Entropy under disintegrations

Juan Pablo Vigneaux Ariztía
jpvigneaux@gmail.com

GSI 2021 - Harmonic Analysis on Lie Groups

Paris, 22 July 2021

Outline

(Generalized) differential entropy

Setting: $(E_X, \mathfrak{B})$ measurable space, μ reference "volume" measure (σ-finite).

1. E_X a countable set, $\mathfrak{B}$ atomic σ-algebra, μ counting measure;
2. E_X euclidean space, $\mathfrak{B}$ its Borel σ-algebra, μ Lebesgue measure;
3. E_X a locally compact topological group, $\mathfrak{B}$ its Borel σ-algebra, μ some (left) Haar measure;
4. $(E_X, \mathfrak{B})$ arbitrary and μ a probability measure on it (prior/initial state).

(Generalized) differential entropy

Setting: $(E_X, \mathfrak{B})$ measurable space, μ reference "volume" measure (σ-finite).

1. E_X a countable set, $\mathfrak{B}$ atomic σ-algebra, μ counting measure;
2. E_X euclidean space, $\mathfrak{B}$ its Borel σ-algebra, μ Lebesgue measure;
3. E_X a locally compact topological group, $\mathfrak{B}$ its Borel σ-algebra, μ some (left) Haar measure;
4. $(E_X, \mathfrak{B})$ arbitrary and μ a probability measure on it (prior/initial state).

Given a probability measure ρ on $(E_X, \mathfrak{B}_X)$ absolutely continuous w.r.t. μ (i.e. $\rho \ll \mu$), we define the **generalized differential entropy** as

$$S_\mu(\rho) := \mathbb{E}_\rho\left(-\ln \frac{\mathrm{d}\rho}{\mathrm{d}\mu}\right) = -\int_{E_X} \frac{\mathrm{d}\rho}{\mathrm{d}\mu}(x) \log \frac{\mathrm{d}\rho}{\mathrm{d}\mu}(x)\, \mathrm{d}\mu(x). \tag{1}$$

Theorem (Asymptotic Equipartition Property (AEP))

$(E_X, \mathfrak{B}, \mu)$ *and* ρ *as before. Suppose that* $S_\mu(\rho)$ *is finite. Let* $\{X_i : (\Omega, \mathfrak{F}, \mathbb{P}) \to (E_X, \mathfrak{B}, \mu)\}_{i\in\mathbb{N}}$ *be a collection of i.i.d random variables with law* ρ*.*

For every $\delta > 0$*, introduce the* **typical set**

$$A_\delta^{(n)}(\rho;\mu) := \left\{ (x_1, ..., x_n) \in E_X^n \;\middle|\; \left| -\frac{1}{n} \log f_{X_1,...,X_n}(x_1, ..., x_n) - S_\mu(\rho) \right| \leq \delta \right\},$$

where $f_{X_1,...,X_n}$ *is the joint density.*

Theorem (Asymptotic Equipartition Property (AEP))

$(E_X, \mathfrak{B}, \mu)$ and ρ as before. Suppose that $S_\mu(\rho)$ is finite. Let $\{X_i : (\Omega, \mathfrak{F}, \mathbb{P}) \to (E_X, \mathfrak{B}, \mu)\}_{i\in\mathbb{N}}$ be a collection of i.i.d random variables with law ρ.
For every $\delta > 0$, introduce the **typical set**

$$A_\delta^{(n)}(\rho; \mu) := \left\{ (x_1, \ldots, x_n) \in E_X^n \mid \left| -\frac{1}{n} \log f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) - S_\mu(\rho) \right| \leq \delta \right\},$$

where $f_{X_1,\ldots,X_n}$ is the joint density. Then, for every $\varepsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$,

$$\mathbb{P}\left(A_\delta^{(n)}(\rho; \mu)\right) > 1 - \varepsilon,$$

Theorem (Asymptotic Equipartition Property (AEP))

