Variance Calculation through Large Deviation Techniques

Arnaud de la Fortelle

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Variance Calculation through Large Deviation Techniques

Arnaud de La Fortelle

N° 4441

April 2002

THÈME 1

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--4441--FR+ENG

Variance Calculation through Large Deviation Techniques

Arnaud de La Fortelle*

Thème 1 — Réseaux et systèmes
Projet PREVAL

Rapport de recherche n° 4441 — April 2002 — 40 pages

Abstract: In this paper, we show how to use the expression of entropy in order to calculate variances. Three cases are analyzed: independent and identically distributed (iid) variables, Markov chains (in discrete time) and jump Markov processes (in continuous time). This framework is valid far beyond these case studies, e.g. in transportation and telecommunication networks and likely in all models where the entropy is explicit.
The method allows to derive the variance from the entropy function, which is a classical quantity in large deviations. Moreover, the entropy has often a rather simple expression (e.g. for networks). Here we show a closed formula expressing the variance in terms of derivatives of the entropy; by-products are also obtained, such as martingales, used in the proof of the central limit theorem. These results might be a good starting point for further developments, e.g. calculation of exact asymptotics (instead of logarithmic ones in large deviation) or solving entropy minimization problems.

Key-words: Variance, central limit theorem, large deviations, empirical measure, entropy.

* Postal address: INRIA - Domaine de Voluceau, Rocquencourt - BP 105 - 78153 Le Chesnay Cedex - FRANCE. e-mail address: `Arnaud.De_La_Fortelle@inria.fr`. This research is supported by the French Ministère de l'Équipement.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France)
Téléphone : +33 1 39 63 55 11 — Télécopie : +33 1 39 63 53 30

Calcul de la variance
par des techniques de grandes déviations

Résumé : Dans cet article, nous montrons comment utiliser l'expression de l'entropie pour calculer des variances. Cette démarche est présentée sur trois cas d'école : les variables indépendantes et identiquement distribuées (i.i.d.), les chaînes de Markov (en temps discret) et les processus de Markov (en temps continu). Cette démarche s'étend aisément à d'autres processus, par exemple les réseaux de transport ou de télécommunications.
La méthode consiste à utiliser l'entropie associée au processus; cette grandeur, classique en grandes déviations, est explicite et relativement simple dans de nombreux cas, notamment les réseaux. On montre ici comment les dérivées de l'entropie permettent d'exprimer la variance, mais aussi d'autres quantités d'intérêt, comme certaines martingales qui sont utilisées pour prouver le théorème central limite correspondant à cette variance. Ces résultats ouvrent des perspectives comme le raffinement des bornes de grandes déviations (obtenir des comportements asymptotiques exacts) ou la résolution exacte de problèmes de minimisation d'entropie.

Mots-clés : Variance, théorème central limite, grandes déviations, mesure empirique, entropie.

1 The i.i.d. framework

In a first step we would like to show which tool have to be set in order to calculate the variance in the i.i.d. case. It could seem useless to analyze again such a well-known case, particularly with these complicated tools, but it turns out that it allows then to produce better large deviations bounds and that this scheme produces a way for calculating variance formulae in much more difficult cases, especially in the Markov case or in continuous time. It also exhibits some particular features of the i.i.d. framework due to special properties.

1.1 Notation

The framework is the following: (E, d) is a Polish space, μ is a probability on E. We shall denote by $\mathcal{M}(\mu)$ the set of finite measures on E absolutely continuous w.r.t. μ and by $\mathcal{M}_1(\mu)$ the subset of probability measures of $\mathcal{M}(\mu)$. For all $\nu \in \mathcal{M}_1(\mu)$, its Radon-Nikodym derivative is denoted by $\mathrm{d}\nu/\mathrm{d}\mu$. More generally, for such a measure ν we shall define the sets:

- $L_1(\nu)$, the real ν–integrable functions;
- $\mathcal{M}(\nu)$, the finite signed measures absolutely continuous w.r.t. ν;
- $L_\infty(\nu) \subset L_1(\nu)$, the real ν–measurable a.s. bounded functions[1];
- $\mathcal{M}_b(\nu) \subset \mathcal{M}(\nu)$, the measures η such that $\mathrm{d}\eta/\mathrm{d}\nu$ is ν-a.s. bounded;
- $\mathcal{M}_0(\nu) \subset \mathcal{M}(\nu)$, the measures, usually denoted by η, ζ, with $\int \mathrm{d}\eta = 0$;
- $\mathcal{F}_0(\nu) \subset L_1(\nu)$, the real ν–integrable functions, usually denoted by ϕ or ψ, verifying $\int \phi \, \mathrm{d}\nu = 0$.

When no confusion arises, we shall drop the indication of the reference probability measure ν. Note that there are classical bijections[2] $\mathcal{M} \leftrightarrow L_1$, $\mathcal{M}_b \leftrightarrow L_\infty$ or $\mathcal{M}_0 \leftrightarrow \mathcal{F}_0$ expressed by:

$$\eta \in \mathcal{M}(\nu) \longleftrightarrow \phi \in \mathcal{F}(\nu), \qquad \text{with} \quad \phi = \frac{\mathrm{d}\eta}{\mathrm{d}\nu} \ \nu\text{-a.s.} \quad \text{or} \quad \eta = \phi\nu.$$

[1] The inclusion $L_\infty(\nu) \subset L_1(\nu)$ is true here since ν is a *probability* measure.

[2] We shall denote by $A \leftrightarrow B$ when the sets A and B are homeomorphic.

1.2 Topology and quotient spaces

Topology The space $L_1 \leftrightarrow \mathcal{M}$ is canonically equipped with the L_1 norm, which is the total variation norm on $\mathcal{M}$. The induced topology is denoted in both cases by $\sigma(L_1)$. The bijection between $\eta \in \mathcal{M}$ and its image $\phi \in L_1$ is an isometry:

$$\|\eta\|_{\mathrm{vt}} = \|\phi\|_{L_1} \qquad \text{when} \quad \eta = \phi\nu. \tag{1.1}$$

The space $L_\infty \leftrightarrow \mathcal{M}_b$ is canonically equipped with the L_∞[3] norm, defined by

$$\|f\|_\infty \stackrel{\text{def}}{=} \inf\{x \in \mathbb{R} : \ \nu(|f| \le x) = 0\}.$$

The set $\mathcal{M}_b$ is equipped with a similar norm, also making the bijection an isometry. The induced topology is denoted by $\sigma(L_\infty)$.
The Banach spaces $\mathcal{M}$ and L_∞ (or L_1 and $\mathcal{M}_b$) are identified as the dual of each other, so that theses spaces are "reflexive[4]". This duality is equivalent to the duality induced by the bilinear form

$$(\eta, f) \to \langle \eta, f\rangle \stackrel{\text{def}}{=} \int f \,\mathrm{d}\nu.$$

Moreover

$$\|\eta\|_{\mathrm{vt}} = |\!|\!|\eta|\!|\!|_{L_\infty^*} \stackrel{\text{def}}{=} \sup_{\|f\|_\infty \le 1} \langle \eta, f\rangle, \tag{1.2}$$

$$\|f\|_\infty = |\!|\!|f|\!|\!|_{L_1^*} \stackrel{\text{def}}{=} \sup_{\|\eta\|_{\mathrm{vt}} \le 1} \langle \eta, f\rangle. \tag{1.3}$$

Quotient spaces Let $N \subset L_1$ be the (closed) vector space generated by $\mathbb{I} : E \mapsto \mathbb{R}$. It is defined so that $\mathcal{M}_0 = N^o$ is the orthogonal of N in $\mathcal{M}$. When confusion may arise, we shall mark the dependence to ν by N_ν.

Topology on quotient spaces Now, $\mathcal{M}_0 = N^o$ is equipped with the total variation norm and the induced topology. The space L_∞/N is equipped with the norm derived from the L_∞ norm: denoting by θ the canonical application from L_∞ onto L_∞/N, for all $\dot{f} \in L_\infty/N$,

$$\|\dot{f}\|_\infty \stackrel{\text{def}}{=} \inf_{\theta(f)=\dot{f}} \|f\|_\infty. \tag{1.4}$$

[3] The canonical construction defines the almost sure equality $f = g$ ν-a.e. and L_∞ is actually the quotient of L_∞ by this equivalence relation. Then, the topology is separable. See, e.g. [5, p. 162].

[4] Using the terminology of Bourbaki [1, IV 5].

This norm defines a topology which is equivalent to the quotient topology of $\sigma(L_\infty)$ by N. The important result now is that L_∞ and $\mathcal{M}$ are reflexive Banach spaces and N is *closed*, so that L_∞/N and $N^o = \mathcal{M}_0$ are also Banach reflexive space *dual of each other.*
This means two things. First, any continuous linear functional g on $\mathcal{M}_0$ can be uniquely represented as an element of L_∞/N; second, and more important, for $f, g \in L_\infty$

$$\langle f, \eta\rangle = \langle g, \eta\rangle, \quad \forall \eta \in \mathcal{M}_0 \implies \exists h \in N: \ f = g + h, \ \nu - \text{a.s.} \tag{1.5}$$

Moreover, we get the identities

$$\|\eta\|_1 = \sup_{\dot{f} \in L_\infty/N} \langle \eta, \dot{f}\rangle, \qquad \forall \eta \in \mathcal{M}_0 \tag{1.6}$$

$$\|\dot{f}\|_\infty = \sup_{\eta \in \mathcal{M}_0} \langle \eta, \dot{f}\rangle = \inf_{\theta(f) = \dot{f}} \|f\|_\infty, \qquad \forall \dot{f} \in L_\infty/N. \tag{1.7}$$

The previous construction is the most natural one, since we obtained $\mathcal{M}_0$. In the sequel, however, we shall use the set $\mathcal{M}_0 \cap \mathcal{M}_b$ (or equivalently $\mathcal{F}_0 \cap L_\infty$). The same construction holds when L_1 and L_∞ (or $\mathcal{M}$ and $\mathcal{M}_b$) are permuted. $N \in L_1$ and $\mathcal{M}_0 \cap \mathcal{M}_b = N^o$ so that $\mathcal{M}_0 \cap \mathcal{M}_b$ and L_1/N are Banach reflexive spaces dual of each other. The relations (1.5)–(1.7) also hold, with L_1 and L_∞ permuted.

Further results One extra feature of the i.i.d. case (by comparison to the Markov case) is to be able to identify the set L_1/N to $\mathcal{F}_0$ by the following identity.

$$\dot{f} \in L_1/N \leftrightarrow f - \langle f, \nu\rangle \in \mathcal{F}_0. \tag{1.8}$$

It is easily checked that (1.8) is a one-to-one mapping. By the same identity L_∞/N is identified to $\mathcal{F}_0 \cap L_\infty$.

1.3 Tangent space

The construction of the derivatives of the entropy function is rather simple when $\mathcal{M}_1(\mu)$ is considered, not as an affine space, but as a configuration space, in which there is a tangent space at each point. The following construction is a transposition of this theory into our framework.

