perp\nu})_{\leftrightarrow}.$$

So we can conclude the proof by applying the definition of f-divergence:

$$\begin{aligned}
D_f(\mu_{\leftrightarrow}, \nu_{\leftrightarrow}) &= \int_{\mathcal{Y}\times\mathcal{X}} f\left(\frac{d\mu_{\leftrightarrow}}{d\nu_{\leftrightarrow}}\right) \mathrm{d}\nu_{\leftrightarrow} + \mu_{\leftrightarrow \perp \nu_{\leftrightarrow}}(\mathcal{Y} \times \mathcal{X}) f'(\infty) \\
&= \int_{\mathcal{Y}\times\mathcal{X}} f\left(\left(\frac{d\mu}{d\nu}\right)_{\leftrightarrow}\right) \mathrm{d}\nu_{\leftrightarrow} + (\mu_{\perp\nu})_{\leftrightarrow}(\mathcal{Y} \times \mathcal{X}) f'(\infty) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} f\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\nu + \mu_{\perp\nu}(\mathcal{X} \times \mathcal{Y}) f'(\infty) \\
&= D_f(\mu, \nu).
\end{aligned}$$

⋕

We are now ready to prove the DPI for Markov kernels in standard Borel spaces.

Theorem 5.2.5 (DPI, standard Borel spaces)**.** *Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces such that $\mathcal{X}$ is standard Borel and $\mathcal{Y}$ is countably generated or $\mathcal{X}$ is countable, let $\mu, \nu \in \mathcal{M}(\mathcal{X})$ be finite measures, $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ a Markov kernel and $f \colon \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}$ a convex function. Then*

$$D_f(\kappa \circ \mu, \kappa \circ \nu) \leq D_f(\mu, \nu).$$

Proof. First, let us remark that, under the current assumptions, the existence of the Bayesian inverses $\kappa_\mu^\dagger$ and $\kappa_\nu^\dagger$ is guaranteed by Remark 2.1.22.

Therefore, we can use Proposition 5.2.1 to obtain:

$$\begin{aligned}
D_f(\kappa \circ \mu, \kappa \circ \nu) &\leq D_f((\kappa \circ \mu) \otimes \kappa_\mu^\dagger, (\kappa \circ \nu) \otimes \kappa_\nu^\dagger) \\
\text{By Lemma 5.2.4} \qquad &= D_f(((\kappa \circ \mu) \otimes \kappa_\mu^\dagger)_\leftrightarrow, ((\kappa \circ \nu) \otimes \kappa_\nu^\dagger)_\leftrightarrow) \\
\text{By Equation (2.3)} \qquad &= D_f(\mu \otimes \kappa, \nu \otimes \kappa) \\
\text{By Proposition 5.2.3} \qquad &= D_f(\mu, \nu).
\end{aligned}$$

∎

Remark 5.2.6. The following is the formalization of this statement of the DPI:

```
lemma fDiv_comp_right_le [Nonempty α] [StandardBorelSpace α]
  [CountableOrCountablyGenerated α β]
  (μ ν : Measure α) [IsFiniteMeasure μ] [IsFiniteMeasure ν]
  (κ : Kernel α β) [IsMarkovKernel κ]
  (hf : StronglyMeasurable f) (hf_cvx : ConvexOn ℝ (Ici 0) f)
  (hf_cont : ContinuousOn f (Ici 0)) :
  fDiv f (κ ∘ₘ μ) (κ ∘ₘ ν) ≤ fDiv f μ ν := by
```

We can notice how in the hypotheses we require the function f to be convex only on the nonnegative real numbers, which is sufficient given that in the definition of f-divergence, f is only applied to nonnegative values. We also require f to be continuous on the nonnegative real numbers and strongly measurable on the entire real line. These hypotheses are needed for technical reasons in the formalization, and they are not implied by the convexity. See also how the hypothesis of the first space being countable or the second one being countably

generated is bundled in the class `CountableOrCountablyGenerated` α β and automatically inferred by the typeclass inference system.

Lastly, we want to remark that the formalized proof is slightly different from the one presented here. In particular, it does not mention the Bayesian inverse explicitly, but instead makes the composition appear from the composition product by taking the second marginal: $\kappa \circ \mu = (\mu \otimes \kappa)_{\mathcal{Y}}$. Then it uses the disintegration to show that the f-divergence of the marginals is always less than the f-divergence of the original measures[4].

This approach is essentially equivalent to the one presented here and requires the same assumptions on the spaces. In fact, as mentioned in Remark 2.1.22, the Bayesian inverse can be seen as a special case of disintegration.

5.3 DPI in general spaces

In this section we will show how the DPI can be extended to general measurable spaces. This proof will adopt a completely different approach than the previous ones, and it is inspired by [LV06; Lie12]. The idea is to use a Taylor formula for convex functions where the error is expressed as the integral of a parametric function $\varphi_{\beta,\gamma}$; using this formula, we will be able to express the f-divergence for a generic convex function in terms of the integral of the f-divergence of $\varphi_{\beta,\gamma}$. Then, we will show how $D_{\varphi_{\beta,\gamma}}$ is closely related to the statistical information, in particular being equal to the statistical information with the addition of a term that is invariant under composition

[4]The idea is that, as mentioned in Remark 2.1.22, $\mu = \mu_{\mathcal{X}} \otimes \mu_{\mathcal{Y}|\mathcal{X}}$. Therefore, using Proposition 5.2.1 we obtain:

$$D_f(\mu, \nu) = D_f(\mu_{\mathcal{X}} \otimes \mu_{\mathcal{Y}|\mathcal{X}}, \nu_{\mathcal{X}} \otimes \nu_{\mathcal{Y}|\mathcal{X}}) \geq D_f(\mu_{\mathcal{X}}, \nu_{\mathcal{X}}).$$

For the second marginal it is enough to use the swap kernel and Lemma 5.2.4:

$$D_f(\mu_{\mathcal{Y}}, \nu_{\mathcal{Y}}) = D_f((\mu_{\leftrightarrow})_{\mathcal{X}}, (\nu_{\leftrightarrow})_{\mathcal{X}}) \leq D_f(\mu_{\leftrightarrow}, \nu_{\leftrightarrow}) = D_f(\mu, \nu).$$

with a constant kernel. This will allow us to use the DPI for the statistical information, which we independently established in Proposition 4.2.12 as a simple consequence of the definition of Bayesian risk, to prove the DPI for $D_{\varphi_{\beta,\gamma}}$ and then for general f-divergences.

We begin by defining the parametric function $\varphi_{\beta,\gamma}$, which will be a piecewise linear function and will be called the hockey-stick function due to its shape.

Definition 5.3.1 (Hockey-stick function)**.** Let $\beta, \gamma \in \mathbb{R}$ be real numbers. We define the *hockey-stick function* $\varphi_{\beta,\gamma} \colon \mathbb{R} \to \mathbb{R}$ as

$$\varphi_{\beta,\gamma}(x) \coloneqq \begin{cases} \max\{0, \gamma - \beta x\} & \text{if } \gamma \leq \beta, \\ \max\{0, \beta x - \gamma\} & \text{if } \gamma > \beta. \end{cases}$$

Remark 5.3.2. The hokey-stick function is continuous and convex on $\mathbb{R}$, moreover it is a piecewise linear function, being zero on one side of γ/β and linear with slope β or $-\beta$ on the other side, depending on whether γ is greater or smaller than β.

This function is well-defined for any real value of β and γ; however, we will primarily consider nonnegative values. This is because our goal is to establish a relation between its f-divergence and the statistical information $\mathcal{I}_{(\beta,\gamma)}$, for which negative values of β and γ do not make sense, since the two parameters define a measure on $\{0, 1\}$. Moreover, in the event that precisely one of the parameters is negative, it is easy to show that $\varphi_{\beta,\gamma}(x) = 0$ for all $x > 0$, rendering its f-divergence trivially zero. We will be especially interested in the case where $\beta = 1$, therefore also γ will be nonnegative.