$(E_X, \mathfrak{B}, \mu)$ and ρ as before. Suppose that $S_\mu(\rho)$ is finite. Let $\{X_i : (\Omega, \mathfrak{F}, \mathbb{P}) \to (E_X, \mathfrak{B}, \mu)\}_{i \in \mathbb{N}}$ be a collection of i.i.d random variables with law ρ.
For every $\delta > 0$, introduce the **typical set**

$$A_\delta^{(n)}(\rho; \mu) := \left\{ (x_1, \ldots, x_n) \in E_X^n \,\middle|\, \left| -\frac{1}{n} \log f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) - S_\mu(\rho) \right| \leq \delta \right\},$$

where $f_{X_1, \ldots, X_n}$ is the joint density. Then, for every $\varepsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$,

$$\mathbb{P}\left(A_\delta^{(n)}(\rho; \mu) \right) > 1 - \varepsilon,$$

and moreover

$$(1 - \varepsilon) \exp\{n(S_\mu(\rho) - \delta)\} \leq \mu^{\otimes n}(A_\delta^{(n)}(\rho; \mu)) \leq \exp\{n(S_\mu(\rho) + \delta)\}.$$

Entropy as exponential growth rate

Corollary

$$\lim_{\delta \to 0} \lim_{n \to \infty} \frac{1}{n} \log \mu^{\otimes n}(A_\delta^{(n)}(\rho; \mu)) = S_\mu(\rho).$$

More imprecisely, but concisely: when n is big and δ small,

$$\mu^{\otimes n}(A_\delta^{(n)}(\rho; \mu)) \approx e^{nS_\mu(\rho)}$$

and the joint density is $\approx e^{-nS_\mu(\rho)}$ on this set.

Entropy as exponential growth rate

Corollary

$$\lim_{\delta\to 0}\lim_{n\to\infty}\frac{1}{n}\log\mu^{\otimes n}(A_\delta^{(n)}(\rho;\mu)) = S_\mu(\rho).$$

More imprecisely, but concisely: when n is big and δ small,

$$\mu^{\otimes n}(A_\delta^{(n)}(\rho;\mu)) \approx e^{nS_\mu(\rho)}$$

and the joint density is $\approx e^{-nS_\mu(\rho)}$ on this set.

Example

$E_X = \mathbb{R}^n$, μ Lebesgue measure, φ automorphism given by matrix A; $\mu(\varphi(V)) = |\det A|\mu(V)$ and $S_\mu(\varphi_*\rho) = S_\mu(\rho) + \log|\det A|$, which is consistent with

$$\mu^{\otimes n}(A_\delta^{(n)}(\varphi_*\rho;\mu)) = \varphi^{\times n}(A_\delta^{(n)}(\rho;\mu)) = |\det A|^n\mu^{\otimes n}(A_\delta^{(n)}(\rho;\mu)). \quad (2)$$

Remarks

1. Certainty in the discrete case (E_X countable): $\rho = \delta_x$ implies $\mu^{\otimes n}(A_\delta^{(n)}(\rho;\mu)) = 1$ and hence $S_\mu(\delta_x) = 0$. Moreover, since for general ρ, $\mu^{\otimes n}(A_\delta^{(n)}) \geq 1$, the entropy $S_\mu(\rho)$ is positive.
2. Certainty in the continuous case: $\mu^{\otimes n}(A_\delta^{(n)}(\rho;\mu)) \to 0$ iff $S_\mu(\rho) \to -\infty$.
3. If μ proba, then $-S_\mu(\rho) = D(\rho||\mu)$, and $\mu^{\otimes n}(A_\delta^{(n)}(\rho;\mu)) \leq 1$ translates into $D(\rho||\mu) \geq 0$.

Outline

Disintegrations

Setting: $T : (E, \mathfrak{B}) \to (E_T, \mathfrak{B}_T)$ measurable map, ν a σ-finite measure on $(E, \mathfrak{B})$, and ξ a σ-finite measure on $(E_T, \mathfrak{B}_T)$.