Definition 1.1 *Let $\nu \in \mathcal{M}_1(\mu)$, and ξ an application from an open neighborhood $O \subset \mathbb{R}$ of 0 onto $\mathcal{M}_1(\mu)$, differentiable*[5] *at 0, with $\xi(0) = \nu$. By definition $\xi'(0)$ is a tangent vector at ν. The tangent space $\mathcal{T}(\nu)$ is the set of such tangent vectors.*

$$\eta \in \mathcal{T}(\nu) \iff \exists \xi : \ O \mapsto \mathcal{M}_1(\mu) \text{ such that } \eta = \xi'(0). \tag{1.9}$$

Proposition 1.2 *Let $\nu \in \mathcal{M}_1(\mu)$. The vector space $\mathcal{T}(\nu) = \mathcal{M}_0(\nu) \cap \mathcal{M}_b(\nu)$ is isomorphic to $\mathcal{F}_0(\nu) \cap L_\infty(\nu)$:*

$$\eta \in \mathcal{T}(\nu) \iff \frac{\mathrm{d}\eta}{\mathrm{d}\nu} \in \mathcal{F}_0(\nu) \cap L_\infty(\nu)$$

Proof : Let $\phi \in \mathcal{F}_0(\nu)$ be bounded: there exists $\varepsilon > 0$ such that $1 + x\phi \geq 0$ for all $|x| < \varepsilon$. Hence it is possible to define $\xi(x) \stackrel{\text{def}}{=} (1 + x\phi)\nu$. Since $\phi \in \mathcal{F}_0(\nu)$, $\xi(x) \in M_1(E)$; ξ is affine, thus differentiable and $\xi' = \phi\nu$.
Reciprocally let $\eta \in \mathcal{T}(\nu)$, and a corresponding function ξ. By definition, for any Borel subset A of E, one can write

$$\nu(A) + x\eta(A) + R_x(A) \geq 0, \tag{1.10}$$

since $\xi(x) \stackrel{\text{def}}{=} \nu + x\eta + R_x \in \mathcal{M}_1(\mu)$. The rest R_x is small, which means that, for all $\varepsilon > 0$, there exists $u > 0$ such that $\|R_x\| \leq \varepsilon x$ for all $|x| \leq u$. Since $|\xi(x)| = |\nu| = 1$, one derives that $x\eta + R_x \in \mathcal{M}_0(E)$, and since $R_x = o(x)$, $|\eta| \in \mathcal{M}_0(E)$.
If $\nu(A) = 0$, (1.10) implies that $\eta(A)$ has the same sign as x, so that $\eta(A) = 0$. Hence $\eta \ll \nu$. If $\nu(A) > 0$, the inequality (1.10) implies that $\nu(A) + |x|\eta(A) + x\nu(A) \geq 0$ (by taking $\varepsilon = \nu(A)$), and

$$|\eta(A)| \leq \frac{\nu(A) + |u|\nu(A)}{|u|} \leq \frac{1 + |u|}{|u|}\nu(A),$$

by taking $x = \pm u$ (with u corresponding to $\varepsilon = \nu(A)$). Hence $\phi \stackrel{\text{def}}{=} \mathrm{d}\eta/\,\mathrm{d}\nu$ is bounded ν–a.s: there exists a bounded version of ϕ. The proposition is proved since $|\eta| \in \mathcal{M}_0(E)$ and $\eta \ll \nu$ is equivalent to $\phi \in \mathcal{F}_0(\nu)$. ■

The tangent space $\mathcal{T}(\nu) \subset \mathcal{M}_b(\nu)$ is canonically equipped with the L_∞ norm and the induced topology. Its dual is L_1/N.

[5]The application ξ is considered as an application on $\mathcal{M}(\mu)$ equipped with the L_1 norm for the notion of derivative. Hence $\xi'(0) \in \mathcal{M}(\mu)$.

Lemma 1.3 *Let* $\nu, \pi \in \mathcal{M}_1(\mu)$.

$$\nu \in \mathcal{M}(\pi) \implies L_\infty(\pi) \subset L_\infty(\nu),$$
$$\nu \in \mathcal{M}_b(\pi) \implies \mathcal{T}(\pi) \subset \mathcal{T}(\nu).$$

Proof : Let $\nu \in \mathcal{M}(\pi)$ which implies $\nu \ll \pi$. By definition, for all Borel set A, $\pi(A) = 0$ implies that $\nu(A) = 0$. Now, if $f \in L_\infty(\pi)$, there exist $c < \infty$ such that

$$\pi\big(\{x \in E : \ |f(x)| > c\}\big) = 0 \quad \text{hence} \quad \nu\big(\{x \in E : \ |f(x)| > c\}\big) = 0,$$

so that $f \in L_\infty(\nu)$. Assume additionally $\mathrm{d}\nu/\,\mathrm{d}\pi$ to be bounded. Let $\eta \in \mathcal{T}(\pi)$, then $\int \mathrm{d}\eta = 0$ and

$$\frac{\mathrm{d}\eta}{\mathrm{d}\pi} \in L_\infty(\pi) \implies \frac{\mathrm{d}\eta}{\mathrm{d}\nu} = \frac{\mathrm{d}\nu}{\mathrm{d}\pi}\frac{\mathrm{d}\eta}{\mathrm{d}\nu} \in L_\infty(\nu),$$

hence $\mathcal{T}(\pi) \subset \mathcal{T}(\nu)$. ■

1.4 Derivatives

The entropy function is $H(.\|\mu)$ is defined by

$$H(\nu\|\mu) \stackrel{\text{def}}{=} \begin{cases} \displaystyle\int \log\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\nu & \forall \nu \in \mathcal{M}_1(\mu), \\ +\infty & \forall \nu \notin \mathcal{M}_1(\mu). \end{cases} \tag{1.11}$$

The entropy function is not continuous and not even finite everywhere, so that it may happen, e.g. that $H \circ \xi$ is always infinite except for $x = 0$. We would like however to define the directional derivative of the entropy function, if it exists, by

$$\nabla H(\nu\|\mu).\eta \stackrel{\text{def}}{=} \frac{\mathrm{d}}{\mathrm{d}x} H \circ \xi(0) \qquad \text{where } \eta = \xi'(0). \tag{1.12}$$

In order to avoid those trajectories for which H is not continuous, we shall only consider affine paths of the form $\xi(x) = \nu + x\eta$ in the equation (1.12). The proof of Proposition 1.2 shows this is always possible.

Definition 1.4 *The gradient* $\nabla H(\nu\|\mu)$ *is defined at* $\nu \in \mathcal{M}_1(\mu)$ *if the directional derivatives* $\nabla H(\nu\|\mu).\eta$ *are defined for all* $\eta \in \mathcal{T}(\nu)$.

Proposition 1.5 *The gradient* $\nabla H(\nu\|\mu)$ *is a linear functional* $\mathcal{T}(\nu) \mapsto \mathbb{R}$ *defined at every points* ν *where the entropy* $H(\nu\|\mu)$ *is finite by:*

$$\nabla H(\nu\|\mu).\eta = \int \log\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\eta, \quad \forall \eta \in \mathcal{T}(\nu). \tag{1.13}$$

Proof : First note that, when $H(\nu\|\mu)$ is finite, $\log(\mathrm{d}\nu/\mathrm{d}\mu)$ is absolutely integrable w.r.t. ν. Moreover $\phi = \mathrm{d}\eta/\mathrm{d}\nu$ is a.s. bounded, by definition of η, hence

$$\int \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\eta = \int \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\phi\,\mathrm{d}\nu$$

is well defined and finite for all $\eta \in \mathcal{T}(\nu)$.
Let $\nu \in \mathcal{M}_1(\mu)$ with finite entropy, $\eta \in \mathcal{T}(\nu)$ and $\phi \stackrel{\text{def}}{=} \mathrm{d}\eta/\mathrm{d}\nu$; $\xi(x) = \nu + x\eta$ is defined for $|x| < u$, with $u > 0$. Since

$$\phi(x) \stackrel{\text{def}}{=} \frac{\mathrm{d}\xi(x)}{\mathrm{d}\nu} = 1 + x\phi$$

is uniformly bounded on $(-u, u)$, the entropy $H(\xi(x)\|\mu)$ is finite on this interval and

$$H\big(\xi(x)\|\mu\big) - H(\nu\|\mu) = x\int \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\eta + \int \log(1 + x\phi)\,\mathrm{d}\xi(x). \tag{1.14}$$

The last term tends to zero faster than x, therefore the gradient is well defined. It is linear as the integral. ■

Corollary 1.6 *When the gradient is defined, it is a continuous functional $\mathcal{T}(\nu) \mapsto \mathbb{R}$. So there exists a unique canonical representation of $\nabla H(\nu\|\mu)$ in $L_1(\nu)/N_\nu$ that will also be denoted by $\nabla H(\nu\|\mu)$. Moreover*

$$f \equiv \nabla H(\nu\|\mu) \pmod{N_\nu} \iff \exists \alpha \in \mathbb{R} :\ f = \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu} + \alpha \quad \nu\text{-a.e.} \tag{1.15}$$

Proof : It is a straight sequel of the topological prelude and of (1.5). ■

> **Remark :** It is worth pointing out that $\nabla H(\nu\|\mu)$ is not necessarily continuous for the usual weak topology; moreover, it is continuous for a topology that varies with ν! This is a stumbling block in usual treatment of differential. On the other hand, here is a definition of gradient for a non-continuous function! It is then natural to possess weak properties, but still enough for the purpose of optimization.

♦♦♦

Second order derivatives The second-order directional derivatives are defined, with the same restrictions as ∇H, by

$$\nabla^2 H(\nu\|\mu).\eta.\zeta \stackrel{\text{def}}{=} \frac{\mathrm{d}}{\mathrm{d}x}\Big[\nabla H\big(\xi(x)\|\mu\big).\eta\Big](0) \qquad \text{where } \zeta = \xi'(0). \tag{1.16}$$

Definition 1.7 *The Hessian $\nabla^2 H(\nu\|\mu)$ is defined at $\nu \in \mathcal{M}_1(\mu)$ if the directional derivatives $\nabla^2 H(\nu\|\mu).\eta.\zeta$ are defined for all $\eta, \zeta \in \mathcal{T}(\nu)$.*

Proposition 1.8 *The Hessian is the bilinear symmetric positive definite functional $\mathcal{T}(\nu) \times \mathcal{T}(\nu) \mapsto \mathbb{R}$ defined at every points ν where the entropy $H(\nu\|P)$ is finite by:*

$$\nabla^2 H(\nu\|\mu).\eta.\zeta = \int \frac{\mathrm{d}\eta}{\mathrm{d}\nu}\frac{\mathrm{d}\zeta}{\mathrm{d}\nu}\,\mathrm{d}\nu, \quad \forall \eta, \zeta \in \mathcal{T}(\nu). \tag{1.17}$$

Proof : The Radon-Nikodym derivatives will be denoted by ϕ for $\mathrm{d}\eta/\mathrm{d}\nu$ and ψ for $\mathrm{d}\zeta/\mathrm{d}\nu$. Since η, ζ belong to $\mathcal{T}(\nu)$, ϕ, ψ are well defined and bounded, hence belong to $L_2(\nu)$, and the Hessian is the canonical scalar product:

$$\nabla^2 H(\nu\|\mu).\eta.\zeta = \int \phi\psi\,\mathrm{d}\nu = \langle \phi, \psi \rangle.$$

In the same way as for the gradient, there exists an open neighborhood $(-u, u)$ of 0 such that $\xi(x) \stackrel{\text{def}}{=} \nu + x\zeta$ belongs to $\mathcal{M}_1(\mu)$. Since $\zeta \ll \nu$, the support of $\xi(x)$ is exactly the same as ν. Moreover the Radon-Nikodym derivatives of ζ w.r.t. ν is a.s. bounded, therefore $\mathcal{T}\big(\xi(x)\big) = \mathcal{T}(\nu)$ and the right-hand side of (1.16) makes sense. Now

$$\nabla H\big(\xi(x)\|\mu\big).\eta - \nabla H(\nu\|\mu).\eta = \int \log(1 + x\psi)\,\mathrm{d}\eta = x\int \psi\phi\,\mathrm{d}\nu + o(x),$$

thus the Hessian exists and is equal to the canonical scalar product on $L_2(\nu)$ which includes $\mathcal{T}(\nu)$. This also proves that the Hessian is bilinear, symmetric and positive definite. ■

Remark : Note that the derivatives depends on ν and μ, but the tangent space does not depend on μ. When no confusion arises, we shall drop μ in $\nabla H(\nu\|\mu)$. For the second derivative it is even better since $\nabla^2 H(\nu\|\mu)$ *does not depend on* μ. As for the spaces L_1, M_1 ... we shall drop the mention of ν in $\mathcal{T}(\nu)$ when no confusion arises.

♦♦♦

Proposition 1.9 *When ν has finite entropy, the Hessian $\nabla^2 H(\nu)$, considered as a linear functional from $\mathcal{T}$ onto $\mathcal{T}^* = L_1/N$ is defined as*

$$\nabla^2 H(\nu).\eta \equiv \frac{\mathrm{d}\eta}{\mathrm{d}\nu} \pmod N \qquad \forall \eta \in \mathcal{T}. \tag{1.18}$$

The Hessian is injective and its inverse, when it exists, is given by

$$\nabla^2 H(\nu)^{-1}.\dot{f} = \big(f - \langle f, \nu\rangle\big)\nu \tag{1.19}$$

for any representative f of $\dot{f} \in L_1/N$.