The following proposition gives us the Taylor formula for convex functions.

Theorem 5.3.3 (Taylor formula for convex functions)**.** *Let $f \colon \mathbb{R} \to \mathbb{R}$ be a convex function and γ_f its curvature measure (see Definition D.0.5). Then for any $a, b \in \mathbb{R}$ we have*

$$f(b) - f(a) - f'(a)(b - a) = \begin{cases} \int_{(a,b]} (b - x) \, \mathrm{d}\gamma_f(x) & \text{if } a \leq b, \\ \int_{(b,a]} (x - b) \, \mathrm{d}\gamma_f(x) & \text{if } a > b. \end{cases}$$

Proof. We will prove the case $a > b$, the other case is analogous.

Let us give a name to the function inside the integral: $g(x) \coloneqq (x - b)$. Recall that the curvature measure is defined as the Lebesgue-Stieltjes measure associated with the right derivative. We can also consider the Lebesgue-Stieltjes measure associated with g, which is just the classical Lebesgue measure, in fact:

$$\Lambda_g((x, y]) = g(y) - g(x) = y - b - (x - b) = y - x.$$

Now we can use integration by parts (Corollary 2.3.6) to change the measure of the integral, making it a classical Lebesgue integral:

$$\begin{aligned}
\int_{(b,a]} (x - b)\,\mathrm{d}\gamma_f(x) &= \int_{(b,a]} g(x)\,\mathrm{d}\gamma_f(x) \\
\text{By Corollary 2.3.6} \qquad &= f'_+(a)g(a) - f'_+(b)g(b) - \int_{(b,a]} f'_+(x)\,\mathrm{d}\Lambda_g(x) \\
&= f'_+(a)(a - b) - \int_b^a f'_+(x)\,\mathrm{d}x \\
\text{Fundamental theorem of calculus}^5 \qquad &= f'_+(a)(a - b) - (f(a) - f(b)).
\end{aligned}$$

⫽

The following corollary demonstrates how the integral in the Taylor formula for $a = 1$ can be interpreted as an integral of the hockey-stick function, and how it simplifies when we use a centered function, i.e. f such that $f(1) = 0$ and $f'(1) = 0$.

Corollary 5.3.4. *Let $f\colon \mathbb{R} \to \mathbb{R}$ be a convex function and γ_f its curvature measure. Then for every $t \in \mathbb{R}$ we have*

$$f(t) - f(1) - f'(1)(t - 1) = \int_{\mathbb{R}} \varphi_{1,y}(t)\,\mathrm{d}\gamma_f(y).$$

[5]For this version of the fundamental theorem of calculus with the right derivative see the lemma `intervalIntegral.integral_eq_sub_of_hasDeriv_right` in Mathlib.

Moreover, if $f(1) = f'(1) = 0$*, then*

$$f(t) = \int_{\mathbb{R}} \varphi_{1,y}(t) \, \mathrm{d}\gamma_f(y).$$

Proof. It is sufficient to observe that if $1 \leq t$, then for every $y \leq 1$, we have $\varphi_{1,y}(t) = \max\{0, y - t\} = 0$, for $1 < y \leq t$ we have $\varphi_{1,y}(t) = \max\{0, t - y\} = t - y$ and for $t < y$ we have $\varphi_{1,y}(t) = \max\{0, t - y\} = 0$. Therefore, the integral coincides with the integral in the Taylor formula:

$$\int_{\mathbb{R}} \varphi_{1,y}(t) \, \mathrm{d}\gamma_f(y) = \int_{(1,t]} (t - y) \, \mathrm{d}\gamma_f(y).$$

Similarly, if $t < 1$ we have

$$\int_{\mathbb{R}} \varphi_{1,y}(t) \, \mathrm{d}\gamma_f(y) = \int_{(t,1]} (y - t) \, \mathrm{d}\gamma_f(y).$$

⁂

We will now use this corollary to write the f-divergence of a convex function as an integral of the f-divergence of $\varphi_{1,y}$. We will restrict our attention to the case where one measure is absolutely continuous with respect to the other and $f(1) = f'(1) = 0$, since this is sufficient to prove the DPI, thanks to Remark 5.0.1 and Proposition 3.1.5. Nevertheless, this representation also holds in the general case with minor modifications, see also Remark 5.3.6 for more details.

Proposition 5.3.5 (Integral representation of f-divergence)**.** *Let* $f \colon \mathbb{R} \to \mathbb{R}$ *be a convex function such that* $f(1) = f'(1) = 0$ *and* $\mu, \nu \in \mathcal{M}(\mathcal{X})$ *be finite measures such that* $\mu \ll \nu$*. Then*

$$D_f(\mu, \nu) = \int_{\mathbb{R}} D_{\varphi_{1,y}}(\mu, \nu) \, \mathrm{d}\gamma_f(y).$$

Proof. Since $\mu \ll \nu$, we only have to deal with the integral part of the f-divergence. Therefore, the proof is essentially a matter of swapping the order of integration:

$$\begin{aligned}
D_f(\mu,\nu) &= \int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}(x)\right) \mathrm{d}\nu(x) \\
\text{By Corollary 5.3.4} \quad &= \int_{\mathcal{X}} \int_{\mathbb{R}} \varphi_{1,y}\left(\frac{d\mu}{d\nu}(x)\right) \mathrm{d}\gamma_f(y)\, \mathrm{d}\nu(x) \\
\text{By Tonelli's theorem}^6 \quad &= \int_{\mathbb{R}} \int_{\mathcal{X}} \varphi_{1,y}\left(\frac{d\mu}{d\nu}(x)\right) \mathrm{d}\nu(x)\, \mathrm{d}\gamma_f(y) \\
&= \int_{\mathbb{R}} D_{\varphi_{1,y}}(\mu,\nu)\, \mathrm{d}\gamma_f(y).
\end{aligned}$$

Remark 5.3.6. This proof serves as an illustrative example of how the formalization of a proof can become considerably longer and more challenging than the informal proof. Indeed, the formalized result has been split into several lemmas, for a total of over 300 lines of code, culminating in the result for the general case `ProbabilityTheory.fDiv_eq_lintegral_fDiv_statInfoFun`. Part of the reason for this is that, in the formalization, we decided to prove this result for general measures and convex functions, but the absolutely continuous and the mutually singular case were still written as separate lemmas, as they were necessary to prove the general case. Some of these intermediate lemmas were also proven for $f(1) = f'(1) = 0$ and subsequently generalized. The decision to prove the result for general measures and functions, rather than postponing the generalization to the proof of the DPI, was made to facilitate its use in other contexts, independently of the DPI.

However, this was not the only reason that made this proof challenging, in fact another tricky aspect was the swapping of the integrals. We usually have 2 tools at our disposal to swap integrals: Fubini's theorem and Tonelli's theorem. In this case, since the integrand is nonnegative, the most appropriate choice would be Tonelli's theorem, as it does not require any integrability condition. However, the version of Tonelli's theorem available in Mathlib only works for the lower Lebesgue integral. In contrast, the definition of f-divergence and

[6] We can use Tonelli's theorem here because the hockey-stick function is nonnegative.

the Taylor formula are based on the Bochner integral, for which only Fubini's theorem is available. At this point, two main options were viable: either stick to the Bochner integral and try to address the integrability condition required by Fubini's theorem, or switch to the lower Lebesgue integral and use Tonelli's theorem. In reality, both options needed some work on the integrability of a function, since also the lemma stating the equality of the Bochner and the lower Lebesgue integral requires them. We decided to proceed with the second option, given that working with the lower Lebesgue integral typically presents fewer challenges. For example, it greatly simplified the proof of the DPI, which leverages the monotonicity of the integral. This property is not true for the Bochner integral, unless both functions are integrable[7], whereas it is always true for the lower Lebesgue integral.

The following result uses one of the formulas for the statistical information from the previous chapter to show how it is related to the f-divergence of the hockey-stick function.

