Information theory in Lean: the DPI

Lorenzo Luccioli

January 15, 2025

Overview

- TLS (Testing lower bounds) project, led by Rémy Degenne.

Overview

- TLS (Testing lower bounds) project, led by Rémy Degenne.
- Information theory studies how to quantify information.

Overview

- TLS (Testing lower bounds) project, led by Rémy Degenne.
- Information theory studies how to quantify information.
- Information divergences: given μ, ν two probability distributions, $D(\mu, \nu)$ measures how easy it is to tell them apart.

Overview

- TLS (Testing lower bounds) project, led by Rémy Degenne.
- Information theory studies how to quantify information.
- Information divergences: given μ, ν two probability distributions, $D(\mu, \nu)$ measures how easy it is to tell them apart.
- Data processing inequality (DPI): if we process the data produced by μ and ν, we cannot make them easier to distinguish.

$$D(\kappa \circ \mu, \kappa \circ \nu) \leq D(\mu, \nu)$$

Overview

1. Transition kernels
2. Information divergences
3. Data Processing Inequality (DPI)

Transition kernels

Definition of kernel

Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, with their respective σ-algebras $\mathscr{F}_{\mathcal{X}}, \mathscr{F}_{\mathcal{Y}}$, also let $\mathcal{M}(\mathcal{X})$ be the set of measures on $\mathcal{X}$.

Definition of kernel

Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, with their respective σ-algebras $\mathscr{F}_{\mathcal{X}}, \mathscr{F}_{\mathcal{Y}}$, also let $\mathcal{M}(\mathcal{X})$ be the set of measures on $\mathcal{X}$.

Definition (Transition kernel)

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:

Definition of kernel

Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, with their respective σ-algebras $\mathscr{F}_\mathcal{X}, \mathscr{F}_\mathcal{Y}$, also let $\mathcal{M}(\mathcal{X})$ be the set of measures on $\mathcal{X}$.

Definition (Transition kernel)

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 $\mathcal{Y}$,

Definition of kernel

Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, with their respective σ-algebras $\mathscr{F}_{\mathcal{X}}, \mathscr{F}_{\mathcal{Y}}$, also let $\mathcal{M}(\mathcal{X})$ be the set of measures on $\mathcal{X}$.

Definition (Transition kernel)

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 $\mathcal{Y}$,
- for every $B \in \mathscr{F}_{\mathcal{Y}}$ the function $x \mapsto \kappa(x, B)$ is measurable.

Definition of kernel

Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, with their respective σ-algebras $\mathscr{F}_{\mathcal{X}}, \mathscr{F}_{\mathcal{Y}}$, also let $\mathcal{M}(\mathcal{X})$ be the set of measures on $\mathcal{X}$.

Definition (Transition kernel)

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 $\mathcal{Y}$,
- for every $B \in \mathscr{F}_{\mathcal{Y}}$ the function $x \mapsto \kappa(x, B)$ is measurable.

We write $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$.

Definition of kernel

Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, with their respective σ-algebras $\mathscr{F}_\mathcal{X}, \mathscr{F}_\mathcal{Y}$, also let $\mathcal{M}(\mathcal{X})$ be the set of measures on $\mathcal{X}$.

Definition (Transition kernel)

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 $\mathcal{Y}$,
- for every $B \in \mathscr{F}_\mathcal{Y}$ the function $x \mapsto \kappa(x, B)$ is measurable.

We write $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$.
If $\kappa(x, \cdot)$ is a probability measure for all x, we say that κ is a *Markov kernel*.

Definition of kernel

Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces, with their respective σ-algebras $\mathscr{F}_{\mathcal{X}}, \mathscr{F}_{\mathcal{Y}}$, also let $\mathcal{M}(\mathcal{X})$ be the set of measures on $\mathcal{X}$.

Definition (Transition kernel)

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 $\mathcal{Y}$,
- for every $B \in \mathscr{F}_{\mathcal{Y}}$ the function $x \mapsto \kappa(x, B)$ is measurable.