Proof : It is a straightforward consequence of Proposition 1.8. ■

Corollary 1.10 *When the Hessian is defined, it is the canonical bijection* Id *between $\mathcal{T}(\nu)$ and $\mathcal{F}_0(\nu) \cap L_\infty(\nu)$. Hence the Hessian is invertible and its inverse $\dot{f} \to \nabla^2 H(\nu)^{-1}.\dot{f}$ is continuous.*

Proof : By Proposition 1.9, the image of the operator $\nabla^2 H(\nu) : \mathcal{T} \mapsto L_1/N$ is exactly L_∞/N. By (1.8), this set is isomorphic to $\mathcal{F}_0 \cap L_\infty \leftrightarrow \mathcal{T}$. Hence the Hessian is an isomorphism, therefore continuous, invertible with its inverse continuous. ■

1.5 Local expansion of the entropy function

The entropy function is not continuous, but the existence of continuous derivatives in a restrained tangent space shows that one can expect a kind of approximation formulæ where the derivatives exist.
A very nice identity verified by the entropy will help us in the sequel. It is the same as (1.14). When the terms are defined

$$H(\nu + \eta \| \mu) = H(\nu\|\mu) + \nabla H(\nu\|\mu).\eta + H(\nu + \eta\|\nu). \tag{1.20}$$

Proposition 1.11 *Let $\nu \in \mathcal{M}_1(\mu)$ have finite entropy. For all $\eta \in \mathcal{T}(\nu)$ such that $\|\eta\|_\infty \le 1/2$ the entropy of $\nu + \eta$ is finite and:*

$$\left| H(\nu + \eta\|\mu) - H(\nu\|\mu) - \nabla H(\nu\|\mu).\eta - \frac{1}{2}\nabla^2 H(\nu).\eta.\eta \right| \le \|\eta\|_\infty^3, \tag{1.21}$$

$$\big\|\nabla H(\nu + \eta\|\mu) - \nabla H(\nu\|\mu) - \nabla^2 H(\nu).\eta\big\|_1 \le 2\|\eta\|_\infty^2. \tag{1.22}$$

Proof : Note that $\mathcal{T}(\nu)$ is canonically endowed with the L_∞ norm. If $\|\eta\|_\infty \le 1/2$, then $\nu+\eta$ is a probability measure, and all terms in (1.20) are well defined. By (1.20), the only term to bound is $H(\nu+\eta\|\nu)$. Using

$$\frac{x^3}{3(1+x)} \le \log(1+x) - x + \frac{x^2}{2} \le \frac{x^3}{3}, \qquad \forall x > -1,$$

and demoting $\mathrm{d}\eta/\mathrm{d}\nu = \phi$, one obtains the bound

$$\frac{-\|\eta\|_\infty^3}{3(1-\|\eta\|_\infty)} \le H(\nu+\eta\|\nu) + \int \left(-\phi + \frac{\phi^2}{2}\right) \mathrm{d}\nu \le \frac{\|\eta\|_\infty^3}{3}. \tag{1.23}$$

Now,

$$\int \phi^2 \,\mathrm{d}\nu = \nabla^2 H(\nu).\eta.\eta \qquad \text{and} \qquad \int \phi \,\mathrm{d}\nu = 0,$$

therefore, combining the bounds (1.23) with (1.20), the uniform expansion (1.21) is obtained. The bound (1.22) is obtained in a similar way by

$$\left|\log(1+x) - x\right| \le \frac{x^2}{1+x}, \qquad \forall x > -1,$$

which implies that

$$\left|\nabla H(\nu+\eta\|\mu).\zeta - \nabla H(\nu\|\mu).\zeta - \nabla^2 H(\nu).\eta.\zeta\right| \le 2\|\eta\|_\infty^2 \|\zeta\|_\infty, \quad \forall \zeta \in \mathcal{T}(\nu).$$

Note that the norm in $\mathcal{T}(\nu)^* \leftrightarrow L_1/N$ is defined by the L_1 version[6] of (1.4). The bound (1.22) follows by an adapted version of (1.2). ∎

Corollary 1.12 *The entropy function is continuous on the set of probability measures ν such that $\mathrm{d}\nu/\mathrm{d}\mu$ is bounded from 0 and from ∞.*

1.6 Optimization with constraints

The topic of this section is to study existence, uniqueness and characterization of the following optimization problem (O1).

Problem (O1) *Let $f : E \mapsto \mathbb{R}^d$ be a vector valued function with coordinates $f_1, \ldots, f_d$ belonging to $L_\infty(\mu)$. Then the optimization problem is*

$$\inf\Big\{ H(\nu\|\mu) \ \ \textit{where} \ \ \langle f, \nu\rangle = z \Big\} \tag{1.24}$$

[6]It is also the L_1 norm for the canonical representative in $\mathcal{F}_0$ defined by (1.8).

The solution of (O1) relies on two hypotheses. The first one is a "structural" hypothese while the second is a condition for the problem not to be ill-defined due to bad values. Assumption (H1) is easily checked to be true in the i.i.d. framework.

Assumption (H1) *For all $\nu \in \mathcal{M}_1(\mu)$, the set $\mathcal{M}_b(\nu) \cap \mathcal{M}_1(\nu)$ is dense in $\mathcal{M}_1(\nu)$.*

Assumption (H2) *$z \in \mathbb{R}^d$ belongs to the interior of $f\big(\mathcal{M}_1(\mu)\big)$.*

Lemma 1.13 *Under Assumptions (H1) and (H2), there exists a measure $\nu \in \mathcal{M}_1(\mu)$ such that*

$$\langle f, \nu \rangle = z \qquad \text{and} \qquad H(\nu \| \mu) < \infty.$$

Proof : Let $\nu \in \mathcal{M}_1(\mu)$. Since $\nu \ll \mu$ and f is μ-a.s. bounded, so f is ν-a.s. bounded hence ν-integrable: $\langle f, \nu \rangle$ is finite for all $\nu \in \mathcal{M}_1(\mu)$.
Take $\varepsilon > 0$ small enough. By (H2), for all y in the ball $B(z, \varepsilon)$ of center z and radius ε, there exists $\nu \in \mathcal{M}_1(\mu)$ such that $\langle f, \nu \rangle = y$. Therefore one can choose $y_0 \ldots y_d \in B(z, \varepsilon)$ such that their convex hull has non-empty interior and contains z. The corresponding measures are denoted by $\nu_0 \ldots \nu_d$.
By (H1), one can find $\nu_i^b \in \mathcal{M}_b(\mu)$ arbitrarily close to ν_i. They are chosen so that the image by f of their convex hull has non-empty interior and contains z.
Now it is easyly checked that the entropy $H(.\|\mu)$ is finite on $\mathcal{M}_b(\mu)$. Since the entropy function is convex, all convex combinations of $\nu_0^b, \ldots, \nu_d^b$ have finite entropy, among which the combination which yields z. The lemma is proved. ■

Corollary 1.14 (Existence) *Under Assumptions (H1) and (H2), there exists a solution $\nu^* \in \mathcal{M}_1(\mu)$ to the optimization problem (O1).*

Proof : It is a classical property (see, e.g. [3]) that the level sets $H^{-1}([0, K])$ are compact, for all $K \geq 0$. Moreover, by Lemma 1.13, the set $\{\langle f, \nu \rangle = z\} \cap H^{-1}([0, \infty))$ is not empty, therefore there exists a solution to the optimization problem (O1). ■

The preceding statements has been separated because they are specific to the i.i.d. framework: e.g. the *existence* of a solution is a hard point in the Markov framework. Now, the tools developped for analysing the entropy function allow to characterize a possible solution as in Theorem 1.15.

Theorem 1.15 *Assume (H1) and (H2) are valid and optimization problem (O1) possesses one solution ν^*. Then this solution has finite entropy, is unique, ν^* and μ*

are equivalent (i.e. mutually absolutely continuous) and there exists $\alpha_1, \dots, \alpha_d$ *such that*

$$\nabla H(\nu^*\|\mu) \equiv \sum_{k=1}^{d} \alpha_k f_k \pmod{N_{\nu^*}}. \tag{1.25}$$

Proof : The solution has finite entropy by Lemma 1.13. It is unique since the entropy is strictly convex.
Since the entropy function (see (1.11)) is infinite outside $\mathcal{M}_1(\mu)$, $\nu^* \ll \mu$. In order to prove the reverse relation, suppose this is not true: there exists a set $A \subset E$ such that

$$\mu(A) > 0 \qquad \text{and} \qquad \nu^*(A) = 0.$$

Denote $w \stackrel{\text{def}}{=} \langle f, \mu \rangle$. For $u > 0$ small enough, $(z - uw)/(1-u)$ belongs to the convex hull of $\langle f, \nu_0^b \rangle, \dots, \langle f, \nu_d^b \rangle$ constructed in Lemma 1.13. Hence there exists $\nu^b \in \mathcal{M}_b(\mu) \cap \mathcal{M}_1$ such that

$$\langle f, \nu^b \rangle = \frac{z - uw}{1 - u} \iff \langle f, (1-u)\nu^b + u\mu \rangle = z.$$

Define $\pi \stackrel{\text{def}}{=} (1-u)\nu^b + u\mu$. Since $\pi \in \mathcal{M}_b(\mu) \cap \mathcal{M}_1$, its entropy is finite and by convexity the entropy of $(1-x)\nu^* + x\pi$ is finite for all $x \in [0,1]$.
The convexity of $a \to a\log(a/b)$ used on A^c yields

$$\begin{aligned}
& H\big((1-x)\nu^* + x\pi \big\| \mu\big) \\
\leq \quad & (1-x)\int_{A^c} \mathrm{d}\nu^* \log\frac{\mathrm{d}\nu^*}{\mathrm{d}\mu} + x\int_{A^c} \mathrm{d}\pi \log\frac{\mathrm{d}\pi}{\mathrm{d}\mu} + x\int_A \mathrm{d}(x\pi)\log\frac{\mathrm{d}(x\pi)}{\mathrm{d}\mu} \\
= \quad & H(\nu^*\|\mu) + \pi(A)x\log x + x\big(H(\pi\|\mu) - H(\nu^*\|\mu)\big).
\end{aligned}$$

Since $\pi(A) > u\mu(A) > 0$, for small x, the right side is strictly less than $H(\nu^*\|\mu)$. This is contradictory with the definition of ν^*. Hence our hypothese on A is false: $\mu \ll \nu^*$. Finally ν^* and μ are equivalent.

The equivalence of ν^* and μ means that $\mathcal{M}(\nu^*) = \mathcal{M}(\mu)$ hence $\mathcal{M}_1(\nu^*) = \mathcal{M}_1(\mu)$. By assumption (H1), $\mathcal{M}_b(\nu^*)$ is dense in $\mathcal{M}_1(\nu^*) = \mathcal{M}_1(\mu)$ so that it is possible to construct $\nu_0^b, \dots, \nu_d^b$ of Lemma 1.13 belonging to $\mathcal{M}_b(\nu^*)$ instead of $\mathcal{M}_b(\mu)$. This will be assumed in the sequel.
Since the convex hull of $\langle f, \nu_0^b \rangle, \dots, \langle f, \nu_d^b \rangle$ has non-empty interior, one can extract from the differences $\langle f, \nu_i^b \rangle - \langle f, \nu_j^b \rangle$ a basis of $\mathbb{R}^d$. By linear combinations one

constructs the canonical basis (e_k). But $\nu_i^b - \nu_j^b$ belongs to $\mathcal{T}(\nu^*)$, so that the k combinations of $\nu_i^b - \nu_j^b$ denoted by η_k also belongs to $\mathcal{T}(\nu^*)$. So we constructed $\eta_k \in \mathcal{T}(\nu^*)$ such that $\left\langle f, \eta_k \right\rangle = e_k$.
Now, Proposition 1.11 and the optimality of ν^* imply

$$\nabla H(\nu^* \| \mu).\eta = 0, \qquad \forall \eta \in \mathcal{T}(\nu^*) \text{ such that } \langle f, \eta \rangle = 0. \tag{1.26}$$

For any $\eta \in \mathcal{T}(\nu^*)$, define

$$\zeta \stackrel{\text{def}}{=} \eta - \sum_{k=1}^{d} \langle f_k, \eta \rangle \eta_k.$$

Then $\langle f, \zeta \rangle = 0$ so that (1.26) can be applied, hence

$$\nabla H(\nu^* \| \mu).\eta = \left\langle \sum_{k=1}^{d} \Big(\nabla H(\nu^* \| \mu).\eta_k \Big) f_k, \eta \right\rangle, \qquad \forall \eta \in \mathcal{T}(\nu^*).$$

Defining $\alpha_k \stackrel{\text{def}}{=} \nabla H(\nu^* \| \mu).\eta_k$, this is exactly equation (1.25). The theorem is proved. ■

> **Remark :** The assumption for z to belong to the interior of $f\big(\mathcal{M}_1(\mu)\big)$ implies that this interior is not empty, hence $f_1, \ldots, f_d$ are *independent* as functionals on $\mathcal{M}_1(\mu)$. The converse is true, i.e. the independence implies that the interior is not empty, but simplifying the condition on z is not that easy. The image of f is a convex set, more precisely a simplex, but its vertices are not simply expressed.
>
> We would like to get an equivalence in (1.25) that is modulo N_ν (i.e. $\mathcal{T}(\nu^*) = \mathcal{T}(\mu)$). Corollary 1.16 proves this is here possible, but in more general settings is it sometimes *not* possible.