Proposition 5.3.7. *Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ finite measures such that $\mu \ll \nu$ and $\beta, \gamma \geq 0$. Then*

$$D_{\varphi_{\beta,\gamma}}(\mu,\nu) = \mathcal{I}_{(\beta,\gamma)}(\mu,\nu) + \frac{1}{2}\left|\beta\mu(\mathcal{X}) - \gamma\nu(\mathcal{X})\right| - \frac{1}{2}\operatorname{sgn}(\beta-\gamma)\left(\beta\mu(\mathcal{X}) - \gamma\nu(\mathcal{X})\right),$$

where $\operatorname{sgn}(t) = 1$ *if* $t \geq 0$ *and* $\operatorname{sgn}(t) = -1$ *if* $t < 0$.

Proof. We initially consider the case where $\gamma \leq \beta$.

We will use the fact that the max can be expressed using the absolute value in the following way: $\max\{a, b\} = \frac{1}{2}(a + b + |a - b|)$.

$$\begin{aligned} D_{\varphi_{\beta,\gamma}}(\mu,\nu) &= \int_{\mathcal{X}} \varphi_{\beta,\gamma}\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\nu \\ &= \int_{\mathcal{X}} \max\left\{0, \gamma - \beta\frac{d\mu}{d\nu}\right\} \mathrm{d}\nu \end{aligned}$$

[7]This is due to the way the Bochner integral handles the non-integrable functions, throwing a junk value of 0.

$$
\begin{aligned}
&\text{By the formula for the max} && = \int_{\mathcal{X}} \frac{1}{2}\left(\gamma - \beta\frac{d\mu}{d\nu} + \left|\gamma - \beta\frac{d\mu}{d\nu}\right|\right) \mathrm{d}\nu \\
&\text{By Proposition B.0.1} && = \frac{1}{2}\left(\gamma\nu(\mathcal{X}) - \beta\mu(\mathcal{X})\right) + \frac{1}{2}\int_{\mathcal{X}} \left|\gamma - \beta\frac{d\mu}{d\nu}\right| \mathrm{d}\nu \\
&\text{By Proposition 4.2.13} && = -\frac{1}{2}\left(\beta\mu(\mathcal{X}) - \gamma\nu(\mathcal{X})\right) + \mathcal{I}_{(\beta,\gamma)}(\mu,\nu) + \frac{1}{2}|\beta\mu(\mathcal{X}) - \gamma\nu(\mathcal{X})|.
\end{aligned}
$$

The case where $\gamma > \beta$ is analogous. ⁂

Remark 5.3.8. In the event that the measures have the same total mass, i.e. $\mu(\mathcal{X}) = \nu(\mathcal{X})$, we can see from Proposition 5.3.7 that the statistical information and the f-divergence of the hockey-stick coincide. This is in particular true when μ and ν are probability measures.

An alternative perspective on this phenomenon can be gained by examining the representation of both divergences in terms of an integral of a max function. In fact, by the definition of f-divergence we have:

$$
D_{\varphi_{\beta,\gamma}}(\mu,\nu) = \begin{cases} \displaystyle\int_{\mathcal{X}} \max\left\{0, \gamma - \beta\frac{d\mu}{d\nu}\right\} \mathrm{d}\nu & \text{if } \gamma \leq \beta, \\ \displaystyle\int_{\mathcal{X}} \max\left\{0, \beta\frac{d\mu}{d\nu} - \gamma\right\} \mathrm{d}\nu & \text{if } \gamma > \beta, \end{cases}
$$

while for the statistical information, using the first formula from Proposition 4.2.13, we have:

$$
\mathcal{I}_{(\beta,\gamma)}(\mu,\nu) = \begin{cases} \displaystyle\int_{\mathcal{X}} \max\left\{0, \gamma - \beta\frac{d\mu}{d\nu}\right\} \mathrm{d}\nu & \text{if } \gamma\nu(\mathcal{X}) \leq \beta\mu(\mathcal{X}), \\ \displaystyle\int_{\mathcal{X}} \max\left\{0, \beta\frac{d\mu}{d\nu} - \gamma\right\} \mathrm{d}\nu & \text{if } \gamma\nu(\mathcal{X}) \geq \beta\mu(\mathcal{X}). \end{cases}
$$

Thus, it becomes evident how the two divergences are defined piecewise, with the definitions differing solely in the condition pertaining to which piece is taken: in the f-divergence what matters is the relation between β and γ, whereas in the statistical information these two quantities are weighted by the total mass of the measures. From this point of view, it is evident that when the total masses are equal the conditions are identical.

Moreover, we can notice how in the statistical information the two pieces coincide when we have equality in the condition, while in the f-divergence

they do not, suggesting that the statistical information is a more natural concept.

At this point, we can finally prove the DPI for general spaces.

Theorem 5.3.9 (DPI, general spaces)**.** *Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ finite measures, $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ a Markov kernel and $f\colon \mathbb{R} \to \mathbb{R}$ a convex function. Then*

$$D_f(\kappa \circ \mu, \kappa \circ \nu) \leq D_f(\mu, \nu).$$

Proof. By Remark 5.0.1, it is possible to assume without loss of generality that $\mu \ll \nu$. Furthermore, invoking Remark 3.1.6 and the fact that the total mass of the measures is invariant under the composition with a Markov kernel (see Proposition 2.1.19), we can assume that $f(1) = f'(1) = 0$.

We begin by demonstrating that the DPI is satisfied for the hockey-stick function, using the formula from Proposition 5.3.7 and the DPI for the statistical information from Proposition 4.2.12:

$$\begin{aligned}
D_{\varphi_{1,y}}(\kappa \circ \mu, \kappa \circ \nu) &= \mathcal{I}_{(1,y)}(\kappa \circ \mu, \kappa \circ \nu) + \frac{1}{2}\left|(\kappa \circ \mu)(\mathcal{X}) - y(\kappa \circ \nu)(\mathcal{X})\right| \\
&\quad - \frac{1}{2}\operatorname{sgn}(1-y)\left((\kappa \circ \mu)(\mathcal{X}) - y(\kappa \circ \nu)(\mathcal{X})\right) \\
&\leq \mathcal{I}_{(1,y)}(\mu, \nu) + \frac{1}{2}\left|(\kappa \circ \mu)(\mathcal{X}) - y(\kappa \circ \nu)(\mathcal{X})\right| && \text{By Proposition 4.2.12} \\
&\quad - \frac{1}{2}\operatorname{sgn}(1-y)\left((\kappa \circ \mu)(\mathcal{X}) - y(\kappa \circ \nu)(\mathcal{X})\right) \\
&= \mathcal{I}_{(1,y)}(\mu, \nu) + \frac{1}{2}\left|\mu(\mathcal{X}) - y\nu(\mathcal{X})\right| - \frac{1}{2}\operatorname{sgn}(1-y)\left(\mu(\mathcal{X}) - y\nu(\mathcal{X})\right) && \text{By Proposition 2.1.19} \\
&= D_{\varphi_{1,y}}(\mu, \nu). && \text{By Proposition 5.3.7}
\end{aligned}$$

We can then use the integral representation of the f-divergence from Proposition 5.3.5, which is applicable given that $\mu \ll \nu$ implies $\kappa \circ \mu \ll \kappa \circ \nu$, and the monotonicity of the integral to get the desired result:

$$D_f(\kappa \circ \mu, \kappa \circ \nu) = \int_{\mathbb{R}} D_{\varphi_{1,y}}(\kappa \circ \mu, \kappa \circ \nu)\, \mathrm{d}\gamma_f(y)$$

$$\begin{aligned} &\leq \int_{\mathbb{R}} D_{\varphi_{1,y}}(\mu,\nu)\,\mathrm{d}\gamma_f(y) \\ &= D_f(\mu,\nu). \end{aligned}$$

⋔

Remark 5.3.10. As previously mentioned in Remark 5.3.6, the formalization of this proof took a slightly different approach than the informal proof. In particular, the generalization to the case where the measures are not absolutely continuous and the function is not centered was not done in the proof of the DPI, but at the level of the integral representation of the f-divergence and the formula linking the statistical information and the f-divergence of the hockey-stick function.