Definition

The measure ν has a disintegration $\{\nu_t\}_{t \in E_T}$ with respect to T and ξ, or a (T, ξ)-**disintegration**, if

1. ν_t is a σ-finite measure on $\mathfrak{B}$ concentrated on $\{T = t\}$, which means that $\nu_t(T \neq t) = 0$ for ξ-almost every t;
2. for each measurable nonnegative function $f : E \to \mathbb{R}$,
    1. $t \mapsto \int_E f \, \mathrm{d}\nu_t$ is measurable,
    2. $\int_E f \, \mathrm{d}\nu = \int_{E_T} \left(\int_E f(x) \, \mathrm{d}\nu_t(x) \right) \mathrm{d}\xi(t)$.

Disintegrations

Setting: $T : (E, \mathfrak{B}) \to (E_T, \mathfrak{B}_T)$ measurable map, ν a σ-finite measure on $(E, \mathfrak{B})$, and ξ a σ-finite measure on $(E_T, \mathfrak{B}_T)$.

Definition

The measure ν has a disintegration $\{\nu_t\}_{t \in E_T}$ with respect to T and ξ, or a (T, ξ)-**disintegration**, if

1. ν_t is a σ-finite measure on $\mathfrak{B}$ concentrated on $\{T = t\}$, which means that $\nu_t(T \neq t) = 0$ for ξ-almost every t;
2. for each measurable nonnegative function $f : E \to \mathbb{R}$,
    1. $t \mapsto \int_E f \, \mathrm{d}\nu_t$ is measurable,
    2. $\int_E f \, \mathrm{d}\nu = \int_{E_T} \left(\int_E f(x) \, \mathrm{d}\nu_t(x) \right) \, \mathrm{d}\xi(t)$.

When ν probability, $x \in E \mapsto \int_E \chi_B(x) \, \mathrm{d}\nu_{T(x)}$ is a *regular version* of the conditional probability $\nu(B|\sigma(T))$ defined by Kolmogorov.

Example

$T : E \to E_T$ a surjection of finite sets, ν and ξ counting measures, with ν_t the counting measure restricted to $T^{-1}(t)$; the disintegration formula is simply $\sum_{x \in E} f(x) = \sum_{t \in E_T} \sum_{x \in T^{-1}(t)} f(x)$.

Example

$T : E \to E_T$ a surjection of finite sets, ν and ξ counting measures, with ν_t the counting measure restricted to $T^{-1}(t)$; the disintegration formula is simply $\sum_{x \in E} f(x) = \sum_{t \in E_T} \sum_{x \in T^{-1}(t)} f(x)$.

Example

$(E, \mathfrak{B}, \nu)$ product of two σ-finite measure spaces: $(E_S, \mathfrak{B}_S, \eta) \otimes (E_T, \mathfrak{B}_T, \xi)$, and T projection. Let ν_t be the image of η under the inclusion $s \mapsto (s, t)$. Then Fubini's theorem implies that ν_t is a (T, ξ)-disintegration of ν:

$$\int_{E_S \times E_T} f \, \mathrm{d}\nu = \int_{E_T} \left(\int_{E_S} f(s, t) \, \mathrm{d}\eta(s) \right) \mathrm{d}\xi(t) = \int_{E_T} \left(\int_E f \, \mathrm{d}\nu_t \right) \mathrm{d}\xi(t).$$

Proposition (Induced disintegration)

Let ν *have a* (T,ξ)*-disintegration* $\{\nu_t\}$ *and let* ρ *be absolutely continuous with respect to* ν *with finite density* $r(x)$*, with each* ν*,* ξ *and* ρ σ*-finite.*

1. *The measure* ρ *has a* (T,ξ)*-disintegration* $\{\tilde{\rho}_t\}$ *with* $\tilde{\rho}_t = r \cdot \nu_t$*.*

Proposition (Induced disintegration)

Let ν have a (T,ξ)-disintegration $\{\nu_t\}$ and let ρ be absolutely continuous with respect to ν with finite density $r(x)$, with each ν, ξ and ρ σ-finite.