♦♦♦

Corollary 1.16 *Under the assumptions of Theorem 1.15, $\mathcal{T}(\nu^*) = \mathcal{T}(\mu)$ and ν^* is given by*[7]

$$\nu^* = \mu e^{\alpha^t f - \Lambda(\alpha^t f)}, \qquad \textit{where} \quad \Lambda(\alpha^t f) \stackrel{\text{def}}{=} \log \int e^{\alpha^t f} \, \mathrm{d}\mu. \tag{1.27}$$

[7] We shall denote with the matrix notation $\alpha^t f \stackrel{\text{def}}{=} \sum_{k=1}^{d} \alpha_k f_k$.

Proof : By Corollary 1.6, (1.25) implies that there exists $\lambda \in \mathbb{R}$ such that

$$\log \frac{\mathrm{d}\nu^*}{\mathrm{d}\mu} = \alpha^t f + \lambda \ \nu^* \text{a.s.}$$

But ν^* is a probability measure, and since ν^* and μ are equivalent, it means that $\lambda = -\Lambda(\alpha^t f)$, where $\Lambda(g)$ is the logarithmic generating function of g. The characterization (1.27) is proved.
Now it is clear that the density $\mathrm{d}\nu^*/\mathrm{d}\mu$ is bounded from zero and infinity, so that by Lemma 1.3, $\mathcal{T}(\nu^*) = \mathcal{T}(\mu)$. ■

1.7 Heuristic in the optimization problem

In order to measure the change around a point ν, the well-known martingale in large deviations is $M_0 = 1$ and for $n \geq 1$

$$M_n \stackrel{\text{def}}{=} \exp\left\{\sum_{i=1}^n \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(X_i)\right\} = \exp\left\langle nL_n, \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right\rangle, \tag{1.28}$$

where L_n is the empirical measure

$$L_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{X_i}.$$

The best changes of measure for deviations for a function $f : E \mapsto \mathbb{R}^d$ (i.e. when the measured quantity is $f(X_i)$) correspond to measures ν^* minimizing the entropy (i.e. optimization problem (O1)). By Theorem 1.15, the solution verifies

$$\nabla H(\nu^* \| \mu) \equiv \alpha^t f \pmod{N_{\nu^*}}, \qquad \text{with } \alpha \in \mathbb{R}^d.$$

When the optimization problem (O1) concerns a small parameter z (assuming f is centered), if we heuristically expand the entropy function around μ, the problem becomes

$$\inf\left\{\nabla^2 H.\eta.\eta \ \text{ where } \ \langle f, \eta\rangle = z\right\}, \tag{1.29}$$

where $\eta \in \mathcal{T}(\mu)$ and $f, z \in \mathbb{R}^d$. Assuming the entropy H is differentiable at the solution η, there exists $\alpha \in \mathbb{R}^d$ such that $\nabla^2 H.\eta \equiv \alpha^t f$. Some computations yield the solution

$$\begin{aligned}
\alpha &= \Sigma^{-1} z & (1.30)\\
\eta &= \nabla^2 H^{-1}.\big(z^t \Sigma^{-1} f\big), & (1.31)\\
\nabla^2 H.\eta.\eta &= z^t \Sigma^{-1} z, & (1.32)\\
\text{where} \quad & \Sigma_{ij} = \langle f_i, \nabla^2 H^{-1}.f_j \rangle. & (1.33)
\end{aligned}$$

For such a solution η, we associate a balanced measure $\nu = \mu + \eta$ and

$$\log \frac{\mathrm{d}\nu}{\mathrm{d}\mu} = \alpha^t f,$$

and the martingale M_n defined by (1.28) becomes heuristically

$$M_n \simeq 1 + \sum_{i=1}^{n} \alpha^t f(X_i).$$

If we remove the constant terms 1, if we suppress the constant multiplicative term α (when z varies, α covers $\mathbb{R}^d$), this strongly suggests to study the multi-dimensional martingale $M_0 = 0$ and for $n \geq 1$

$$M_n = \sum_{i=1}^{n} f(X_i) \in \mathbb{R}^d. \tag{1.34}$$

First, it is indeed a martingale, since f is supposed centered; second, this is precisely a martingale which is classically used in the proof of the CLT for i.i.d. variables, as shown by the following theorem.

1.8 Variance calculation

We formulate the result (a proof using the martingale method and implying this theorem is given for Theorem 2.8).

Theorem 1.17 *Let $f_1, \ldots, f_d$ be bounded continuous functions and denote $(f_1, \ldots, f_d)$ by $f : \mathcal{M}_1(E) \mapsto \mathbb{R}^d$. Assume furthermore that the functions are independent and $\langle f, \mu \rangle = 0$.*
Let $\{X_n,\ n \geq 1\}$ be an i.i.d. sequence with law μ. Then $\{f(X_n),\ n \geq 1\}$ verifies a central limit theorem with variance Σ such that

$$\Sigma_{ij} \stackrel{\text{def}}{=} \big\langle \nabla^2 H(\mu)^{-1}.f_i, f_j \big\rangle = \int f_i f_j \,\mathrm{d}\mu.$$

Since Theorem 1.17 is valid for any positive integer d, it yields the following functional theorem.

Theorem 1.18 *Let $\{X_n,\ n \geq 1\}$ be an i.i.d. sequence with law μ and L_n its empirical measure. Then $\sqrt{n}(L_n - \mu)$ converges in $\mathcal{T}(\mu)$ to a gaussian process with covariance $\nabla^2 H^{-1}(\mu)$.*

2 The Markov framework

We now turn the tools developped in the i.i.d. case to the Markov case. The framework looks very much like the i.i.d. one, but there is a new concept of local balance. This means that a chain enters a state as often as it exits it, in means (i.e. with respect to a stationary measure). However there are some problems of even defining the irreducibility in continuous spaces, and we will try to bypass these difficulties. Contrary to the i.i.d. case, we have to restrict the class of functions for which a central limit theorem holds, but this is naturally done.

2.1 Notation

The framework is the following: (E, d) is a Polish space and $P(.,.)$ is a transition kernel on E, not necessarily ergodic but irreducible in the sense [6] of ϕ-irreducibility:

Definition 2.1 (ϕ-irreducibility) *The transition kernel P is said ϕ-irreducible if there exists a σ-finite positive mesure ϕ with support E such that*

$$\phi(A) > 0 \implies x \rightsquigarrow A, \quad \forall x \in E,$$

where $x \rightsquigarrow A$ means that there exists n such that $P^{(n)}(x, A) > 0$. The measure ϕ is called an irreducible measure. We shall denote by π a maximal irreducible measure, i.e. an irreducible measure such that $\pi P \ll \pi$.

We shall consider *finite* measures in $E^2 = E \times E$ and their projections. Let $\nu \in \mathcal{M}(E^2)$, we shall denote its one-dimensional projections by

$$\begin{cases} \nu_1(.) \stackrel{\text{def}}{=} \nu(., E) \in \mathcal{M}(E), \\ \nu_2(.) \stackrel{\text{def}}{=} \nu(E, .) \in \mathcal{M}(E). \end{cases} \tag{2.1}$$

If such a measure is additionally a ***probability*** measure, it defines a Markov kernel N: $\mathrm{Supp}(\nu_1) \mapsto \mathrm{Supp}(\nu_2)$ and a reverse kernel $\tilde{N}$: $\mathrm{Supp}(\nu_2) \mapsto \mathrm{Supp}(\nu_1)$ by

$$\begin{cases} N(x,.) \stackrel{\text{def}}{=} \dfrac{\mathrm{d}\nu(x,.)}{\mathrm{d}\nu_1}, \\ \tilde{N}(y,.) \stackrel{\text{def}}{=} \dfrac{\mathrm{d}\nu(.,y)}{\mathrm{d}\nu_2}. \end{cases}$$

These kernels are defined ν_1-almost surely [resp. ν_2] and satisfy

$$\nu_1 N(\mathrm{d}x, \mathrm{d}y) = \nu(\mathrm{d}x, \mathrm{d}y) = \nu_2 \tilde{N}(\mathrm{d}y, \mathrm{d}x).$$

In Section 1, we had to restrict to probability measures ν absolutely continuous w.r.t. the measure μ. Since P replaces μ in the Markov settings we define the "regular" measures ν w.r.t. P by

$$\nu \in \mathcal{M}(P) \iff \nu \ll \nu_1 \otimes P \iff N(x,.) \ll P(x,.), \quad \nu_1 \ a.s. \tag{2.2}$$

The notation $\nu_1 \otimes P$ stands for the measure $\nu_1(\mathrm{d}x)P(x, \mathrm{d}y)$. In order to prevent singular behaviour, it is necessary to impose an additional condition[8] on ν: $\nu_1 \ll \pi$. This notion of regularity is transitive: $\nu \in \mathcal{M}(P)$ and $\nu' \in \mathcal{M}(N)$ implies that $\nu' \in \mathcal{M}(P)$. Moreover,

$$\nu' \in \mathcal{M}(N) \iff \nu' \ll \nu.$$

In the i.i.d case, there was only one constraint for the measures: the mass be one. In the Markov case, we emphasize that there is furthermore a balance constraint, so that we shall consider only *balanced measures.*

Definition 2.2 (Balanced measures) *A balance measure $\nu \in \mathcal{M}(E^2)$ is a finite measure such that $\nu_1 \stackrel{\text{law}}{=} \nu_2$ in (2.1). The set of balanced measures is denoted by $\mathcal{M}_s(E^2)$ and the set of* balanced probability measures *by $\mathcal{M}_1(E^2)$. If furthermore the measures are* regular *w.r.t. a kernel P as in (2.2), the sets will be denoted by $\mathcal{M}_s(P)$ and $\mathcal{M}_1(P)$ respectively.*

For a measure $\nu \in \mathcal{M}_1(P)$, we shall define the sets:

- $L_1(\nu)$, the real ν–integrable functions;
- $\mathcal{M}(\nu)$, the finite signed measures absolutely continuous w.r.t. ν;

[8] In this condition we use only the *maximality* of π. For a measure ν, ν_1 is maximal for N on $\mathrm{Supp}(\nu_1)$ but not necessarily irreducible.

- $L_\infty(\nu) \subset L_1(\nu)$, the real ν–measurable a.s. bounded functions;
- $\mathcal{M}_b(\nu) \subset \mathcal{M}(\nu)$, the measures η such that $\mathrm{d}\eta/\mathrm{d}\nu$ is ν-a.s. bounded;
- $\mathcal{M}_0(N) \subset \mathcal{M}_s(N)$, the balanced, centered ($\int \mathrm{d}\eta = 0$), N-regular measures, usually denoted by η, ζ;
- $\mathcal{F}_0(\nu) \subset L_1(\nu)$, the real ν–integrable functions, usually denoted by ϕ or ψ, verifying $\phi\nu \in \mathcal{M}_0(\nu)$.

As for the i.i.d. framework, when no confusion arises, we shall drop the indication of the reference probability measure ν. Note that there are classical bijections $\mathcal{M} \leftrightarrow L_1$, $\mathcal{M}_b \leftrightarrow L_\infty$ or $\mathcal{M}_0 \leftrightarrow \mathcal{F}_0$ expressed by:

$$\eta \in \mathcal{M}(\nu) \longleftrightarrow \phi \in \mathcal{F}(\nu), \qquad \text{with} \quad \phi = \frac{\mathrm{d}\eta}{\mathrm{d}\nu} \ \nu\text{-a.s.} \quad \text{or} \quad \eta = \phi\nu.$$

2.2 Topology

Topology The paragraph "Topology" of Section 1.2 applies verbatim here.