We write $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$.
If $\kappa(x, \cdot)$ is a probability measure for all x, we say that κ is a *Markov kernel*.

A kernel $\mathcal{X} \rightsquigarrow \mathcal{Y}$ can also be seen as a measurable function $\mathcal{X} \to \mathcal{M}(\mathcal{Y})$.

Lean implementation

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

Lean implementation

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

The measurable structure on `Measure β` is the canonical one given by the projection maps $\mu \mapsto \mu(s)$ for $s \subseteq \beta$ measurable, and is automatically inferred by Lean.

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

What are kernels?

Kernels are a way of representing the idea of a transformation that involves some randomness, in other words, a stochastic function.

What are kernels?

Kernels are a way of representing the idea of a transformation that involves some randomness, in other words, a stochastic function.

Kernels are a generalization of:

What are kernels?

Kernels are a way of representing the idea of a transformation that involves some randomness, in other words, a stochastic function.

Kernels are a generalization of:

- measures: $x \mapsto \mu$ is a *constant kernel*

What are kernels?

Kernels are a way of representing the idea of a transformation that involves some randomness, in other words, a stochastic function.

Kernels are a generalization of:

- measures: $x \mapsto \mu$ is a *constant kernel*
- measurable functions: $x \mapsto \delta_{f(x)}$ is a *deterministic kernel*

Kernel operations

Let $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $\mu \in \mathcal{M}(\mathcal{X})$. We can combine κ and μ in various ways.

Kernel operations

Let $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $\mu \in \mathcal{M}(\mathcal{X})$. We can combine κ and μ in various ways.

- Composition: $\kappa \circ \mu \in \mathcal{M}(\mathcal{Y})$. Also $\eta \circ \kappa : \mathcal{X} \rightsquigarrow \mathcal{Z}$ for $\eta : \mathcal{Y} \rightsquigarrow \mathcal{Z}$.

Kernel operations

Let $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $\mu \in \mathcal{M}(\mathcal{X})$. We can combine κ and μ in various ways.

- Composition: $\kappa \circ \mu \in \mathcal{M}(\mathcal{Y})$. Also $\eta \circ \kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Z}$ for $\eta\colon \mathcal{Y} \rightsquigarrow \mathcal{Z}$.
- Posterior kernel (or Bayesian inverse): $\kappa_\mu^\dagger\colon \mathcal{Y} \rightsquigarrow \mathcal{X}$.

Kernel operations

Let $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $\mu \in \mathcal{M}(\mathcal{X})$. We can combine κ and μ in various ways.

- Composition: $\kappa \circ \mu \in \mathcal{M}(\mathcal{Y})$. Also $\eta \circ \kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Z}$ for $\eta \colon \mathcal{Y} \rightsquigarrow \mathcal{Z}$.
- Posterior kernel (or Bayesian inverse): $\kappa_\mu^\dagger \colon \mathcal{Y} \rightsquigarrow \mathcal{X}$.
  - $\kappa_\mu^\dagger$ is not guaranteed to exist. Some sufficient hypotheses are: κ, μ finite, $\mathcal{X}$ standard Borel space (complete, separable metric space with Borel σ-algebra).

Kernel operations

Let $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $\mu \in \mathcal{M}(\mathcal{X})$. We can combine κ and μ in various ways.

- Composition: $\kappa \circ \mu \in \mathcal{M}(\mathcal{Y})$. Also $\eta \circ \kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Z}$ for $\eta\colon \mathcal{Y} \rightsquigarrow \mathcal{Z}$.
- Posterior kernel (or Bayesian inverse): $\kappa_\mu^\dagger\colon \mathcal{Y} \rightsquigarrow \mathcal{X}$.
  - $\kappa_\mu^\dagger$ is not guaranteed to exist. Some sufficient hypotheses are: κ, μ finite, $\mathcal{X}$ standard Borel space (complete, separable metric space with Borel σ-algebra).
- Composition product: $\kappa \otimes \mu \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$. Also $\kappa \otimes \eta\colon \mathcal{X} \rightsquigarrow \mathcal{Y} \times \mathcal{Z}$ for $\eta\colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$.