Remark 5.3.11. Note that the current proof of the DPI for general spaces is not sorry-free, as it depends on a result whose proof is not completely formalized. In particular, the proof of Remark 5.3.6 relies on the integration by parts theorem (Theorem 2.3.1), which has not yet been formalized in Mathlib (see Remark 2.3.7). Nevertheless, this is a well-established result that can be relied upon with a reasonable degree of confidence (see also Remark C.0.5), and we are confident that it will be formalized in the future.

Remark 5.3.12. The following is the formalization of this statement of the DPI:

```
lemma fDiv_comp_right_le' (η : Kernel 𝒳 𝒳') [IsMarkovKernel η]
    (μ ν : Measure α) [IsFiniteMeasure μ] [IsFiniteMeasure ν]
    (hf_cvx : ConvexOn ℝ univ f) (hf_cont : Continuous f) :
    fDiv f (η ∘ₘ μ) (η ∘ₘ ν) ≤ fDiv f μ ν := by
```

We can compare it with the formalized statement for the DPI for standard Borel spaces in Remark 5.2.6 and see how the latter requires the first space to be standard Borel, as well as some additional countability condition. In contrast, this statement applies to every measurable space. However, we can also notice how the convexity hypotheses are slightly different. In this case, the function is required to be convex on the entire real line, whereas in the

standard Borel case, it is only required to be convex on the interval $[0, \infty)$. The first condition is more stringent than the second one. This issue can be partially mitigated by noting that the function can be modified at will on the negative numbers; this allows us to extend any convex function on $[0, \infty)$ that has a finite right derivative at 0 to a convex function on the whole real line. Nevertheless, if the right derivative at 0 is $-\infty$ (which is the case for $x \mapsto x \log x$, used to define the Kullback-Leibler divergence), then it is not possible to extend it to a convex function on the whole real line.

This limitation of the current proof arises from the fact that some of the employed tools, in particular the definitions of Stieltjes function and curvature measure, are implemented in Mathlib only for functions that have certain properties on the entire real line. To extend the result, further work is required. One possible approach is to generalize the definition of Stieltjes function to allow the function to take infinite values. This would enable any function that is monotone and right continuous on an interval to be extended to a Stieltjes function. In fact, the function could be defined to be $-\infty$ on the left of the interval, and $+\infty$ on the right.

5.4 Consequences of the DPI

In this section we will examine some consequences of the DPI for f-divergences. In particular, we will see how other inequalities involving the operations between kernels and measures can be derived from it and how it implies the DPI for the Rényi divergence.

Lemma 5.4.1. *Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, $\mu, \nu \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ finite measures and $f \colon \mathbb{R} \to \mathbb{R}$ a convex function. Then we have:*

i) $D_f(\mu_{\mathcal{X}}, \nu_{\mathcal{X}}) \leq D_f(\mu, \nu)$,

ii) $D_f(\mu_{\mathcal{Y}}, \nu_{\mathcal{Y}}) \leq D_f(\mu, \nu)$.

Proof. Notice how we can write the marginals of a measure as the composition with the deterministic kernels associated with the projections, which will be

denoted by $\kappa_{\mathcal{X}} \coloneqq (x, y) \mapsto x$ and $\kappa_{\mathcal{Y}} \coloneqq (x, y) \mapsto y$. The DPI can then be employed to obtain the desired result:

$$D_f(\mu_{\mathcal{X}}, \nu_{\mathcal{X}}) = D_f(\kappa_{\mathcal{X}} \circ \mu, \kappa_{\mathcal{X}} \circ \nu) \leq D_f(\mu, \nu).$$

And similarly:

$$D_f(\mu_{\mathcal{Y}}, \nu_{\mathcal{Y}}) = D_f(\kappa_{\mathcal{Y}} \circ \mu, \kappa_{\mathcal{Y}} \circ \nu) \leq D_f(\mu, \nu).$$

⋔

We can use this lemma to prove two other inequalities and, indeed, an equality involving the composition product.

Proposition 5.4.2. *Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ finite measures, $\kappa, \eta \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ a Markov kernel and $f \colon \mathbb{R} \to \mathbb{R}$ a convex function. Then we have:*

i) $D_f(\mu, \nu) \leq D_f(\mu \otimes \kappa, \nu \otimes \eta)$,

ii) $D_f(\kappa \circ \mu, \eta \circ \nu) \leq D_f(\mu \otimes \kappa, \nu \otimes \eta)$,

iii) $D_f(\mu \otimes \kappa, \nu \otimes \kappa) = D_f(\mu, \nu)$.

Proof. **(i), (ii)** Using Proposition A.0.1, we can use the marginals of the composition product to write the measure and the composition, then we conclude using Lemma 5.4.1:

$$D_f(\mu, \nu) = D_f((\mu \otimes \kappa)_{\mathcal{X}}, (\nu \otimes \eta)_{\mathcal{X}}) \leq D_f(\mu \otimes \kappa, \nu \otimes \eta).$$

And similarly for the second part:

$$D_f(\kappa \circ \mu, \eta \circ \nu) = D_f((\mu \otimes \kappa)_{\mathcal{Y}}, (\nu \otimes \eta)_{\mathcal{Y}}) \leq D_f(\mu \otimes \kappa, \nu \otimes \eta).$$

(iii) We have already established one inequality from (i), so we only need to prove the other one. To do that, recall that in Remark 2.1.18 we saw how it is possible to write the composition product as a composition: $\mu \otimes \kappa = (id_{\mathcal{X}} \times \kappa) \circ \mu$. We can therefore use the DPI once again:

$$D_f(\mu \otimes \kappa, \nu \otimes \kappa) = D_f((id_{\mathcal{X}} \times \kappa) \circ \mu, (id_{\mathcal{X}} \times \kappa) \circ \nu) \leq D_f(\mu, \nu).$$

⋔

Remark 5.4.3. It is interesting to note that in this case we deduce the inequality (i) of the previous proposition as a corollary of the DPI. However, in the proof of the DPI for standard Borel spaces from Section 5.2, the same inequality (i) was proven separately (Proposition 5.2.1) and then used as the first step to prove the DPI. The same observation is true for the equality $D_f(\mu_{\leftrightarrow}, \nu_{\leftrightarrow}) = D_f(\mu, \nu)$ from Lemma 5.2.4, which can easily be demonstrated by noticing that $(\cdot)_{\leftrightarrow}$ is an involution and can be expressed as the composition with a deterministic kernel, and then using the DPI twice.

Different combinations of those properties and the DPI imply each other in various ways, making it possible to prove them in different orders. For example, as we just saw in the proof of Proposition 5.4.2, the DPI implies all the other properties. However, if we have (i), (iii), the equality for the swap kernel and the existence of the Bayesian inverse then we can prove the DPI (this is basically the proof of the DPI for standard Borel spaces from Section 5.2). Moreover, if we have (ii) and (iii) we can bypass certain steps in that same proof and directly establish the DPI, obviating the necessity for the Bayesian inverse.

Finally, let us demonstrate how the DPI for f-divergences, in particular for the Hellinger divergence, implies the DPI for the Rényi divergence.

Theorem 5.4.4 (DPI for Rényi divergence)**.** *Let $\mathcal{X}$ be a measurable space, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ finite measures, $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ a Markov kernel and $\alpha \in [0, \infty)$ such that the DPI for the Hellinger α-divergence holds. Then*

$$\mathrm{R}_\alpha(\kappa \circ \mu, \kappa \circ \nu) \leq \mathrm{R}_\alpha(\mu, \nu).$$

Proof. For $\alpha = 1$ we have that $\mathrm{R}_1 = \mathrm{KL} = \mathrm{H}_1$, so there is nothing to prove.