Quotient spaces The problem now is to construct a set $N \subset L_\infty$ such that $\mathcal{M}_0 = N^o$. This means to find a way to make a measure centered and balanced. In the i.i.d. case N was the vector space generated by $\mathbb{I}$. In the Markov case, the balance condition is more difficult to handle.
Let $\nu \in \mathcal{M}(E^2)$. For any function $\beta \in L_\infty(\nu_1)$, we define

$$\begin{aligned} K_\beta : \quad \mathcal{M}(\nu) &\mapsto \mathbb{R} \\ \eta \quad &\to \langle \eta_1, \beta \rangle - \langle \eta_2, \beta \rangle. \end{aligned} \tag{2.3}$$

The application K_β is a continuous linear functional, so that its kernel is closed. Moreover, if the terms in the right hand side of (2.3) are null for all bounded $\beta \in L_\infty(\nu)$, then $\eta_1 = \eta_2$ and the measure η is balanced, hence

$$\mathcal{M}_s(N) = \bigcap_{\beta \in L_\infty(\nu)} \ker K_\beta.$$

Now, the continuous linear functional K_β is identified to an element of the dual of $\mathcal{M}(\nu)$, namely $K_\beta \in L_\infty(\nu)$ with

$$K_\beta(x, y) = \beta(x) - \beta(y) \quad \nu - \text{a.s.}$$

The set N is the closed vector subspace of $L_\infty(\nu)$:

$$N \stackrel{\text{def}}{=} \left\{ \alpha\mathbb{I} + K_\beta, \ \alpha \in \mathbb{R}, \ \beta \in L_\infty(\nu) \right\}.$$

Topology on quotient spaces The paragraph "Topology on quotient spaces" of Section 1.2 applies here.
While permuting L_1 and L_∞, we need $N \in L_1$ so that K_β is identically defined but the only change $\beta \in L_1$ and

$$N \stackrel{\text{def}}{=} \left\{ \alpha \mathbb{1} + K_\beta,\ \alpha \in \mathbb{R},\ \beta \in L_1(\nu) \right\}.$$

> **Remark :** Unfortunately, there is no such isomorphism $L_1/N \leftrightarrow \mathcal{F}_0$: in the i.i.d. case, we have $L_1 = \mathcal{F}_0 \oplus N$. This means that, for a function $f \in L_1$, it may happens that there is no decomposition $f = \phi + f'$ with $\phi \in \mathcal{F}_0$ and $f' \in N$. This kind of decomposition depends strongly of the structure of P and of f. This remark also holds in the L_1 space.
>
> ♦♦♦

2.3 Calculus of variations

Tangent space We briefly recall the major steps of the construction of tangent spaces. It is merely a repetition of Section 1.3 in the context of Markov chains by replacing $\mathcal{M}_1(\mu)$ by $\mathcal{M}_1(P)$. Definition 1.1 and Proposition 1.2 holds. In short:

Proposition 1.2 *For all $\nu \in \mathcal{M}_1(P)$, $\mathcal{T}(\nu) = \mathcal{M}_0(\nu) \cap \mathcal{M}_b(\nu)$.*

The tangent space $\mathcal{T}(\nu)$ is canonically equipped with the L_∞ norm and the induced topology. Its dual is L_1/N. Lemma 1.3 also holds.

Derivatives The derivatives are defined exactly the same manner as in Section 1.4, but the expression of the gradient is a bit different.
We consider the entropy function on the pairs defined by

$$\begin{aligned} H(\nu \| P) &\stackrel{\text{def}}{=} \iint \nu(\,\mathrm{d}x,\,\mathrm{d}y) \log \frac{\mathrm{d}\nu}{\mathrm{d}(\nu_1 \otimes P)}(x,y) \quad \forall \nu \in \mathcal{M}_1(P) \\ &= \iint \nu(\,\mathrm{d}x,\,\mathrm{d}y) \log \frac{\mathrm{d}N(x,.)}{\mathrm{d}P(x,.)}(y) \end{aligned}$$

and $H(.\|P) \stackrel{\text{def}}{=} \infty$ outside $\mathcal{M}_1(P)$. For the sake of brevity, we shall note:

$$H(\nu \| P) = \int \log \frac{\mathrm{d}N}{\mathrm{d}P}\,\mathrm{d}\nu.$$

Using Definition 1.4, Proposition 1.5 becomes:

Proposition 2.3 *The gradient $\nabla H(\nu\|P)$ is a linear functional $\mathcal{T}(\nu) \mapsto \mathbb{R}$ defined where the entropy $H(\nu\|P)$ is finite by:*

$$\nabla H(\nu\|P).\eta = \int \log \frac{\mathrm{d}N}{\mathrm{d}P}\,\mathrm{d}\eta, \quad \forall \eta \in \mathcal{T}(\nu). \tag{2.4}$$

Proof : It is similar to the proof of Proposition 1.5. The only difference is the expansion of H. For $\eta \in \mathcal{T}(\nu)$, one defines $\xi(t) \stackrel{\text{def}}{=} \nu + t\eta$, the affine path in $\mathcal{M}_1(P)$ defining η (for t sufficiently small) and

$$\phi \stackrel{\text{def}}{=} \frac{\mathrm{d}\eta}{\mathrm{d}\nu} \in L_\infty(\nu), \qquad \phi_1 \stackrel{\text{def}}{=} \frac{\mathrm{d}\eta_1}{\mathrm{d}\nu_1} \in L_\infty(\nu_1).$$

Now, a short calculation shows that

$$\frac{\mathrm{d}\Xi}{\mathrm{d}P}(x,y) \stackrel{\text{def}}{=} \frac{\mathrm{d}\xi}{\mathrm{d}\xi_1 \otimes P}(x,y) = \frac{\mathrm{d}N}{\mathrm{d}P}(x,y)\frac{1+t\phi(x,y)}{1+t\phi_1(x)}. \tag{2.5}$$

Then the equation (1.14) becomes

$$H(\xi(t)\|P) - H(\nu\|P) = x\int \log \frac{\mathrm{d}N}{\mathrm{d}P}\,\mathrm{d}\eta + \int \log \frac{1+t\phi}{1+t\phi_1}\,\mathrm{d}\xi(t). \tag{2.6}$$

Since $\eta \in \mathcal{T}(\nu)$, $\int \phi\,\mathrm{d}\xi(t)$ and $\int \phi_1\,\mathrm{d}\xi(t)$ tend to 0 and the last term in (2.6) tends to zero faster than t. Equation (2.4) is proved. ∎

Corollary 1.6 holds with characterization (1.15) becoming (2.7):

Corollary 2.4 *When the gradient is defined, it is a continuous functional $\mathcal{T}(\nu) \mapsto \mathbb{R}$. Hence there exists a unique canonical representation of $\nabla H(\nu)$ in $L_1(\nu)/N$ that will also be denoted by $\nabla H(\nu)$. Moreover $f \equiv \nabla H(\nu)$ if, and only if,*

$$\exists \alpha \in \mathbb{R},\ \beta \in L_1(E):\ f = \log \frac{\mathrm{d}N}{\mathrm{d}P} + \alpha + K_\beta \quad \nu\text{-a.e.} \tag{2.7}$$

Note that $\beta \in L_1(\nu)$ is not unique. It is unique up to a constant.

Second order derivatives They are defined as in Section 1.4, first the directional derivatives by equation (1.16) and then the Hessian by Definition 1.7. It is described as in Proposition 1.8 with (1.17) modified in (2.8).

Proposition 2.5 *The Hessian* $\nabla^2 H(\nu)$ *does not depend on* P. *It is the bilinear symmetric positive definite functional* $\mathcal{T}(\nu) \times \mathcal{T}(\nu) \mapsto \mathbb{R}$ *defined at every points* ν *where the entropy* $H(\nu\|P)$ *is finite by:*

$$\nabla^2 H(\nu).\eta.\zeta \stackrel{\text{def}}{=} \int \frac{\mathrm{d}\eta}{\mathrm{d}\nu}\frac{\mathrm{d}\zeta}{\mathrm{d}\nu}\,\mathrm{d}\nu - \int \frac{\mathrm{d}\eta_1}{\mathrm{d}\nu_1}\frac{\mathrm{d}\zeta_1}{\mathrm{d}\nu_1}\,\mathrm{d}\nu_1, \quad \forall \eta,\zeta \in \mathcal{T}(\nu). \tag{2.8}$$

Proof : Let $\eta, \zeta \in \mathcal{T}(\nu)$: the Radon-Nikodym derivatives will be denoted by ϕ for $\mathrm{d}\eta/\mathrm{d}\nu$ and by ψ for $\mathrm{d}\zeta/\mathrm{d}\nu$. They are bounded and their projections ϕ_1 and ψ_1 also are bounded (in $L_1(\nu_1)$) so that the Hessian is well defined:

$$\nabla^2 H(\nu\|P).\eta.\zeta = \int \phi\psi\,\mathrm{d}\nu - \int \phi_1\psi_1\,\mathrm{d}\nu_1.$$

In the same way as for the gradient, there exists an open neighborhood $(-u,u)$ of 0 such that $\xi(x) \stackrel{\text{def}}{=} \nu + t\zeta$ belongs to $\mathcal{M}_1(E)$. Since $\zeta \ll \nu$, the support of $\xi(t)$ is exactly the same as ν. Moreover the Radon-Nikodym derivatives of ζ w.r.t. ν is a.s. bounded, therefore $\mathcal{T}\big(\xi(t)\big) = \mathcal{T}(\nu)$ and the right-hand side of (1.16) makes sense. Now, using (2.5),

$$\nabla H\big(\xi(t)\|P\big).\eta - \nabla H(\nu\|P).\eta = \int \log\frac{1+t\psi}{1+t\psi_1}\,\mathrm{d}\eta = t\int(\psi-\psi_1)\,\mathrm{d}\eta + o(t),$$

thus the Hessian exists and is equal to equation (2.8). This also proves that the Hessian is bilinear and symmetric. It is positive definite[9] since, by Schwartz' inequality,

$$\int \psi^2\,\mathrm{d}\nu \geq \int\left(\int \psi(x,y)N(x,\,\mathrm{d}y)\right)^2 \nu_1(\,\mathrm{d}x) = \int \psi_1^2\,\mathrm{d}\nu_1. \tag{2.9}$$

There is equality if, only if, $\psi(x,y)$ is constant in y for ν_1 almost all x. By symmetry, using $\nu_1(\,\mathrm{d}x)N(x,\,\mathrm{d}y) \stackrel{\text{law}}{=} \nu_1(\,\mathrm{d}y)\tilde{N}(y,\,\mathrm{d}x)$ in an equation similar to (2.9), $\psi(x,y)$ is constant in x for ν_1 almost all y. Finally ψ is constant ν-a.s. and since ψ is centered, this constant must be 0. Finally, there is equality if, and only if, $\psi = 0$ ν-a.s. ■

Proposition 2.6 *When* ν *has finite entropy, the Hessian* $\nabla^2 H(\nu)$, *considered as a linear functional from* $\mathcal{T}$ *onto* $\mathcal{T}^* = L_1/N$ *is defined as*

$$\nabla^2 H(\nu).\eta \equiv \frac{\mathrm{d}\eta}{\mathrm{d}\nu}(x,y) - \frac{\mathrm{d}\eta_1}{\mathrm{d}\nu_1}(x) \qquad \forall \eta \in \mathcal{T}. \tag{2.10}$$

[9] It is necessarily positive, since the entropy is convex.

The Hessian is injective and its inverse, when it exists, is given by[10]

$$\frac{\mathrm{d}(\nabla^2 H(\nu)^{-1}.f)}{\mathrm{d}\nu}(x,y) = f(x,y) + (I-N)^{-1}f_1(y) + (I-\tilde{N})^{-1}f_2(x) \tag{2.11}$$

$$\text{where} \qquad f_1(x) \stackrel{\text{def}}{=} \int f(x,y)N(x,\,\mathrm{d}y), \tag{2.12}$$

$$f_2(y) \stackrel{\text{def}}{=} \int f(x,y)\tilde{N}(y,\,\mathrm{d}x), \tag{2.13}$$

where the function f is supposed centered[11]*, i.e.* $\langle f,\nu\rangle = 0$.