Kernel operations

Let $\kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $\mu \in \mathcal{M}(\mathcal{X})$. We can combine κ and μ in various ways.

- Composition: $\kappa \circ \mu \in \mathcal{M}(\mathcal{Y})$. Also $\eta \circ \kappa\colon \mathcal{X} \rightsquigarrow \mathcal{Z}$ for $\eta\colon \mathcal{Y} \rightsquigarrow \mathcal{Z}$.
- Posterior kernel (or Bayesian inverse): $\kappa_\mu^\dagger\colon \mathcal{Y} \rightsquigarrow \mathcal{X}$.
  - $\kappa_\mu^\dagger$ is not guaranteed to exist. Some sufficient hypotheses are: κ, μ finite, $\mathcal{X}$ standard Borel space (complete, separable metric space with Borel σ-algebra).
- Composition product: $\kappa \otimes \mu \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$. Also $\kappa \otimes \eta\colon \mathcal{X} \rightsquigarrow \mathcal{Y} \times \mathcal{Z}$ for $\eta\colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$.
- Product: $\kappa \times \eta\colon \mathcal{X} \rightsquigarrow \mathcal{Y} \times \mathcal{Z}$ for $\eta\colon \mathcal{X} \rightsquigarrow \mathcal{Z}$.

Kernel operations

Let $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $\mu \in \mathcal{M}(\mathcal{X})$. We can combine κ and μ in various ways.

- Composition: $\kappa \circ \mu \in \mathcal{M}(\mathcal{Y})$. Also $\eta \circ \kappa : \mathcal{X} \rightsquigarrow \mathcal{Z}$ for $\eta : \mathcal{Y} \rightsquigarrow \mathcal{Z}$.
- Posterior kernel (or Bayesian inverse): $\kappa_\mu^\dagger : \mathcal{Y} \rightsquigarrow \mathcal{X}$.
    - $\kappa_\mu^\dagger$ is not guaranteed to exist. Some sufficient hypotheses are: κ, μ finite, $\mathcal{X}$ standard Borel space (complete, separable metric space with Borel σ-algebra).
- Composition product: $\kappa \otimes \mu \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$. Also $\kappa \otimes \eta : \mathcal{X} \rightsquigarrow \mathcal{Y} \times \mathcal{Z}$ for $\eta : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$.
- Product: $\kappa \times \eta : \mathcal{X} \rightsquigarrow \mathcal{Y} \times \mathcal{Z}$ for $\eta : \mathcal{X} \rightsquigarrow \mathcal{Z}$.
- Parallel product: $\kappa \parallel \eta : \mathcal{X} \times \mathcal{Z} \rightsquigarrow \mathcal{Y} \times \mathcal{W}$ for $\eta : \mathcal{Z} \rightsquigarrow \mathcal{W}$.

Kernel operations

Let $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $\mu \in \mathcal{M}(\mathcal{X})$. We can combine κ and μ in various ways.