1. *The measure ρ has a (T,ξ)-disintegration $\{\tilde{\rho}_t\}$ with $\tilde{\rho}_t = r \cdot \nu_t$.*
2. *$T_*\rho \ll \xi$, with density $\int_E r \, \mathrm{d}\nu_t$.*

Proposition (Induced disintegration)

Let ν have a (T,ξ)-disintegration $\{\nu_t\}$ and let ρ be absolutely continuous with respect to ν with finite density $r(x)$, with each ν, ξ and ρ σ-finite.

1. *The measure ρ has a (T,ξ)-disintegration $\{\tilde{\rho}_t\}$ with $\tilde{\rho}_t = r \cdot \nu_t$.*
2. $T_*\rho \ll \xi$, *with density* $\int_E r \, \mathrm{d}\nu_t$.
3. *If $T_*\rho$ is σ-finite then $0 < \nu_t r < \infty$ $T_*\nu$-almost surely, and the measures $\{\rho_t\}$ given by*

$$\rho_t = \frac{r}{\int_E r \, \mathrm{d}\nu_t} \cdot \nu_t$$

are probabilities that give a $(T, T_*\rho)$-disintegration of ρ.

Example: Discrete case

$T : E \to E_T$ a surjection of finite set, ν and ξ counting measures, with ν_t the counting measure restricted to $T^{-1}(t)$; the disintegration formula is simply $\sum_{x \in A \subset E} f(x) = \sum_{t \in T(A)} \sum_{x \in T^{-1}(t)} f(x)$.

Let ρ be a proba on E; its density is $d_\rho(x) = \rho(\{x\})$. There is a disintegration

$$\int_E \chi_A \,\mathrm{d}\rho = \rho(A) = \sum_{t \in E_T} \rho_t(A) d_{T_*\rho}(t) = \int_{E_T} \left(\int_E \chi_A \,\mathrm{d}\rho_t \right) \mathrm{d}T_*\rho.$$

which is the *theorem of total probability*. The conditional probabilities are the classical, $A \mapsto \rho_t(A) = \rho(A \cap T^{-1}(t))/\rho(T^{-1}(t))$, with density $d_\rho / d_{T_*\rho}(t)$ w.r.t. the counting measure on $T^{-1}(t)$.

Example: Product spaces

$(E, \mathfrak{B}, \nu)$ product of two σ-finite measure spaces: $(E_S, \mathfrak{B}_S, \eta) \otimes (E_T, \mathfrak{B}_T, \xi)$. Let ν_t be the image of η under the inclusion $s \mapsto (s, t)$. Then Fubini's theorem implies that ν_t is a (T, ξ)-disintegration of ν.

If $r(s, t)$ is the density of a probability ρ on $(E, \mathfrak{B})$, then $\rho_t \ll \nu_t$ with density $r(s, t)$, the value of t being fixed, and $\tilde{\rho}_t$ is a probability supported on $\{T = t\}$ with density $r(s, t) / \int_{E_S} r(s, t)\, \mathrm{d}\eta(s)$. Then,

$$
\begin{aligned}
\rho(A) &= \int_{E_T \times E_S} \chi_A r \, \mathrm{d}\nu \\
&= \int_{E_T} \left(\int_{E_S} \chi_A \frac{r}{\int_{E_S} r(s, t)\, \mathrm{d}\eta(s)} \, \mathrm{d}\eta \right) \int_{E_S} r(s, t)\, \mathrm{d}\eta(s)\, \mathrm{d}\xi(t) \\
&= \int_{E_T} \left(\int_E \chi_A \, \mathrm{d}\rho_t \right) \mathrm{d}T_* \rho(t).
\end{aligned}
$$

Outline

Chain rule

Setting: $T : (E_X, \mathfrak{B}_X) \to (E_Y, \mathfrak{B}_Y)$ measurable map between measurable spaces; μ (respectively ν) a σ-finite measure on $(E_X, \mathfrak{B}_X)$ (resp. $(E_Y, \mathfrak{B}_Y)$), and $\{\mu_y\}$ a (T, ν)-disintegration of μ.