Proof : Using the balance property, it is easily checked that

$$\left\langle \frac{\mathrm{d}\eta}{\mathrm{d}\nu}(x,y) - \frac{\mathrm{d}\eta_1}{\mathrm{d}\nu_1}(x), \zeta \right\rangle = \int \frac{\mathrm{d}\eta}{\mathrm{d}\nu}\,\mathrm{d}\zeta - \int \frac{\mathrm{d}\eta_1}{\mathrm{d}\nu_1}(x)\,\mathrm{d}\zeta_1 = \nabla^2 H(\nu).\eta.\zeta,$$

for all $\zeta \in \mathcal{T}(\nu)$, hence (2.10) holds. The Hessian is injective because the associated bilinear form is definite positive: if $\nabla^2 H(\nu).\eta \equiv 0$ then $\nabla^2 H(\nu).\eta.\eta = 0$ and $\eta = 0$.

Before calculating the inverse $\nabla^2 H(\nu)^{-1}$, it must be checked that the right-hand side of equation (2.11) defines a balanced centered measure. The $(I-N)^{-1}$ and $(I-\tilde{N})^{-1}$ operators represent Dinkyn's transform of centered functions: it is not always defined but, when it is, it defines a unique centered function. So first, the inverse is uniquely defined (this could have been deduced from the injectivity). The next calculations prove that it is balanced and centered.
We shall denote by ψ the density of η w.r.t. ν. The expression of the inverse is obtained by using necessary conditions: *the inverse does not necessarily exist.*
The identity $\nabla^2 H(\nu).\eta \equiv f$ implies that there exist $\alpha \in \mathbb{R}$ and $\beta \in L_1(\nu_1)$ (defined up to a constant) such that

$$\psi(x,y) - \psi_1(x) = f(x,y) + \alpha + \beta(x) - \beta(y) \quad \nu\text{-a.s.} \tag{2.14}$$

If we integrate equation (2.14) w.r.t. ν, we get

$$0 = \langle f,\nu\rangle + \alpha,$$

[10]The notation f_1 and f_2 is consistent with (2.1) since we have $\nu_1 f_1 = (\nu f)_1 \stackrel{\text{def}}{=} \int f(.,y)\nu(.,\,\mathrm{d}y)$ and symmetrically $\nu_2 f_2 = (\nu f)_2$ (with $\nu_1 = \nu_2$ since ν is balanced).

[11]If not centered, f has to be replaced by $f - \langle f,\nu\rangle$ in the equations.

so that $\alpha = 0$, since we assumed f is centered. If we integrate partially equation (2.14) w.r.t. $N(x, \mathrm{d}y)$ and $\tilde{N}(y, \mathrm{d}x)$, we get, using (2.12)–(2.13),

$$0 = f_1 + \beta - N\beta \quad \nu_1\text{-a.s.} \tag{2.15}$$
$$\psi_1 - \tilde{N}\psi_1 = f_2 + \tilde{N}\beta - \beta \quad \nu_1\text{-a.s.} \tag{2.16}$$

Assuming that equations (2.15)–(2.16) can be inverted, the solution is necessarily

$$\beta = -(I-N)^{-1} f_1 \quad \nu_1\text{-a.s.} \tag{2.17}$$
$$\psi_1 = (I-\tilde{N})^{-1} f_2 - \beta \quad \nu_1\text{-a.s.} \tag{2.18}$$

Note that β is here implicitly chosen centered (by convention, Dynkin's transforms are chosen centered), which is possible since it was defined up to a constant and ν_1-integrable. Putting the β and ψ_1 as defined in (2.17)–(2.18) into (2.14) yields the result: if it exists, necessarily

$$\psi(x,y) = f(x,y) + (I-N)^{-1} f_1(y) + (I-\tilde{N})^{-1} f_2(x) \quad \nu\text{-a.s.}$$

■

Note that these expressions are valid for all irreducible finite Markov chains P, at a point ν which is irreducible (i.e. N and $\tilde{N}$ are irreducible), e.g. if ν is the stationary measure. It is also valid for any Markov chain P at points ν which are irreducible with finite support.

Remark : One can easily tranform the reciprocal of the Hessian (2.11) into a formulation very close to Appendix B. First note that

$$f_1(x) \stackrel{\text{def}}{=} \int f(x,y) N(x, \mathrm{d}y) = \mathbb{E}\left[f(X_0, X_1) \middle| X_0 = x \right],$$

and similarly $f_2(X_0) = \mathbb{E}\left[f(X_{-1}, X_0) \,|\, X_0\right]$. Then, replacing the Dynkin's transform by its probabilistic counterpart

$$(I-N)^{-1} f_1(X_1) = \mathbb{E}\left[\sum_{i=1}^{\infty} f_1(X_i) \middle| X_1 \right],$$

formula (2.11) becomes for short

$$\nabla^2 H(\nu)^{-1}.f = \mathbb{E}\left[\sum_{i=-\infty}^{\infty} f(X_i, X_{i+1}) \,\middle|\, X_0, X_1 \right] \tag{2.19}$$

where the expectation is taken w.r.t. the process defined by ν and the left-hand side is the density of the measure $\nabla^2 H(\nu)^{-1}.f$ w.r.t. ν, taken as a function of the variables X_0, X_1.

The formula (2.19) is very useful for the verification of the validity of the variance formula

$$\sigma^2 = \left\langle \nabla^2 H(\nu)^{-1}.f, f \right\rangle. \tag{2.20}$$

It yields immediately

$$\sigma^2 = \mathbb{E}\left[f(X_0, X_{-1}) \sum_{i=-\infty}^{\infty} f(X_i, X_{i+1}) \right],$$

which is exactly the general formula (B.1). For this reason, it is expected that the validity of (2.19) and (2.20) goes beyond the scope of Markov chains.

♦♦♦

2.4 Optimization with constraints

This section is the transposition into the Markov framework of Section 1.6, i.e. the following optimization problem (O2).

Problem (O2) *Let $f : E^2 \mapsto \mathbb{R}^d$ be a vector valued function with coordinates $f_1, \ldots, f_d$ belonging*[12] *to $L_\infty(E^2)$. Then the optimization problem is*

$$\inf\Big\{ H(\nu \| P) \quad \textit{where} \quad \langle f, \nu \rangle = z \Big\} \tag{2.21}$$

Replacing μ by P, Assumptions (H1) and (H2) are the same for problem (O2) and Lemma 1.13 holds here. A natural question is: are there settings where Assumptions (H1) and (H2) holds? The answer is yes. Assumptions (H2) can hold in every settings, depending on the function F. Assumptions (H1) is the most difficult one to prove. It is easily checked when E is finite. It also holds if E is countable by [4, Proposition C.2].

Unfortunately, since the entropy $H(.\|P)$ does not necessarily have compact level sets, Corollary 1.14 does not hold. However, Theorem 1.15 still holds when properly modified.

[12] Actually we are interested only to the part of f_i that belongs to the support of P so that we should write $L_\infty(P)$. We do not for simplicity.

Theorem 2.7 *Assume (H1) and (H2) are valid and optimization problem (O2) possesses one solution ν^*. Then this solution has finite entropy, is unique, ν^* and $\pi \otimes P$ are equivalent (i.e. mutually absolutely continuous) and there exists $\alpha_1, \ldots, \alpha_d$ such that*

$$\nabla H(\nu^* \| P) \equiv \sum_{k=1}^{d} \alpha_k f_k \pmod{N_{\nu^*}}. \tag{2.22}$$

Proof : The sketch of the proof is the same as in Theorem 1.15. The difference lies in the sets used to prove the equivalence between ν^* and $\pi \otimes P$.
By Assumption (H1), one can find $\mu \in \mathcal{M}_b \cap \mathcal{M}_1$ equivalent to $\pi \otimes P$ (which is not balanced). Recall that ν^* is regular (its entropy is finite) so that $\nu^* \ll \mu$. To prove the reverse relation, assume that there exists a set $A \subset E^2$ such that

$$\mu(A) > 0 \qquad \text{and} \qquad \nu^*(A) = 0.$$

Then it is possible to construct a set $A' \subset A$ with projection on the first coordinate $A'_1 = \mathrm{proj_1(A')}$ such that

$$\mu(A') > 0, \quad \nu^*(A') = 0 \quad \text{and} \quad \nu_1^*(A'_1) > 0.$$

Finally it is possible to construct A and $A_1 = \mathrm{proj_1(A)}$ such that there exists $\varepsilon > 0$ with

$$\mu(A') > 0, \quad \nu^*(A') = 0, \quad \text{and} \quad \nu_1^* \geq \varepsilon \mu_1 \quad \text{on } A_1.$$

Then one proceeds as in the proof of Theorem 1.15, but the measure which is denoted by π there is still denoted by μ here. It verifies $\mu \in \mathcal{M}_b \cap \mathcal{M}_1$, its entropy is finite, it is equivalent to $\pi \otimes P$ and $\langle f, \mu \rangle = z$.
The convexity of $(a, b) \to a \log(a/b)$ used on A^c yields ($x + y = 1$ with $x, y \geq 0$)

$$\begin{aligned} & H\big(y\nu^* + x\mu \big\| P\big) \\ \leq \;& y \int_{A^c} \log \frac{\mathrm{d}N^*}{\mathrm{d}P} \,\mathrm{d}\nu^* + x \int_{A^c} \log \frac{\mathrm{d}M}{\mathrm{d}P} \,\mathrm{d}\mu + x \int_{A} \log \frac{\mathrm{d}(x\mu)}{\mathrm{d}(y\nu_1^* + x\mu_1) \otimes P} \,\mathrm{d}(x\mu) \\ = \;& yH(\nu^* \| P) + \mu(A) x \log x + x \left(H(\mu \| P) + \frac{1-\varepsilon}{\varepsilon} \int_A \log \frac{\mathrm{d}M}{\mathrm{d}P} \,\mathrm{d}\mu \right). \end{aligned}$$

Since $\mu(A) > 0$, for small x the right side is strictly less than $H(\nu^* \| \mu)$. This is contradictory with the definition of ν^*. Hence our hypothese on A is false: $\mu \ll \nu^*$. Finally ν^* and μ are equivalent.
The rest of the proof is the same as in Theorem 1.15. ■

2.5 Heuristic in the optimization problem

This section is a transposition of Section 1.7 to the Markov framework. We assume P is irreducible, ergodic and π is the stationary probability (hence also a maximal irreducibility measure).
In order to measure the change around a point ν, the well-known martingale in large deviations is $M_0 = 1$ and for $n \geq 1$

$$M_n \stackrel{\text{def}}{=} \exp\left\{\sum_{i=0}^{n-1} \log \frac{\mathrm{d}N}{\mathrm{d}P}(X_i, X_{i+1})\right\} = \exp\left\langle nL_n, \log \frac{\mathrm{d}N}{\mathrm{d}P}\right\rangle, \tag{2.23}$$

where L_n is the pair empirical measure

$$L_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=0}^{n-1} \delta_{X_i, X_{i+1}}.$$

We are interested in small deviations for a function $f : E^2 \mapsto \mathbb{R}^d$. Assuming Theorem 2.7 applies, one heuristically derives equations (1.30)–(1.33) for the optimal change of measure ν^* and its expansion η around $\pi \otimes P$. For such a solution η, we associate a balanced measure $\nu = \pi \otimes P + \eta$ and

$$\log \frac{\mathrm{d}N}{\mathrm{d}P}(x, y) = \alpha^t f \sim \psi(x, y) - \psi_1(x),$$

where ψ is the density of η w.r.t. $\pi \otimes P$. Now, denoting $\alpha^t f$ by F and reminding that F is small, equation (1.31) and Proposition 2.6 yield the expression of ψ, equation (2.18) yields the expression of ψ_1, so that

$$\log \frac{\mathrm{d}N}{\mathrm{d}P}(x, y) \sim F(x, y) + (I - P)^{-1}F_1(y) - (I - P)^{-1}F_1(x),$$

and the martingale M_n defined by (2.23) becomes heuristically

$$M_n \simeq 1 + \sum_{i=0}^{n-1} F(X_i, X_{i+1}) + (I - P)^{-1}F_1(X_n) - (I - P)^{-1}F_1(X_0).$$

If we remove the constant terms 1 and $(I - P)^{-1}F_1(X_0)$, if we suppress the constant multiplicative term α, this strongly suggests to study the multi-dimensional martingale $M_0 = 0$ and for $n \geq 1$

$$M_n = \sum_{i=0}^{n-1} f(X_i, X_{i+1}) + (I - P)^{-1}f_1(X_n) \in \mathbb{R}^d. \tag{2.24}$$

First, it is indeed a martingale; second, this is precisely a martingale which is classically used in the proof of the CLT for Markov chains, as shown by the following theorem.