Let us now consider the case where $\alpha \neq 1$. Then, once we fix ν, the function $t \mapsto \frac{1}{\alpha-1}\log(\nu(\mathcal{X}) + (\alpha-1)t)$ is nondecreasing. Therefore, we can apply the DPI for the Hellinger α-divergence to get the result:

$$\begin{aligned}
\mathrm{R}_\alpha(\kappa \circ \mu, \kappa \circ \nu) &= \frac{1}{\alpha-1}\log\left((\kappa \circ \nu)(\mathcal{X}) + (\alpha-1)\,\mathrm{H}_\alpha(\kappa \circ \mu, \kappa \circ \nu)\right) \\
\text{By Proposition 2.1.19} &= \frac{1}{\alpha-1}\log\left(\nu(\mathcal{X}) + (\alpha-1)\,\mathrm{H}_\alpha(\kappa \circ \mu, \kappa \circ \nu)\right)
\end{aligned}$$

$$\begin{aligned} &\leq \frac{1}{\alpha-1}\log\left(\nu(\mathcal{X})+(\alpha-1)\,\mathrm{H}_\alpha(\mu,\nu)\right) && \text{DPI for } \mathrm{H}_\alpha \\ &= \mathrm{R}_\alpha(\mu,\nu). \end{aligned}$$

⋔

Remark 5.4.5. For $\alpha > 1$, we have that $(f_\alpha)'_+(0)$ is finite, thus the Hellinger function can be extended to a convex function over the entire real line, by defining it to be linear with slope $(f_\alpha)'_+(0)$ for $t < 0$. Therefore, we can apply Theorem 5.3.9 to have the DPI for the Hellinger α-divergence and then extend it to the Rényi divergence with Theorem 5.4.4.

However, when $\alpha \leq 1$ it is not possible to extend the Hellinger function to the entire real line in a convex manner. Therefore, with the current formalization we can prove the DPI for the Hellinger and the Rényi divergences for $\alpha \leq 1$ exclusively within the context of standard Borel spaces, using Theorem 5.2.5. Nonetheless, it should be feasible, with some additional effort, to extend the proof of Theorem 5.3.9 to functions that are only convex on the nonnegative reals (see Remark 5.3.12).

Conclusions

This thesis has presented a formalization project aimed at contributing to the understanding of information theory and its applications to hypothesis testing. We leveraged the power of the interactive theorem prover Lean 4 to formalize key concepts, including f-divergences, the framework for estimation problems, and the data processing inequality.

Our work resulted in several key achievements:

- **Formalization of f-divergences**: We formalized a comprehensive API for f-divergences, treating them in a general setting that extends beyond probability measures. This allows for broader applications and theoretical insights.
- **Formalization of the DPI**: We presented three distinct proofs of the DPI for f-divergences. These proofs vary in their generality and the assumptions they require, thereby offering different perspectives on this fundamental result.
- **Contribution to Mathlib**: The results that we have incorporated to the Mathlib library during the project, and the ones we will add in the near future, solidify the foundations of probability and information theory in Lean 4, and pave the way for further research in these areas.

The chosen approach of using an interactive theorem prover offers a number of advantages:

- **Increased Rigor**: Formal proofs eliminate errors and ambiguities that can arise in traditional mathematical reasoning.

- **Replicability**: All code and proofs are readily available for verification and reusable by others.
- **Generalization**: The proof assistant helped us to keep track of very fine-grained assumptions, allowing us to generalize results in a more systematic way and gain insights that would have been harder to obtain otherwise.
- **Facilitated Collaboration**: The use of a shared repository and the automatic verification of proofs enabled a seamless collaboration, allowing us to focus our discussions on high-level concepts and insights.

One of the most interesting observations that emerged the project is the seemingly unexpected connection between the DPI and the statistical information. While the DPI is a fundamental result in information theory, statistical information arises from the domain of hypothesis testing. As we proved in Section 5.3, the DPI can be regarded as a consequence of a similar inequality for the statistical information. However, the relationship between these two domains goes in the opposite direction, too. Indeed, inequalities involving the information divergences can be used to provide lower bounds on the error of a statistical estimator in a hypothesis testing problem. Moreover, if the test involves a variable number of samples, as in Example 4.0.4, we can use those bounds to obtain results about the sample complexity, quantifying how many observations necessary to achieve a specified level of confidence.

In summary, this thesis illustrates the efficacy of interactive theorem provers in formalizing intricate mathematical concepts and contributes to the expanding corpus of formalized knowledge in information theory and its applications.

Bibliography

[Apo74] Tom M Apostol. *Mathematical analysis; 2nd ed.* Addison-Wesley series in mathematics. Reading, MA: Addison-Wesley, 1974. URL: https://cds.cern.ch/record/105425.

[Avi24] Jeremy Avigad. "Mathematics and the formal turn". In: *Bulletin of the American Mathematical Society* 61 (2024), pp. 225–240.

[CA10] Andrzej Cichocki and Shun-ichi Amari. "Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities". In: *Entropy* 12.6 (2010), pp. 1532–1568. ISSN: 1099-4300. DOI: 10.3390/e12061532. URL: https://www.mdpi.com/1099-4300/12/6/1532.

[Car19] Mario Carneiro. "The Type Theory of Lean". In: (2019). URL: https://github.com/digama0/lean-type-theory/releases.

[CH88] Thierry Coquand and Gérard Huet. "The calculus of constructions". In: *Information and Computation* 76.2 (1988), pp. 95–120. ISSN: 0890-5401. DOI: https://doi.org/10.1016/0890-5401(88)90005-3. URL: https://www.sciencedirect.com/science/article/pii/0890540188900053.

[Cle+17] Florence Clerc et al. "Pointless learning". English. In: *20th International Conference on Foundations of Software Science and Computation Structures (FoSSaCS 2017)*. Lecture Notes in Computer Science. 20th International Conference on Foundations of Software Science and Computation Structures, FoSSaCS 2017 ;

Conference date: 22-04-2017 Through 29-04-2017. Springer, Mar. 2017, pp. 355–369. ISBN: 978-3-662-54457-0. DOI: 10.1007/978-3-662-54458-7_21. URL: https://www.etaps.org/index.php/2017/fossacs.

[Dah+18] Fredrik Dahlqvist et al. "Borel Kernels and their Approximation, Categorically". In: *Electronic Notes in Theoretical Computer Science* 341 (2018). Proceedings of the Thirty-Fourth Conference on the Mathematical Foundations of Programming Semantics (MFPS XXXIV), pp. 91–119. ISSN: 1571-0661. DOI: https://doi.org/10.1016/j.entcs.2018.11.006. URL: https://www.sciencedirect.com/science/article/pii/S1571066118300860.

[Dud02] R. M. Dudley. *Real Analysis and Probability*. 2nd ed. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002.

[EH14] Tim van Erven and Peter Harremos. "Rényi Divergence and Kullback-Leibler Divergence". In: *IEEE Transactions on Information Theory* 60.7 (2014), pp. 3797–3820. DOI: 10.1109/TIT.2014.2320500.

[Fri+23] Tobias Fritz et al. "Weakly Markov Categories and Weakly Affine Monads". In: *10th Conference on Algebra and Coalgebra in Computer Science (CALCO 2023)*. Ed. by Paolo Baldan and Valeria de Paiva. Vol. 270. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023, 16:1–16:17. ISBN: 978-3-95977-287-7. DOI: 10.4230/LIPIcs.CALCO.2023.16. URL: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.CALCO.2023.16.

[Hyt+16] Tuomas Hytönen et al. *Analysis in Banach Spaces: Volume I: Martingales and Littlewood-Paley Theory*. English. Ergebnisse

der Mathematik und ihrer Grenzgebiete. 3. Folge. Springer, 2016. ISBN: 978-3-319-48519-5. DOI: 10.1007/978-3-319-48520-1.

[Kal21] Olav Kallenberg. *Foundations of Modern Probability*. Jan. 2021. ISBN: 978-3-030-61870-4. DOI: 10.1007/978-3-030-61871-1.

[Lie12] Friedrich Liese. "ϕPHI-divergences, sufficiency, Bayes sufficiency, and deficiency". eng. In: *Kybernetika* 48.4 (2012), pp. 690–713. URL: http://eudml.org/doc/246799.