- Composition: $\kappa \circ \mu \in \mathcal{M}(\mathcal{Y})$. Also $\eta \circ \kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Z}$ for $\eta \colon \mathcal{Y} \rightsquigarrow \mathcal{Z}$.
- Posterior kernel (or Bayesian inverse): $\kappa_\mu^\dagger \colon \mathcal{Y} \rightsquigarrow \mathcal{X}$.
  - $\kappa_\mu^\dagger$ is not guaranteed to exist. Some sufficient hypotheses are: κ, μ finite, $\mathcal{X}$ standard Borel space (complete, separable metric space with Borel σ-algebra).
- Composition product: $\kappa \otimes \mu \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$. Also $\kappa \otimes \eta \colon \mathcal{X} \rightsquigarrow \mathcal{Y} \times \mathcal{Z}$ for $\eta \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$.
- Product: $\kappa \times \eta \colon \mathcal{X} \rightsquigarrow \mathcal{Y} \times \mathcal{Z}$ for $\eta \colon \mathcal{X} \rightsquigarrow \mathcal{Z}$.
- Parallel product: $\kappa \parallel \eta \colon \mathcal{X} \times \mathcal{Z} \rightsquigarrow \mathcal{Y} \times \mathcal{W}$ for $\eta \colon \mathcal{Z} \rightsquigarrow \mathcal{W}$.

Remark (Lean notation)

- Composition : $\kappa \circ_m \mu$, $\eta \circ_k \kappa$
- Comp. prod. : $\mu \otimes_m \kappa$, $\kappa \otimes_k \eta$
- Product : $\kappa \times_k \eta$
- Parallel product : $\kappa \parallel_k \eta$
- Posterior kernel : $\kappa\dagger\mu$

Information divergences

Information divergence

- A measure of dissimilarity between two probability distributions, in particular it is a function of the following shape:

$$D\colon \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \overline{\mathbb{R}}$$

Information divergence

- A measure of dissimilarity between two probability distributions, in particular it is a function of the following shape:

$$D\colon \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \overline{\mathbb{R}}$$

- Not necessarily a metric, can be asymmetric.

Information divergence

- A measure of dissimilarity between two probability distributions, in particular it is a function of the following shape:

$$D \colon \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \overline{\mathbb{R}}$$

- Not necessarily a metric, can be asymmetric.
- Some properties that a good divergence should verify, at least for probability measures μ, ν:

Information divergence

- A measure of dissimilarity between two probability distributions, in particular it is a function of the following shape:

$$D\colon \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \overline{\mathbb{R}}$$

- Not necessarily a metric, can be asymmetric.
- Some properties that a good divergence should verify, at least for probability measures μ, ν:
  - $D(\mu, \mu) = 0$.

Information divergence

- A measure of dissimilarity between two probability distributions, in particular it is a function of the following shape:

$$D \colon \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \overline{\mathbb{R}}$$

- Not necessarily a metric, can be asymmetric.
- Some properties that a good divergence should verify, at least for probability measures μ, ν:
  - $D(\mu, \mu) = 0$.
  - Non-negativity: $D(\mu, \nu) \geq 0$.

Information divergence

- A measure of dissimilarity between two probability distributions, in particular it is a function of the following shape:

$$D\colon \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \overline{\mathbb{R}}$$

- Not necessarily a metric, can be asymmetric.
- Some properties that a good divergence should verify, at least for probability measures μ, ν:
  - $D(\mu, \mu) = 0$.
  - Non-negativity: $D(\mu, \nu) \geq 0$.
  - Data processing inequality: $D(\kappa \circ \mu, \kappa \circ \nu) \leq D(\mu, \nu)$ for every Markov kernel κ.

Definition of f-divergence

Definition (f-divergence)

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

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

Definition of f-divergence

Definition (f-divergence)

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

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

Reminder: Lebesgue decomposition

If we have two (sigma finite) measures $\mu, \nu \in \mathcal{M}(\mathcal{X})$, we can decompose μ into two parts, respectively absolutely continuous and singular w.r.t. ν:

$$\mu = \frac{d\mu}{d\nu} \cdot \nu + \mu_{\perp\nu}.$$

Definition of f-divergence

Definition (f-divergence)

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

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

Reminder: Lebesgue decomposition

If we have two (sigma finite) measures $\mu, \nu \in \mathcal{M}(\mathcal{X})$, we can decompose μ into two parts, respectively absolutely continuous and singular w.r.t. ν:

$$\mu = \frac{d\mu}{d\nu} \cdot \nu + \mu_{\perp\nu}.$$

$\frac{d\mu}{d\nu}$ and $\mu_{\perp\nu}$ are known respectively as the *Radon-Nikodym derivative* of and the *singular part* of μ with respect to ν.