Theorem 2.8 (Central limit theorem) *Let $\{X_n,\ n \geq 1\}$ be a Markov chain with ergodic irreducible transition kernel P and stationary measure π. Let $f : E^2 \mapsto \mathbb{R}^d$ be a centered functional (i.e. $\langle \pi \otimes P, f\rangle = 0$) such that $h = (I-P)^{-1} f_1 \in \mathbb{R}^d$ exists, is π-integrable and centered. Then the sum S_n converges to a normal law $\mathcal{N}(0, n\Sigma)$:*

$$\frac{S_n}{\sqrt{n}} \stackrel{\text{def}}{=} \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} f(X_k, X_{k+1}) \xrightarrow{law} \mathcal{N}(0, \Sigma). \tag{2.25}$$

Moreover, $\tilde{h} \stackrel{\text{def}}{=} (I - \tilde{P})^{-1} f_2$ and the inverse $\nabla^2 H^{-1}.f$ exist, and the variance Σ is expressed by

$$\Sigma_{ij} = \langle f_i, \nabla^2 H^{-1}.f_j \rangle \tag{2.26}$$

$$= \int f_i(x,y)\big(f_j(x,y) + h_j(y)\big) + h_i(y) f_j(x,y) \pi(\mathrm{d}x) P(x, \mathrm{d}y). \tag{2.27}$$

Proof : We consider the martingale M_n defined by (2.24). Since f is centered, the ergodic theorem ensures that M_n/n tends to 0 a.s. and in L_2.
By definition

$$\langle M\rangle_n \stackrel{\text{def}}{=} \sum_{k=0}^{n-1} \mathbb{E}\left[(M_{k+1} - M_k)(M_{k+1} - M_k)^t \,\middle|\, \mathcal{F}_k\right] \in \mathbb{R}^{d^2}.$$

A simple calculation yields (denoting zz^t by z^2 and $uv^t + vu^t$ by $u \diamond v$)

$$\begin{aligned}
& \mathbb{E}\left[\big(M_{k+1} - M_k\big)^2 \,\middle|\, \mathcal{F}_k\right] \\
= \ & \mathbb{E}\left[\big(h(X_{k+1}) - h(X_k) + f(X_k, X_{k+1})\big)^2 \,\middle|\, \mathcal{F}_k\right] \\
= \ & \big[P(h^2) - (Ph)^2 - h \diamond Ph - h \diamond f_1\big](X_k) \\
& + \mathbb{E}\left[f^2(X_k, X_{k+1}) + f(X_k, X_{k+1}) \diamond h(X_{k+1}) \,\middle|\, \mathcal{F}_k\right].
\end{aligned}$$

By the ergodic theorem, we deduce that

$$\begin{aligned}
\frac{\langle M\rangle_n}{n} \xrightarrow{a.s.} \Sigma &= \mathbb{E}^*\left[\big[P(h^2) - (Ph)^2 - h \diamond Ph - h \diamond f_1\big](X_0)\right] \\
&\quad + \mathbb{E}^*\left[f^2(X_0, X_1) + f(X_0, X_1) \diamond h(X_1)\right] \\
&= \int \big[f^2(x,y) + f(x,y) \diamond h(y)\big] \pi(\,\mathrm{d}x) P(x,\,\mathrm{d}y),
\end{aligned}$$

where $\mathbb{P}^*$ is the stationary probability for the Markov process $\{X_k,\ k \in \mathbb{Z}\}$ with transition probabilities P and $\mathbb{E}^*$ the associated expectation. One easily checks that the first term in Σ is null. Now, (2.25) and (2.27) is the consequence of the central limit theorem for martingales.
The expression of Σ in (2.26) is the sequel of following formulæ:

$$\begin{aligned} & \int h_i(y) f_j(x,y)\pi(\,\mathrm{d}x)P(x,\,\mathrm{d}y) \\ = \quad & \mathbb{E}^*\left[\sum_{k=1}^{\infty} f_i(X_k, X_{k+1}) f_j(X_0, X_1)\right] \\ = \quad & \mathbb{E}^*\left[\sum_{k=-\infty}^{-1} f_i(X_0, X_1) f_j(X_k, X_{k+1})\right] \\ = \quad & \int f_i(x,y)\tilde{h}_j(x)\pi(\,\mathrm{d}x)P(x,\,\mathrm{d}y). \end{aligned}$$

Indeed, using the ergodicity of the Markov chain and that f is centered, one can verify that convergence occurs in

$$\begin{aligned} h(x) &= \mathbb{E}^*\left[\sum_{k=0}^{\infty} f(X_k, X_{k+1}) \,\Big|\, X_0 = x\right], \\ \tilde{h}(x) &= \mathbb{E}^*\left[\sum_{k=-\infty}^{0} f(X_{k-1}, X_k) \,\Big|\, X_0 = x\right], \end{aligned}$$

so that $\tilde{h}$ exists when h exists and all upper formulæ are valid. The expression (2.26) is straightforwardly derived since, by Proposition 2.6, the density of $\nabla^2 H^{-1}.f(x,y)$ w.r.t. $\pi \otimes P$ is $f(x,y) + h(y) + \tilde{h}(x)$. ■

Martingale expression : The “good” martingale to prove that the variance of a sum $S_n = \langle f, nL_n\rangle$ defined as a functional of an empirical measure L_n (where the underlying process is not necessarily Markov) is indeed

$$\sigma^2 = \big\langle \nabla^2 H^{-1}.f, f \big\rangle,$$

is probably, following the upper heuristic calculation,

$$M_n \stackrel{\mathrm{def}}{=} \big\langle nL_n, \nabla\phi.(\nabla^2 H^{-1}.f)\big\rangle,$$

where ϕ (actually $\phi(G,R)$) is the functional defining the general change of measure and the entropy, i.e $H(G\|R) \stackrel{\text{def}}{=} \langle\phi(G,R),G\rangle$. It can be sometimes simpler to use

$$M_n \stackrel{\text{def}}{=} \mathbb{E}\left[S_\infty \middle| \mathcal{F}_n\right].$$

♦♦♦

Functional theorem Since Theorem 2.8 is valid for any positive integer d, one could derive an analog of Theorem 1.18. The problem is that the corresponding empirical measure

$$L_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=0}^{n-1} \delta_{X_k,X_{k+1}}$$

does not generally belong to $\mathcal{T}(\pi\otimes P)$ which is the limit space (the distance between L_n and $\mathcal{T}(\pi\otimes P)$ tends to 0). This difficulty can be overcome but the formulation of the theorem, though very close to Theorem 1.18, becomes heavy.

3 The continuous time Markov framework

This section aims at giving a quick look at the continuous time framework. All subsequent propositions are rigorous but details of proofs are left to the reader.

3.1 Framework

For the sake of simplicity, we will consider a denumerable state space E (so we do not need to write Radon-Nykodim derivatives and processes are pure jumps processes). We shall work with *generators* in continuous time, instead of pair measures (for discrete time).

Definition 3.1 *The set of balanced generators is*

$$\mathcal{G}_s \stackrel{\text{def}}{=} \mathcal{M}_1(E)\times\mathcal{M}_s(E_*^2),$$

where $E_*^2 \stackrel{\text{def}}{=} \{(x,y)\in E\times E,\ x\neq y\}$ *is the set of off-diagonal pairs. For* $G\in\mathcal{G}_s$*, we shall conventionally write*[13] $G=(\dot g,\ddot g)$

[13] The dots have nothing to do with derivatives. It it simply useful to signal the mono-dimensional measure $\dot g$ and the bi-dimensional measure $\ddot g$.

The balanced generator G is closely related to the generator of Markov processes. We shall denote both by G and writes the intensities of jumps by a capital letter:

$$G_{xy} \stackrel{\text{def}}{=} \frac{\ddot{g}_{xy}}{\dot{g}_x} \text{ if } x \neq y \quad \text{and} \quad G_{xx} \stackrel{\text{def}}{=} -\frac{\ddot{g}_x}{\dot{g}_x}$$

with the convention $\ddot{g}_x \stackrel{\text{def}}{=} \sum_y \ddot{g}_{xy} = \sum_z \ddot{g}_{zx}$ (since $\ddot{g}$ is balanced). Note that such generators, when irreducible (i.e. when $\ddot{g}$ is irreducible), are ergodic and that $\ddot{g}_{xy}$ represents the *stationary number of jumps* from x to y per unit time while $\dot{g}$ is the *stationary measure.* For more details about these generators, see [2].

3.2 Entropy

Definition 3.2 (Entropy) *Let $R = (R_{xy})$ be the generator of a Markov process on E and $G \in \mathcal{G}_s$. The entropy of G relatively to R is defined by*

$$\begin{aligned} H(G\|R) &\stackrel{\text{def}}{=} \sum_{x\neq y} \dot{g}_x \left(G_{xy} \log \frac{G_{xy}}{R_{xy}} - G_{xy} + R_{xy} \right) \\ &= \sum_{x \neq y} \ddot{g}_{xy} \log \frac{\ddot{g}_{xy}}{\dot{g}_x R_{xy}} - \ddot{g}_{xy} + \dot{g}_x R_{xy}. \end{aligned}$$

This entropy has a quite natural interpretation in terms of information theory. It appears in large deviations for continuous Markov chains [2].

3.3 Tangent space

Note that the measure $\ddot{g}$ evolves without constraint on its sum: $\ddot{g}$ belongs to $\mathcal{M}_s(E^2)$, not to $\mathcal{M}_1(E^2)$ as in the discrete time framework. This is reflected in the structure of the tangent space.

Proposition 3.3 *Let $G \in \mathcal{G}_s$. The vector tangent space $\mathcal{T}(G)$ is defined by*

$$\mathcal{T}(G) = \mathcal{M}_0^b(\dot{g}) \times \mathcal{M}_s^b(\ddot{g}).$$

It is equipped with the L_∞ norm. Its dual[14] *is*

$$\mathcal{T}^*(G) = L_1(\dot{g})/N_1 \times L_1(\ddot{g})/N_2.$$

where $N_1 \stackrel{\text{def}}{=} \{\alpha \mathbb{1},\ \alpha \in \mathbb{R}\}$ and $N_2 \stackrel{\text{def}}{=} \{K_\beta,\ \beta \in L_1(\ddot{g}_1)\}$.

We shall denote a tangent vector by $\eta = (\dot{\eta}, \ddot{\eta})$ or $\zeta = (\dot{\zeta}, \ddot{\zeta})$. Note that the dual $\mathcal{T}^*(G)$ can be equivalently written $(L_1(\dot{g}) \times L_1(\ddot{g}))/(N_1 \times N_2)$.

[14] [1, IV 2.8] use the direct sum $\oplus$ which is equivalent here to a product.

3.4 Derivatives

Calculations very similar to those of Propositions 2.3 and 2.5 yield the expressions of first and second order derivatives.