[LV06] F. Liese and I. Vajda. "On Divergences and Informations in Statistics and Information Theory". In: *IEEE Transactions on Information Theory* 52.10 (2006), pp. 4394–4412. DOI: 10.1109/TIT.2006.881731.

[Mou+15] Leonardo de Moura et al. "The lean theorem prover (system description)". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9195 (2015). Cited by: 250, pp. 378–388. DOI: 10.1007/978-3-319-21401-6_26. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84984612110&doi=10.1007%2f978-3-319-21401-6_26&partnerID=40&md5=41b8c0e54a38e3d1e5710fa2167dcc80.

[MU21] Leonardo de Moura and Sebastian Ullrich. "The Lean 4 Theorem Prover and Programming Language". In: *Automated Deduction – CADE 28*. Ed. by André Platzer and Geoff Sutcliffe. Cham: Springer International Publishing, 2021, pp. 625–635. ISBN: 978-3-030-79876-5.

[Per24] Paolo Perrone. "Markov Categories and Entropy". In: *IEEE Transactions on Information Theory* 70.3 (2024), pp. 1671–1692. DOI: 10.1109/TIT.2023.3328825.

[PW24] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2024.

[Rén65] A. Rényi. "On the Foundations of Information Theory". In: *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* 33.1 (1965), pp. 1–14. ISSN: 03731138. URL: http://www.jstor.org/stable/1401301 (visited on 08/28/2024).

[SUM20] Daniel Selsam, Sebastian Ullrich, and Leonardo de Moura. *Tabled Typeclass Resolution*. 2020. arXiv: 2001.04301 [cs.PL]. URL: https://arxiv.org/abs/2001.04301.

[Tao23a] Terence Tao. *A slightly longer Lean 4 proof tour*. Dec. 5, 2023. URL: https://terrytao.wordpress.com/2023/12/05/a-slightly-longer-lean-4-proof-tour/ (visited on 09/10/2024).

[Tao23b] Terence Tao. *Formalizing the proof of PFR in Lean4 using Blueprint: a short tour*. Nov. 18, 2023. URL: https://terrytao.wordpress.com/2023/11/18/formalizing-the-proof-of-pfr-in-lean4-using-blueprint-a-short-tour/ (visited on 09/10/2024).

[VO18] Matthijs Vákár and Luke Ong. *On S-Finite Measures and Kernels*. 2018. arXiv: 1810.01837 [math.PR].

[Whe15] R.L. Wheeden. *Measure and Integral: An Introduction to Real Analysis, Second Edition*. Chapman & Hall/CRC Pure and Applied Mathematics. CRC Press, 2015. ISBN: 9781498702904. URL: https://books.google.fr/books?id=X3d3CAAAQBAJ.

Appendix A

Properties of kernel operations

This appendix will present some properties of kernels that we use in the main text. In particular, we are going to examine an interesting relation between the composition and the composition product of a Markov kernel and a measure. We will also investigate how the composition and the parallel product of kernels behave in relation to the composition and the swap kernel, and what happens when we compose a constant kernel.

Proposition A.0.1. *Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $\mu \in \mathcal{M}(\mathcal{X})$. Then:*

i) $(\mu \otimes \kappa)_{\mathcal{X}} = \mu$,

ii) $(\mu \otimes \kappa)_{\mathcal{Y}} = \kappa \circ \mu$,

where $(\mu \otimes \kappa)_{\mathcal{X}}$ and $(\mu \otimes \kappa)_{\mathcal{Y}}$ are the marginals of the composition product $\mu \otimes \kappa$, respectively on $\mathcal{X}$ and $\mathcal{Y}$.

Proof. **(i)** Let $A \in \mathscr{F}_{\mathcal{X}}$, then by definition of marginal measure and composition product we have

$$(\mu \otimes \kappa)_{\mathcal{X}}(A) = (\mu \otimes \kappa)(A \times \mathcal{Y}) = \int_A \kappa(x, \mathcal{Y})\mu(dx) = \int_A 1\mu(dx) = \mu(A).$$

(ii) Let $B \in \mathscr{F}_{\mathcal{Y}}$, then by definition of marginal measure and composition product we have

$$(\mu \otimes \kappa)_{\mathcal{Y}}(B) = (\mu \otimes \kappa)(\mathcal{X} \times B) = \int_{\mathcal{X}} \kappa(x, B)\mu(dx) = \kappa \circ \mu(B).$$

⫷

Proposition A.0.2. *Let $\mathcal{X}, \mathcal{X}', \mathcal{X}'', \mathcal{Y}, \mathcal{Y}', \mathcal{Y}''$ be measurable spaces, $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{X}'$, $\eta\colon \mathcal{Y} \rightsquigarrow \mathcal{Y}'$, $\kappa'\colon \mathcal{X}' \rightsquigarrow \mathcal{X}''$ and $\eta'\colon \mathcal{Y}' \rightsquigarrow \mathcal{Y}''$. Then:*

- $(\kappa' \parallel \eta') \circ (\kappa \parallel \eta) = (\kappa' \circ \kappa) \parallel (\eta' \circ \eta)$,
- $(\kappa \parallel \eta)_{\leftrightarrow} = \eta \parallel \kappa$.

Moreover, if $\mathcal{X} = \mathcal{Y}$ then:

- $(\kappa' \parallel \eta') \circ (\kappa \times \eta) = (\kappa' \circ \kappa) \times (\eta' \circ \eta)$,
- $(\kappa \times \eta)_{\leftrightarrow} = \eta \times \kappa$.

Proof. For the proof refer to the Lean code. ⫲

Proposition A.0.3. *Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be measurable spaces, $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$, $\eta \colon \mathcal{Y} \rightsquigarrow \mathcal{Z}$ a constant kernel (i.e. $\eta(x, \cdot) = \tilde{\eta}$ for every x) and $\mu \in \mathcal{M}(\mathcal{Y})$. Then:*

i) $\eta \circ \kappa(x) = \kappa(x, \mathcal{Y})\tilde{\eta}$, for every $x \in \mathcal{X}$,

ii) $\eta \circ \mu = \mu(\mathcal{Y})\tilde{\eta}$.

Proof. **(i)** For every $B \subseteq \mathcal{Z}$ measurable we have

$$(\eta \circ \kappa)(x, B) = \int_{\mathcal{Y}} \eta(y, B)\kappa(x, dy) = \int_{\mathcal{Y}} \tilde{\eta}(B)\kappa(x, dy) = \kappa(x, \mathcal{Y})\tilde{\eta}(B).$$

(ii) This is just a special case of (i) with κ also constant. ⫲

Appendix B

Lebesgue decomposition

This appendix presents two propositions that demonstrate some useful properties of the Lebesgue decomposition of measures and kernels. For the proofs we refer to [Kal21] and the Lean code.