Definition of f-divergence

Definition (f-divergence)

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

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

Remark

If μ is absolutely continuous with respect to ν ($\mu \ll \nu$), then the definition simplifies to

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

Definition of f-divergence

Definition (f-divergence)

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

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

The class of f-divergences is very broad, and includes many well-known divergences.

- Kullback-Leibler divergence, with $f(x) = x \log x$.
- Total variation, with $f(x) = \frac{1}{2}|x-1|$.
- χ^2-divergence, with $f(x) = (x-1)^2$.
- Hellinger α-divergence, with $f_\alpha(x) = \frac{x^\alpha - 1}{\alpha - 1}$ and $\alpha \in (0, +\infty) \setminus \{1\}$.

Definition of f-divergence

Definition (f-divergence)

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

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

Lean definition (old)

```
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 univ
```

Refactor (ongoing)

Old definition

```
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 univ
```

Refactor (ongoing)

Old definition

```
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 univ
```

Some issues with this definition:

Refactor (ongoing)

Old definition

```
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 univ
```

Some issues with this definition:

- Domain of f: $\mathbb{R}$ instead of $\mathbb{R}_{\geq 0 \infty}$ (`ENNReal`), forces us to use `.toReal`. Derivatives and convexity in Mathlib work better with real functions.

Refactor (ongoing)

Old definition

```
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 univ
```

Some issues with this definition:

- Domain of f: $\mathbb{R}$ instead of $\mathbb{R}{\geq}0\infty$ (`ENNReal`), forces us to use `.toReal`. Derivatives and convexity in Mathlib work better with real functions.
- Codomain of f: we would like to allow infinite values.

Refactor (ongoing)

Old definition

```
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 univ
```

Some issues with this definition:

- Domain of f: $\mathbb{R}$ instead of $\mathbb{R}_{\geq 0 \infty}$ (`ENNReal`), forces us to use `.toReal`. Derivatives and convexity in Mathlib work better with real functions.
- Codomain of f: we would like to allow infinite values.
- Bochner integral: forces us to use an if statement, cumbersome to use.

Refactor (ongoing)

Old definition

```
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 univ
```

Some issues with this definition:

- Domain of f: $\mathbb{R}$ instead of $\mathbb{R}_{\geq 0 \infty}$ (`ENNReal`), forces us to use `.toReal`. Derivatives and convexity in Mathlib work better with real functions.
- Codomain of f: we would like to allow infinite values.
- Bochner integral: forces us to use an if statement, cumbersome to use.
- Output type: `EReal`. Do we really need/want to allow negative values?

Refactor (ongoing)

Old definition

```
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 univ
```

New definition

```
structure DivFunction where
  toFun : ℝ≥0∞ → ℝ≥0∞
  one : toFun 1 = 0
  convexOn' : ConvexOn ℝ≥0 univ toFun
  continuous' : Continuous toFun

def fDiv (f : DivFunction) (μ ν : Measure α) : ℝ≥0∞ :=
  ∫⁻ x, f ((∂μ/∂ν) x) ∂ν + f.derivAtTop * μ.singularPart ν .univ
```

Data Processing Inequality (DPI)

Data Processing Inequality for f-divergences

Theorem (Data Processing Inequality)

Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ finite measures, $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ a Markov kernel, and $f : \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}$ a convex function such that $f(1) = 0$. Then

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

Data Processing Inequality for f-divergences

Theorem (Data Processing Inequality)

Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ finite measures, $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ a Markov kernel, and $f : \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}$ a convex function such that $f(1) = 0$. Then

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

- A fundamental result in information theory.