Theorem (Chain rule for general disintegrations)

Any probability measure ρ absolutely continuous w.r.t. μ, with density r, has a $(T, T_*\rho)$-disintegration $\{\rho_y\}_{y \in Y}$ such that each ρ_y is a probability measure with density $r / \int_{E_X} r \,\mathrm{d}\mu_y$ w.r.t. μ_y, and the following chain rule holds:

$$S_\mu(\rho) = S_\nu(T_*\rho) + \int_{E_Y} S_{\mu_y}(\rho_y) \,\mathrm{d} T_*\rho(y). \tag{3}$$

Chain rule

Setting: $T : (E_X, \mathfrak{B}_X) \to (E_Y, \mathfrak{B}_Y)$ measurable map between measurable spaces; μ (respectively ν) a σ-finite measure on $(E_X, \mathfrak{B}_X)$ (resp. $(E_Y, \mathfrak{B}_Y)$), and $\{\mu_y\}$ a (T, ν)-disintegration of μ.

Theorem (Chain rule for general disintegrations)

Any probability measure ρ absolutely continuous w.r.t. μ, with density r, has a $(T, T_*\rho)$-disintegration $\{\rho_y\}_{y \in Y}$ such that each ρ_y is a probability measure with density $r / \int_{E_X} r \, \mathrm{d}\mu_y$ w.r.t. μ_y, and the following chain rule holds:

$$S_\mu(\rho) = S_\nu(T_*\rho) + \int_{E_Y} S_{\mu_y}(\rho_y) \, \mathrm{d} T_*\rho(y). \tag{3}$$

Proof.

$$S_\mu(\rho) = \mathbb{E}_\rho(-\log r) = \mathbb{E}_\rho(-\log \frac{r}{\int_{E_X} r \, \mathrm{d}\mu_y} - \log \int_{E_X} r \, \mathrm{d}\mu_y) = \dots$$

□

Example: Polar coordinates

Set $R : \mathbb{R}^2 \to [0, \infty[$, $x \mapsto ||x||$. The change-of-variables formula

$$\int_{\mathbb{R}^2} \varphi(x, y)\, \mathrm{d}x\, \mathrm{d}y = \int_{[0,\infty[} \left(\int_0^{2\pi} \varphi(r, \theta) r\, \mathrm{d}\theta \right) \mathrm{d}r, \qquad (4)$$

means that the Lebesgue measure $\lambda = \mathrm{d}x\, \mathrm{d}y$ on $\mathbb{R}^2$ has an $(R, \mathrm{d}r)$-disintegration into measures $\lambda_r = r\, \mathrm{d}\theta$ on the level sets $\{R = r\}$. If ρ has density f, its $(R, R_* \rho)$-disintegration are probabilities ρ_r with density $f / \int_0^{2\pi} f(r, \theta) r\, \mathrm{d}\theta$ w.r.t. λ_r, and we obtain an exact chain rule:

$$S_\lambda(\rho) = S_{dr}(R_* \rho) + \int_{[0,\infty[} S_{\lambda_r}(\rho_r)\, \mathrm{d}R_* \rho(r).$$

It has to be contrasted with a deformed chain rule in the literature, with an extra term $\mathbb{E}_\rho(\log R)$.

Interpretation of the conditional term

If $\{\mu_y\}_y$ is a (T, ν)-disintegration of μ, then $\{\mu_y^{\otimes n}\}_y$ is a $(T^{\times n}, \nu^{\otimes n})$-disintegration of $\mu^{\otimes n}$.