Proposition B.0.1. *Let $\mathcal{X}$ be a measurable space and $\mu, \nu, \xi \in \mathcal{M}(\mathcal{X})$ such that a Lebesgue decomposition between μ and ν exists. Then the following properties hold:*

i) $\mu \ll \nu$ *if and only if* $\mu_{\perp\nu} = 0$.

ii) $\mu \perp \nu$ *if and only if* $\frac{d\mu}{d\nu} \cdot \nu = 0$, *that is* $\frac{d\mu}{d\nu} = 0$ *almost everywhere with respect to* ν. *This is also equivalent to* $\mu_{\perp\nu} = \mu$.

iii) If $\mu \ll \nu$, *then for every* μ*-integrable function* f *we have*

$$\int f \,\mathrm{d}\mu = \int f \frac{d\mu}{d\nu} \,\mathrm{d}\nu.$$

In particular $\mu(A) = \int_A \frac{d\mu}{d\nu} \,\mathrm{d}\nu$ *for every measurable set* A.

iv) If μ, ν *and* ξ *are* σ*-finite and* $\mu \ll \nu$, *then for almost every* $x \in \mathcal{X}$ *with respect to* ξ *we have*

$$\frac{d\mu}{d\xi}(x) = \frac{d\mu}{d\nu}(x) \frac{d\nu}{d\xi}(x).$$

Proposition B.0.2. *Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces such that $\mathcal{X}$ is countable or $\mathcal{Y}$ is countably generated, let $\kappa, \eta \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ be finite kernels and $\mu, \nu \in \mathcal{M}(\mathcal{X})$ finite measures. Then the following properties hold:*

i) For every $x \in \mathcal{X}$ *and for almost every* $y \in \mathcal{Y}$ *with respect to* $\eta(x)$

$$\frac{d\kappa}{d\eta}(x,y) = \frac{d\kappa(x)}{d\eta(x)}(y).$$

ii) For every $x \in \mathcal{X}$, $\kappa_{\perp\eta}(x) = \kappa(x)_{\perp\eta(x)}$.

iii) For almost every $(x,y) \in \mathcal{X} \times \mathcal{Y}$ *with respect to* $\nu \otimes \eta$

$$\frac{d(\mu \otimes \kappa)}{d(\nu \otimes \eta)}(x,y) = \frac{d\mu}{d\nu}(x)\frac{d\kappa}{d\eta}(x,y).$$

iv) For almost every $(x,y) \in \mathcal{X} \times \mathcal{Y}$ *with respect to* $\nu \otimes \eta$

$$\frac{d(\mu \otimes \kappa)}{d(\nu \otimes \kappa)}(x,y) = \frac{d\mu}{d\nu}(x).$$

v) $(\mu \otimes \kappa)_{\perp(\nu\otimes\eta)} = \mu_{\perp\nu} \otimes \kappa + \left(\frac{d\mu}{d\nu} \cdot \nu\right) \otimes \kappa_{\perp\eta}$.

Appendix C

Riemann-Stieltjes integral

In this appendix we will introduce the definition of the Riemann-Stieltjes integral, and present some basic properties associated with it. For a more detailed discussion see [Apo74, Chapter 7], see also [Whe15, Chapter 11.3] for a more complete introduction to the Lebesgue-Stieltjes integral.

The Riemann-Stieltjes integral represents a generalization of the Riemann integral. Their definitions are, in fact, quite similar, the only difference being that in the Riemann-Stieltjes integral the weights for the Riemann sum are not the lengths of the intervals, but rather the differences between the values of a function at the endpoints. Indeed, if we take that function to be the identity, we recover the classic Riemann integral.

Definition C.0.1 (Riemann-Stieltjes integral)**.** Let $a, b \in \mathbb{R}$ such that $a < b$, $f, g\colon [a,b] \to \mathbb{R}$ be bounded functions. We call a *partition* of $[a,b]$ a finite sequence of points $P = \{x_0, x_1, \dots, x_n\}$ such that $a = x_0 < x_1 < \dots < x_n = b$; moreover we denote by $t_P = \{t_1, \dots, t_n\}$ a set of points such that $t_k \in [x_{k-1}, x_k]$ for every $k = 1, \dots, n$ and $\Delta g_k \coloneqq g(t_k) - g(t_{k-1})$. We define the *Riemann-Stieltjes sum* of f with respect to g over the partition P as

$$S(P, f, g) \coloneqq \sum_{k=1}^{n} f(t_k)\Delta g_k.$$

Then, f is said to be *Riemann-Stieltjes integrable* with respect to g on $[a,b]$ if there exists a number $A \in \mathbb{R}$ such that for every $\varepsilon > 0$ there exists a partition P_ε of $[a,b]$ such that for every partition $P \subseteq P_\varepsilon$ and every choice of points t_P we have $|S(P,f,g) - A| < \varepsilon$. In this case, we denote the number A by $\int_a^b f \,\mathrm{d}g$, or $\int_a^b f(x)\,\mathrm{d}g(x)$, and we call it the *Riemann-Stieltjes integral* of f with respect to g on $[a,b]$.

Theorem C.0.2. *Let $a, b \in \mathbb{R}$ such that $a < b$, $f, g\colon [a,b] \to \mathbb{R}$ such that f is continuous and g has bounded variation. Then f is Riemann-Stieltjes integrable with respect to g.*

Proof. See [Apo74, Theorem 7.27]. ⁂

The following theorem guarantees that the Lebesgue-Stieltjes integral and the Riemann-Stieltjes integral coincide when both are defined.

Theorem C.0.3. *Let $a, b \in \mathbb{R}$ such that $a < b$, $f, g\colon [a,b] \to \mathbb{R}$ such that f is nondecreasing and right continuous and g is bounded, measurable and Riemann-Stieltjes integrable with respect to f. Then*

$$\int_{(a,b]} g \,\mathrm{d}\Lambda_f = \int_a^b g \,\mathrm{d}f.$$

Proof. See [Whe15, Theorem 11.11]. ⁂

Remark C.0.4. Note that, while the Riemann-Stieltjes integral is a generalization of the Riemann integral, the Lebesgue-Stieltjes integral is not a generalization of the Lebesgue integral, but rather a special case of it, since it is a Lebesgue integral with respect to a particular measure.

Indeed, the Lebesgue integral is strictly more general than the Lebesgue-Stieltjes integral, as there exist measures that are not of the form Λ_f for any Stieltjes function f. This can be readily observed by noting that any Lebesgue-Stieltjes measure is finite on bounded intervals, but not all measures satisfy this property (for instance the counting measure on the real line does not).

Remark C.0.5. We will now demonstrate that the statement of the integration by parts theorem `integral_stieltjes_meas_by_parts` that we have formalized (see Remark 2.3.7) is a consequence of Theorem 2.3.1.

First of all, note that the integrals in the Lean statement are actually Lebesgue-Stieltjes integrals, in fact `f` and `g` are Stieltjes functions, and the notation ∂`f.measure` indicates the Lebesgue-Stieltjes measure associated with `f`, which is defined in Mathlib. Moreover, f is continuous by the hypothesis

`hf` and g is of bounded variation on $[a, b]$, since it is a Stieltjes function; therefore, by Theorem C.0.2 f is Riemann-Stieltjes integrable with respect to g; by Theorem 2.3.1 we also have that g is Riemann-Stieltjes integrable with respect to f. These integrability conditions are a prerequisite to apply the other theorems, and also serve to guarantee us that the integrals in the Lean code are well-defined and do not take junk values. We now conclude using Theorem C.0.3 (which is applicable, since g is a Stieltjes function and thus bounded on $[a, b]$ and measurable) to show that the integrals in the statement of Theorem 2.3.1 are actually equal to Lebesgue-Stieltjes integrals. Therefore, under our hypotheses, the statement of Theorem 2.3.1 is equivalent to the one we have formalized.

This remark also serves as a proof of Corollary 2.3.6.

Appendix D

Convex functions

In this appendix we will present a straightforward inequality for convex functions and introduce the concept of curvature measure associated with a convex function.

First, let us establish the necessary notation.

Definition D.0.1. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function. We denote by f'_+ the right derivative of f, whenever it is defined. Moreover, we denote by $f'(\infty) \coloneqq \lim_{t\to\infty} f'_+(t)$ the limit of the right derivative at infinity, whenever it exists.

Remark D.0.2. If f is convex, then the right derivative exists at every point. Moreover, f'_+ is nondecreasing, thus ensuring the existence of the limit $f'(\infty)$.

In Lean, we define $f'(\infty)$ as follows:

```
def derivAtTop (f : ℝ → ℝ) : EReal :=
  limsup (fun x ↦ (rightDeriv f x : EReal)) atTop
```

where `limsup ... atTop` is the limsup of a function as the argument goes to infinity, whereas `rightDeriv f` is the right derivative of f, which takes the junk value 0 at the points where the right derivative does not exist. This, in conjunction with the fact that the limsup always exists, guarantees `derivAtTop f` to be well-defined for every function $f\colon \mathbb{R} \to \mathbb{R}$, regardless of its properties.