Data Processing Inequality for f-divergences

Theorem (Data Processing Inequality)

Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ finite measures, $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ a Markov kernel, and $f : \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}$ a convex function such that $f(1) = 0$. Then

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

- A fundamental result in information theory.
- The amount of information cannot be increased by processing data through a (potentially random) transformation.

Data Processing Inequality for f-divergences

Theorem (Data Processing Inequality)

Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ finite measures, $\kappa \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ a Markov kernel, and $f : \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}$ a convex function such that $f(1) = 0$. Then

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

- A fundamental result in information theory.
- The amount of information cannot be increased by processing data through a (potentially random) transformation.
- If we want to distinguish two distributions based on data sampled from them, we cannot make our job easier by processing the samples.

Data Processing Inequality for f-divergences

Theorem (Data Processing Inequality)

Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ finite measures, $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ a Markov kernel, and $f : \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}$ a convex function such that $f(1) = 0$. Then

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

- A fundamental result in information theory.
- The amount of information cannot be increased by processing data through a (potentially random) transformation.
- If we want to distinguish two distributions based on data sampled from them, we cannot make our job easier by processing the samples.
- We formalized several proofs of this result, along with many auxiliary results, like posterior kernels, hypothesis testing, lemmas about `EReal`, convex functions, `rnDeriv`, etc.

Different proofs

1. Deterministic kernels

Different proofs

1. Deterministic kernels
   - Only applies if κ is a deterministic kernel.

Different proofs

1. Deterministic kernels
   - Only applies if κ is a deterministic kernel.
   - Use the conditional Jensen's inequality (still sorried).

Different proofs

1. Deterministic kernels
   - Only applies if κ is a deterministic kernel.
   - Use the conditional Jensen's inequality (still sorried).
2. Standard Borel spaces

Different proofs

1. Deterministic kernels
    - Only applies if κ is a deterministic kernel.
    - Use the conditional Jensen's inequality (still sorried).
2. Standard Borel spaces
    - Only applies if $\mathcal{X}, \mathcal{Y}$ are complete, separable metric spaces with Borel σ-algebras.

Different proofs

1. Deterministic kernels
    - Only applies if κ is a deterministic kernel.
    - Use the conditional Jensen's inequality (still sorried).
2. Standard Borel spaces
    - Only applies if $\mathcal{X}, \mathcal{Y}$ are complete, separable metric spaces with Borel σ-algebras.
    - Use Jensen's inequality.

Different proofs

1. Deterministic kernels
   - Only applies if κ is a deterministic kernel.
   - Use the conditional Jensen's inequality (still sorried).
2. Standard Borel spaces
   - Only applies if $\mathcal{X}, \mathcal{Y}$ are complete, separable metric spaces with Borel σ-algebras.
   - Use Jensen's inequality.
   - Use the existence of the posterior kernels $\kappa_\mu^\dagger$ and $\kappa_\nu^\dagger$.

Different proofs

1. Deterministic kernels
   - Only applies if κ is a deterministic kernel.
   - Use the conditional Jensen's inequality (still sorried).
2. Standard Borel spaces
   - Only applies if $\mathcal{X}, \mathcal{Y}$ are complete, separable metric spaces with Borel σ-algebras.
   - Use Jensen's inequality.
   - Use the existence of the posterior kernels $\kappa_\mu^\dagger$ and $\kappa_\nu^\dagger$.
3. General spaces using hypothesis testing

Different proofs

1. Deterministic kernels
   - Only applies if κ is a deterministic kernel.
   - Use the conditional Jensen's inequality (still sorried).
2. Standard Borel spaces
   - Only applies if $\mathcal{X}, \mathcal{Y}$ are complete, separable metric spaces with Borel σ-algebras.
   - Use Jensen's inequality.
   - Use the existence of the posterior kernels $\kappa_\mu^\dagger$ and $\kappa_\nu^\dagger$.
3. General spaces using hypothesis testing
   - Use a parametric family of divergences $\mathcal{I}_{(\beta,\gamma)}$ (*statistical information*) that naturally emerges from the hypothesis testing framework.