Proposition

Keeping the setting of the previous theorem,

$$\lim_{\delta \to 0} \lim_{n \to \infty} \frac{1}{n} \log \left(\frac{\int_{E_Y} \mu_y^{\otimes n}(A_\delta^{(n)}(\rho; \mu)) \, \mathrm{d}\nu^{\otimes n}(y)}{\nu^{\otimes n}(A_\delta^{(n)}(T_* \rho; \nu))} \right) = \int_{E_Y} S_{\mu_y}(\rho_y) \, \mathrm{d} T_* \rho(y).$$

Outline

Weil's formula

Given a locally compact topological group G, there exists a left invariant Haar measure that we denote λ^G.

Proposition (Weil's formula)

Let G be a locally compact group and H a closed normal subgroup of G. Given Haar measures on two groups among G, H and G/H, there is a Haar measure on the third one such that, for any integrable function $f : G \to \mathbb{R}$,

$$\int_G f(x) d\lambda^G(x) = \int_{G/H} \left(\int_H f(xy) \, \mathrm{d}\lambda^H(y) \right) \mathrm{d}\lambda^{G/H}(xH). \tag{5}$$

Let us denote by $\lambda^H_{[g]}$ the image of λ^H under the inclusion $\iota_g : H \to G$, $h \mapsto gH$. The proposition shows that $\{\lambda^H_{[g]}\}_{[g] \in G/H}$ is a $(T, \lambda^{G/H})$-disintegration of λ^G.

Proposition (Chain rule, Haar case)

Let G be a locally compact group, H a closed normal subgroup of G, and λ^G, λ^H, and $\lambda^{G/H}$ Haar measures in canonical relation. Let ρ be a probability measure on G. Denote by $T : G \to G/H$ the canonical projection. Then, there is T-disintegration $\{\rho_{[g]}\}_{[g] \in G/H}$ of ρ such that each $\rho_{[g]}$ is a probability measure, and

$$S_{\lambda^G}(\rho) = S_{\lambda^{G/H}}(\pi_* \rho) + \int_{G/H} S_{\lambda^H_{[g]}}(\rho_{[g]})\, \mathrm{d}\pi_* \rho([g]). \tag{6}$$

Conclusions and perspectives

1. There is a notion of differential entropy on any measure space, which has a probabilistic interpretation in terms of concentration of measure.
2. The entropy always depend on a choice of reference measure ("volume").
3. Several information inequalities have a "volumetric" interpretation (or at least an analogue).
4. Disintegration of measures give regular versions of conditional probabilities in topological situations (e.g. Radon measures).
5. Every disintegration of reference measures induces a chain rule for the corresponding entropies.
6. In particular, the results apply to Haar measures in canonical relation. Information inequalities in this group case? Chain rule for mutual information and higher analogues?

Coarea formula

Let $f : \mathbb{R}^M \to \mathbb{R}^k$ be a Lipschitz function and let E be a countably $\mathscr{H}^N$-rectifiable subset of $\mathbb{R}^M$.

The function $t \mapsto \mathscr{H}^{N-k}(E \cap f^{-1}(t))$ is $\mathscr{L}^k$-measurable, $E \cap f^{-1}(t)$ is countably $\mathscr{H}^{N-k}$-rectifiable for $\mathscr{L}^k$-almost every t, and

$$\int_E g(x)\sqrt{\det(d^E f_x \circ d^E f_x^*)}\,\mathrm{d}\mathscr{H}^N(x) = \int_{\mathbb{R}^k}\left(\int_{E\cap\{f=t\}} g(y)\,\mathrm{d}\mathscr{H}^{N-k}(y)\right)\,\mathrm{d}t.$$

When $f : \mathbb{R}^2 \to R$ maps (x, y) to $\sqrt{x^2+y^2}$ and $E = \mathbb{R}^2$, then $d^E f_{(x,y)} = \left[x/\sqrt{x^2+y^2} \quad y/\sqrt{x^2+y^2}\right]$, $\sqrt{\det(d^E f_x \circ d^E f_x^*)} = 1$, and H^{N-k} is $r\,\mathrm{d}\theta$.