Since in Lean there is currently no definition of right (or left) derivative, we had to define it ourselves, based on the more general `derivWithin`, and prove some API lemmas about it, including linearity lemmas, results about the right derivative of convex functions and its relation with the slope.

Now we proceed to prove a simple inequality for convex functions that involves $f'(\infty)$.

Lemma D.0.3. *Let $f\colon \mathbb{R} \to \mathbb{R}$ be a convex function. Then for every $x, y \in \mathbb{R}$ we have*

$$f(x+y) \leq f(x) + yf'(\infty).$$

Proof. If $y = 0$ the inequality is trivial.

Let us assume that $y > 0$. Since the slope of a convex function is always less than its right derivative at the right endpoint, we have that

$$\frac{f(x+y) - f(x)}{y} \leq f'_+(x+y) \leq \lim_{t\to\infty} f'_+(t) = f'(\infty),$$

where the second inequality comes from the fact that the right derivative of a convex function is nondecreasing.

Analogously, if $y < 0$ we have that

$$\frac{f(x+y) - f(x)}{y} \leq f'_+(x) \leq \lim_{t\to\infty} f'_+(t) = f'(\infty).$$

⁂

Remark D.0.4. Note that the same lemma holds true under less restrictive assumptions on the convexity of f, for example see `le_add_derivAtTop''` in the Lean code for a proof in the case where f is convex only on $[0, +\infty)$ and $x, y \geq 0$.

Convex functions are also linked to Stieltjes functions. Indeed, if f is convex, then its right derivative is defined everywhere, nondecreasing and right continuous, therefore it is a Stieltjes function, and we can define the Lebesgue-Stieltjes measure associated with it.

Definition D.0.5 (Curvature measure)**.** Let $f\colon \mathbb{R} \to \mathbb{R}$ be a convex function. Then the Lebesgue-Stieltjes measure associated with the right derivative of f is referred to as the *curvature measure* of f, and we denote it by $\gamma_f \coloneqq \Lambda_{f'_+}$.

Remark D.0.6. The name *curvature measure* comes from the fact that if f is twice differentiable then

$$\gamma_f(A) = \int_A f''(x)\,\mathrm{d}x.$$

Appendix E

Properties of f-divergences

In this appendix we list some additional properties of the f-divergences defined in Section 3.1.

Proposition E.0.1. *Let $\mathcal{X}$ be a measurable space, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ and f a convex function such that $f(1) = 0$. Then the following hold:*

i) if $\mu \ll \nu$, we have that $D_f(\mu, \nu) = \int_{\mathcal{X}} f\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\nu$,

ii) if $\mu \perp \nu$, we have that $D_f(\mu, \nu) = f(0)\nu(\mathcal{X}) + f'(\infty)\mu(\mathcal{X})$,

iii) $D_f(\mu, \mu) = 0$.

Proof. **((i), (ii))** They follow trivially from Proposition B.0.1.
(iii) This is a special case of (i), using the fact that $f(1) = 0$. ⁜

Proposition E.0.2. *Let $\mathcal{X}$ be a measurable space, $\mu, \mu_1, \mu_2, \nu \in \mathcal{M}(\mathcal{X})$ finite measures and f a convex function such that $f(1) = 0$. Then the following hold:*

i) if $\mu_1 \ll \nu$ and $\mu_2 \perp \nu$, then $D_f(\mu_1 + \mu_2, \nu) = D_f(\mu_1, \nu) + \mu_2(\mathcal{X}) f'(\infty)$,

ii) $D_f(\mu, \nu) = D_f\left(\frac{d\mu}{d\nu} \cdot \nu, \nu\right) + \mu_{\perp\nu}(\mathcal{X}) f'(\infty)$,

iii) if $\mu_1 \ll \nu$ and $\mu_2 \ll \nu$, then $D_f(\mu_1 + \mu_2, \nu) \leq D_f(\mu_1, \nu) + \mu_2(\mathcal{X}) f'(\infty)$,

iv) $D_f(\mu_1 + \mu_2, \nu) \leq D_f(\mu_1, \nu) + \mu_2(\mathcal{X}) f'(\infty)$.

Proof. **(i)** From Proposition B.0.1 it follows that $\frac{d(\mu_1 + \mu_2)}{d\nu} = \frac{d\mu_1}{d\nu} + \frac{d\mu_2}{d\nu} = \frac{d\mu_1}{d\nu}$ ν-a.e., and $(\mu_1 + \mu_2)_{\perp\nu} = \mu_{1\perp\nu} + \mu_{2\perp\nu} = \mu_2$. Therefore, by definition of f-divergence we have

$$D_f(\mu_1 + \mu_2, \nu) = \int_{\mathcal{X}} f\left(\frac{d\mu_1 + \mu_2}{d\nu}\right) \mathrm{d}\nu + f'(\infty)(\mu_1 + \mu_2)_{\perp\nu}(\mathcal{X})$$

$$
\begin{aligned}
&= \int_{\mathcal{X}} f\left(\frac{d\mu_1}{d\nu}\right) \mathrm{d}\nu + f'(\infty)\mu_2(\mathcal{X}) \\
&= D_f(\mu_1, \nu) + \mu_2(\mathcal{X}) f'(\infty).
\end{aligned}
$$

(ii) This is an immediate consequence of (i).

(iii) From Lemma D.0.3 we have that for every $x \in \mathcal{X}$

$$
f\left(\frac{d\mu_1}{d\nu}(x) + \frac{d\mu_2}{d\nu}(x)\right) \leq f\left(\frac{d\mu_1}{d\nu}(x)\right) + \frac{d\mu_2}{d\nu}(x) f'(\infty),
$$

moreover $\mu_1 + \mu_2 \ll \nu$, therefore

$$
\begin{aligned}
D_f(\mu_1 + \mu_2, \nu) &= \int_{\mathcal{X}} f\left(\frac{d\mu_1 + \mu_2}{d\nu}\right) \mathrm{d}\nu \\
&= \int_{\mathcal{X}} f\left(\frac{d\mu_1}{d\nu} + \frac{d\mu_2}{d\nu}\right) \mathrm{d}\nu \\
&\leq \int_{\mathcal{X}} f\left(\frac{d\mu_1}{d\nu}\right) \mathrm{d}\nu + \int_{\mathcal{X}} \frac{d\mu_2}{d\nu} f'(\infty)\, \mathrm{d}\nu \\
&= D_f(\mu_1, \nu) + \mu_2(\mathcal{X}) f'(\infty).
\end{aligned}
$$

(iv) Using the Lebesgue decomposition (Theorem 2.2.3) we have that

$$
\begin{aligned}
D_f(\mu_1 + \mu_2, \nu) &= D_f\left(\frac{d\mu_1}{d\nu} \cdot \nu + \mu_{1\perp\nu} + \frac{d\mu_2}{d\nu} \cdot \nu + \mu_{2\perp\nu}, \nu\right) \\
\text{By (i)} \qquad &= D_f\left(\frac{d\mu_1}{d\nu} \cdot \nu + \frac{d\mu_2}{d\nu} \cdot \nu, \nu\right) + \mu_{1\perp\nu}(\mathcal{X}) f'(\infty) + \mu_{2\perp\nu}(\mathcal{X}) f'(\infty) \\
\text{By (iii)} \qquad &\leq D_f\left(\frac{d\mu_1}{d\nu} \cdot \nu, \nu\right) + \left(\frac{d\mu_2}{d\nu} \cdot \nu\right)(\mathcal{X}) f'(\infty) + \mu_{1\perp\nu}(\mathcal{X}) f'(\infty) + \mu_{2\perp\nu}(\mathcal{X}) f'(\infty) \\
\text{By (i)} \qquad &= D_f(\mu_1, \nu) + \left(\frac{d\mu_2}{d\nu} \cdot \nu + \mu_{2\perp\nu}\right)(\mathcal{X}) f'(\infty) \\
&= D_f(\mu_1, \nu) + \mu_2(\mathcal{X}) f'(\infty).
\end{aligned}
$$

⋕