Different proofs

1. Deterministic kernels
   - Only applies if κ is a deterministic kernel.
   - Use the conditional Jensen's inequality (still sorried).
2. Standard Borel spaces
   - Only applies if $\mathcal{X}, \mathcal{Y}$ are complete, separable metric spaces with Borel σ-algebras.
   - Use Jensen's inequality.
   - Use the existence of the posterior kernels $\kappa_\mu^\dagger$ and $\kappa_\nu^\dagger$.
3. General spaces using hypothesis testing
   - Use a parametric family of divergences $\mathcal{I}_{(\beta,\gamma)}$ (*statistical information*) that naturally emerges from the hypothesis testing framework.
   - $\mathcal{I}_{(\beta,\gamma)}$ naturally satisfies the DPI.

Different proofs

1. Deterministic kernels
   - Only applies if κ is a deterministic kernel.
   - Use the conditional Jensen's inequality (still sorried).
2. Standard Borel spaces
   - Only applies if $\mathcal{X}, \mathcal{Y}$ are complete, separable metric spaces with Borel σ-algebras.
   - Use Jensen's inequality.
   - Use the existence of the posterior kernels $\kappa_\mu^\dagger$ and $\kappa_\nu^\dagger$.
3. General spaces using hypothesis testing
   - Use a parametric family of divergences $\mathcal{I}_{(\beta,\gamma)}$ (*statistical information*) that naturally emerges from the hypothesis testing framework.
   - $\mathcal{I}_{(\beta,\gamma)}$ naturally satisfies the DPI.
   - Use a representation of f-divergences in terms of an integral of $\mathcal{I}_{(1,\gamma)}$ (this needs a generalization of the integration by parts, still sorried).

Different proofs

1. Deterministic kernels
   - Only applies if κ is a deterministic kernel.
   - Use the conditional Jensen's inequality (still sorried).
2. Standard Borel spaces
   - Only applies if $\mathcal{X}, \mathcal{Y}$ are complete, separable metric spaces with Borel σ-algebras.
   - Use Jensen's inequality.
   - Use the existence of the posterior kernels $\kappa_\mu^\dagger$ and $\kappa_\nu^\dagger$.
3. General spaces using hypothesis testing
   - Use a parametric family of divergences $\mathcal{I}_{(\beta,\gamma)}$ (*statistical information*) that naturally emerges from the hypothesis testing framework.
   - $\mathcal{I}_{(\beta,\gamma)}$ naturally satisfies the DPI.
   - Use a representation of f-divergences in terms of an integral of $\mathcal{I}_{(1,\gamma)}$ (this needs a generalization of the integration by parts, still sorried).
4. General spaces

Different proofs

1. Deterministic kernels
   - Only applies if κ is a deterministic kernel.
   - Use the conditional Jensen's inequality (still sorried).
2. Standard Borel spaces
   - Only applies if $\mathcal{X}, \mathcal{Y}$ are complete, separable metric spaces with Borel σ-algebras.
   - Use Jensen's inequality.
   - Use the existence of the posterior kernels $\kappa_\mu^\dagger$ and $\kappa_\nu^\dagger$.
3. General spaces using hypothesis testing
   - Use a parametric family of divergences $\mathcal{I}_{(\beta,\gamma)}$ (*statistical information*) that naturally emerges from the hypothesis testing framework.
   - $\mathcal{I}_{(\beta,\gamma)}$ naturally satisfies the DPI.
   - Use a representation of f-divergences in terms of an integral of $\mathcal{I}_{(1,\gamma)}$ (this needs a generalization of the integration by parts, still sorried).
4. General spaces
   - Similar to 2.

Different proofs

1. Deterministic kernels
   - Only applies if κ is a deterministic kernel.
   - Use the conditional Jensen's inequality (still sorried).
2. Standard Borel spaces
   - Only applies if $\mathcal{X}, \mathcal{Y}$ are complete, separable metric spaces with Borel σ-algebras.
   - Use Jensen's inequality.
   - Use the existence of the posterior kernels $\kappa_\mu^\dagger$ and $\kappa_\nu^\dagger$.
3. General spaces using hypothesis testing
   - Use a parametric family of divergences $\mathcal{I}_{(\beta,\gamma)}$ (*statistical information*) that naturally emerges from the hypothesis testing framework.
   - $\mathcal{I}_{(\beta,\gamma)}$ naturally satisfies the DPI.
   - Use a representation of f-divergences in terms of an integral of $\mathcal{I}_{(1,\gamma)}$ (this needs a generalization of the integration by parts, still sorried).
4. General spaces
   - Similar to 2.
   - Use 1 to skip part of the proof, avoiding the need for $\kappa_\mu^\dagger$ and $\kappa_\nu^\dagger$.

Integral represantation of f-divergences

- Given two probability measures μ, ν and a prior $\pi = (\beta, \gamma)$, an estimator is a function that tries to distinguish between μ and ν based on a data sample.

Integral represantation of f-divergences

- Given two probability measures μ, ν and a prior $\pi = (\beta, \gamma)$, an estimator is a function that tries to distinguish between μ and ν based on a data sample.
- $\mathcal{I}_{(\beta,\gamma)}(\mu, \nu)$ is the difference in performance between the best possible estimator and a "blind" estimator that cannot look at the data.

Integral representation of f-divergences

- Given two probability measures μ, ν and a prior $\pi = (\beta, \gamma)$, an estimator is a function that tries to distinguish between μ and ν based on a data sample.
- $\mathcal{I}_{(\beta,\gamma)}(\mu, \nu)$ is the difference in performance between the best possible estimator and a "blind" estimator that cannot look at the data.

Theorem (Integral representation of f-divergences)

Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function such that $f(1) = f'_+(1) = 0$, $\mu, \nu \in \mathcal{M}(\mathcal{X})$ probability measures such that $\mu \ll \nu$. Then

$$D_f(\mu, \nu) = \int_{\mathbb{R}_+} \mathcal{I}_{(1,y)}(\mu, \nu) \, \mathrm{d}\gamma_f(y).$$

Where γ_f is the curvature measure of f.

Summary and Conclusions

Done

- `fDiv`, definition and API
- Exampes of information divergences: Kulback-Leibler, Hellinger, Rényi, Statistical Information.
- Hypothesis testing, definitions and connection with f-divergences
- Data Processing Inequality for f-divergences
- Posterior kernels and other operations on kernels, results about existing operations
- Other results about `EReal`, convex functions, `rnDeriv`, etc.

Summary and Conclusions

Done

- `fDiv`, definition and API
- Exampes of information divergences: Kulback-Leibler, Hellinger, Rényi, Statistical Information.
- Hypothesis testing, definitions and connection with f-divergences
- Data Processing Inequality for f-divergences
- Posterior kernels and other operations on kernels, results about existing operations
- Other results about `EReal`, convex functions, `rnDeriv`, etc.

Still To Do

- Generalized integration by parts (Riemann-Stieltjes integral)
- Conditional Jensen's inequality
- Complete the refactor of `fDiv`
- Porting to Mathlib

Summary and Conclusions

Done

- `fDiv`, definition and API
- Exampes of information divergences: Kulback-Leibler, Hellinger, Rényi, Statistical Information.
- Hypothesis testing, definitions and connection with f-divergences
- Data Processing Inequality for f-divergences
- Posterior kernels and other operations on kernels, results about existing operations
- Other results about `EReal`, convex functions, `rnDeriv`, etc.

Still To Do

- Generalized integration by parts (Riemann-Stieltjes integral)
- Conditional Jensen's inequality
- Complete the refactor of `fDiv`
- Porting to Mathlib

Thank you for your attention :)