Proposition 3.4 *The gradient* $\nabla H(G\|R)$ *is a linear functional* $\mathcal{T}(G) \mapsto \mathbb{R}$ *defined where the entropy* $H(G\|P)$ *is finite by:*

$$\nabla H(G\|R).\eta = \sum_x \dot\eta_x \left(R_x - \frac{\ddot g_x}{\dot g_x} \right) + \sum_{xy} \ddot\eta_{xy} \log \frac{\ddot g_{xy}}{\dot g_x R_{xy}}, \qquad \forall \eta \in \mathcal{T}(G).$$

The Hessian $\nabla^2 H(G\|R)$ *does not depend on* R*. It is the bilinear symmetric positive definite functional* $\mathcal{T}(G) \times \mathcal{T}(G) \mapsto \mathbb{R}$ *defined at every points* G *where the entropy* $H(G\|P)$ *is finite by:*

$$\begin{aligned}
\nabla^2 H(G).\eta.\zeta &= \sum_x \frac{\dot\eta_x \dot\zeta_x}{\dot g_x^2} \ddot g_x - \sum_{xy} \frac{\dot\zeta_x \ddot\eta_{xy} + \dot\eta_x \ddot\zeta_{xy}}{\dot g_x} + \sum_{xy} \frac{\ddot\eta_{xy} \ddot\zeta_{xy}}{\ddot g_{xy}} \\
&= \sum_x \frac{\dot\eta_x \dot\zeta_x}{\dot g_x^2} \ddot g_x - \sum_{x} \frac{\dot\zeta_x \ddot\eta_{x} + \dot\eta_x \ddot\zeta_{x}}{\dot g_x} + \sum_{xy} \frac{\ddot\eta_{xy} \ddot\zeta_{xy}}{\ddot g_{xy}}.
\end{aligned}$$

For $f = (\dot f, \ddot f) \in L_1(\dot g) \times L_1(\ddot g)$, we shall write $f \equiv \nabla H(G)$ if

$$\langle f, \eta \rangle \stackrel{\text{def}}{=} \sum_x \dot f(x) \dot\eta_x + \sum_{xy} \ddot f(x,y) \ddot\eta_{xy} = \nabla H(G).\eta, \qquad \forall \eta \in \mathcal{T}(G).$$

One can then deduce the following corollary from Propositions 3.3 and 3.4.

Corollary 3.5 *When* $H(G\|R)$ *is finite, the gradient is a continuous functional* $\mathcal{T}(G) \mapsto \mathbb{R}$*. Hence there exists a unique canonical representation of* $\nabla H(G)$ *in* $L_1(\dot g)/N_1 \oplus L_1(\ddot g)/N_2$ *that will also be denoted by* $\nabla H(G)$*. Moreover* $f \equiv \nabla H(G)$ *if, and only if,*

$$\exists \alpha \in \mathbb{R} : \qquad \dot f(x) = R_x - \frac{\ddot g_x}{\dot g_x} + \alpha \tag{3.1}$$

$$\exists \beta \in L_1(\ddot g_1) : \quad \ddot f(x,y) = \log \frac{\ddot g_{xy}}{\dot g_x R_{xy}} + \beta(x) - \beta(y). \tag{3.2}$$

3.5 Inverse of Hessian

Proposition 3.6 *When G has finite entropy, the Hessian $\nabla^2 H(\nu)$, considered as a linear functional from $\mathcal{T}(G)$ onto $\mathcal{T}^*(G)$ is defined as*

$$\nabla^2 H(\nu).\eta \equiv \left(\frac{\dot\eta_x \ddot g_x}{\dot g_x^2} - \frac{\ddot\eta_x}{\dot g_x}\ ,\ -\frac{\dot\eta_x}{\dot g_x} + \frac{\ddot\eta_{xy}}{\ddot g_{xy}}\right) \qquad \forall \eta \in \mathcal{T}(G). \tag{3.3}$$

The Hessian is injective and its inverse, when it exists, is given by $\nabla^2 H(G)^{-1}.f = \eta$ with

$$\frac{\dot\eta_x}{\dot g_x} = (I-\tilde P)^{-1}\ddot f_2(x) + (I-P)^{-1}\ddot f_1(x) \tag{3.4}$$
$$-G^{-1}\dot f(x) - \tilde G^{-1}\dot f(x)$$
$$\frac{\ddot\eta_{xy}}{\ddot g_{xy}} = (I-\tilde P)^{-1}\ddot f_2(x) + (I-P)^{-1}\ddot f_1(y) \tag{3.5}$$
$$-\tilde G^{-1}\dot f(x) - G^{-1}\dot f(y) + \ddot f(x,y)$$

where $P_{xy} \stackrel{\text{def}}{=} \ddot g_{xy}/\ddot g_x$ is the embedded Markov transition matrix, $\tilde P_{xy} \stackrel{\text{def}}{=} \ddot g_{yx}/\ddot g_x$ is the reversed *Markov transition matrix; $(I-P)^{-1}\ddot f_1$ and $(I-\tilde P)^{-1}\ddot f_2$ are Dynkin's transforms; $\tilde G \stackrel{\text{def}}{=} (\dot g_x, \ddot g_{yx})$ is the reverse Markov process; $G^{-1}\dot f$ and $\tilde G^{-1}\dot f$ are transforms defined by $G^{-1}\dot f \stackrel{\text{def}}{=} \int_0^\infty T(t)\dot f\,\mathrm{d}t$ where $T(t)$ is the Markov semi-group defined by G. Moreover the function f is taken centered. If it is not, it should be replaced by $\dot f - \langle \dot g, \dot f\rangle$ and $\ddot f - \langle \ddot g, \ddot f\rangle/|\ddot g|$ (see (3.10) and (3.11)).*

Proof : This result is a purely algebraic exercise, once all existence hypotheses are done. Here is the sketch of the calculations.
One search for the solution η of the problem:

$$\left(\frac{\dot\eta_x \ddot g_x}{\dot g_x^2} - \frac{\ddot\eta_x}{\dot g_x}\ ,\ -\frac{\dot\eta_x}{\dot g_x} + \frac{\ddot\eta_{xy}}{\ddot g_{xy}}\right) \equiv \left(\dot f(x), \ddot f(x,y)\right).$$

If such a solution exists, there exist α and β such that

$$\frac{\dot\eta_x \ddot g_x}{\dot g_x^2} - \frac{\ddot\eta_x}{\dot g_x} = \dot f(x) + \alpha \tag{3.6}$$
$$-\frac{\dot\eta_x}{\dot g_x} + \frac{\ddot\eta_{xy}}{\ddot g_{xy}} = \ddot f(x,y) + \beta(x) - \beta(y). \tag{3.7}$$

Combining both equations (3.6) and (3.7) in order to eliminate $\dot\eta$ and using $\ddot g_{xy}/\ddot g_x = P_{xy}$ yields:

$$-\ddot\eta_x P_{xy} + \ddot\eta_{xy} = \ddot g_{xy}\big(\ddot f(x,y) + \beta(x) - \beta(y)\big) + \dot g_x P_{xy}\big(\dot f(x) + \alpha\big). \tag{3.8}$$

Summing up (3.8) over x and y yields:

$$0 = \sum_{x,y} \ddot g_{xy}\ddot f(x,y) + \sum_x \dot g_x \dot f(x) + \alpha. \tag{3.9}$$

Hence we can split α between the double sum and the single sum so as to have each one null. This means centering both functions $\ddot f$ and $\dot f$:

$$\dot f \quad\leftarrow\quad \dot f - \sum_x \dot g_x \dot f(x), \tag{3.10}$$

$$\ddot f \quad\leftarrow\quad \ddot f - \sum_{x,y} \ddot g_{xy}\ddot f(x,y)\Big(\sum_{x,y}\ddot g_{xy}\Big)^{-1}. \tag{3.11}$$

For the sake of simplicity, we shall assume in the sequel $\ddot f$ and $\dot f$ to be centered and $\alpha = 0$.
Now, (3.8) is summed over y:

$$0 = \sum_y \ddot g_{xy}\ddot f(x,y) + \ddot g_x\big(\beta(x) - P\beta(x)\big) + \dot g_x \dot f(x),$$

which yields the expression of β:

$$\beta(x) = -(I-P)^{-1}\ddot f_1(x) + G^{-1}\dot f(x). \tag{3.12}$$

where $\ddot f_1(x) \stackrel{\text{def}}{=} \sum_y P_{xy}\ddot f(x,y)$. The inverse operator G^{-1} appears since

$$G\beta(x) = \sum_y G_{xy}\beta_y = \sum_y \frac{\ddot g_{xy}}{\dot g_x}\beta(y) - \frac{\ddot g_x}{\dot g_x}\beta(x) = -\frac{\ddot g_x}{\dot g_x}(I-P)\beta(x)$$

so that one can write, with a slight abuse of notation,

$$-(I-P)^{-1}\left(\frac{\dot g_x}{\ddot g_x}\dot f(x)\right) = G^{-1}\dot f(x). \tag{3.13}$$

Note that, for any function h, $(I-P)h$ and Gh are centered (resp. for the measures $\ddot g$ and $\dot g$). This is why it is important to reduce the problem to *centered* $\ddot f$ and $\dot f$.

Under some mild conditions (irreducibility ...) equation (3.12) possesses one, and only one, solution up to a constant (but it was already known β is defined up to a constant).
Equation (3.8) is expressed *up in time*, i.e. with the upward transition matrix P. It is in fact a time non-oriented equation[15] which can be re-written *down in time*:

$$-\ddot{g}_y \tilde{P}_{yx} \frac{\ddot{\eta}_x}{\ddot{g}_x} + \ddot{\eta}_{xy} = \ddot{g}_{xy} \big(\ddot{f}(x,y) + \beta(x) - \beta(y) \big) + \dot{g}_y \tilde{P}_{yx} \frac{\dot{g}_x}{\ddot{g}_x} \dot{f}(x). \tag{3.14}$$

Summing up (3.14) over x, identifying the $\ddot{\eta}_y$ part and simplifying (using $(I - \tilde{P})^{-1}\tilde{P} = -I + (I - \tilde{P})^{-1}$) yields:

$$\frac{\ddot{\eta}_y}{\ddot{g}_y} = (I - \tilde{P})^{-1} \ddot{f}_2(y) - \beta(y) - \tilde{G}^{-1} \dot{f}(y) - \dot{f}(y). \tag{3.15}$$

Finally, the expression (3.4) of $\dot{\eta}$ is deduced from (3.6) and (3.15); the expression (3.5) of $\ddot{\eta}$ with both variables is deduced from (3.4) combined with (3.7). The identification of the inverse is completed. ■

3.6 Martingale and variance

Assuming our theory applies in the continuous framework the same way as in the discrete one, denoting by Y_t a Markov process with generator G and by X_i the embedded process, the variance of

$$S_t \stackrel{\text{def}}{=} \int_0^t \dot{f}(Y_s)\, \mathrm{d}s + \sum_{i=1}^{N_t - 1} \ddot{f}(X_i, X_{i+1}) \tag{3.16}$$

should asymptotically behave as $\sigma^2 t$ with

$$\begin{aligned} \sigma^2 \quad & \stackrel{\text{def}}{=} \langle \nabla^2 H^{-1} f, f \rangle = \sum_{x,y} \ddot{g}_{xy} \ddot{f}^2(x,y) \\ & + \sum_x \big(\dot{g}_x \dot{f}(x) + \ddot{g}_x \ddot{f}_1(x) \big) \Big((I - \tilde{P})^{-1} \ddot{f}_2(x) - \tilde{G}^{-1} \dot{f}(x) \Big) \\ & + \sum_x \big(\dot{g}_x \dot{f}(x) + \ddot{g}_x \ddot{f}_2(x) \big) \Big((I - P)^{-1} \ddot{f}_1(x) - G^{-1} \dot{f}(x) \Big). \end{aligned} \tag{3.17}$$

[15] The symmetry is destroyed by the choice we made to write the second component of $\nabla^2 H$ asymmetrically: $-\dot{\eta}_x/\dot{g}_x + \ddot{\eta}_{xy}/\ddot{g}_{xy}$ but it could have been written $-\dot{\eta}_y/\dot{g}_y + \ddot{\eta}_{xy}/\ddot{g}_{xy}$ as well.

Note that, for a function without jump part $f = (\dot{f}, 0)$, this expression reduces to the well-known variance for the integral part of (3.16):

$$\sigma^2 = -\sum_x \dot{g}_x \dot{f}(x) \tilde{G}^{-1} \dot{f}(x) - \sum_x \dot{g}_x \dot{f}(x) G^{-1} \dot{f}(x) = -2\langle \dot{g}, fh \rangle$$

where $Gh = f$. We give here the equivalent of Theorem 2.8 with the sketch of the proof given for $d = 1$ for the sake of simplicity.

Theorem 3.7 (Central limit theorem) *Let* $\{Y_t,\ t \geq 0\}$ *be a jump Markov process with ergodic irreducible generator* G*. Let* $f : E^2 \mapsto \mathbb{R}^d$ *be a centered functional (see (3.10)–(3.11)). Conditionally to the existence of the solution of* $\nabla^2 H^{-1}.f$*, then* $S_t/\sqrt{t}$ *converges to a normal law* $\mathcal{N}(0, \Sigma)$ *where*

$$\Sigma_{ij} \quad = \quad \langle f_i, \nabla^2 H^{-1}.f_j \rangle. \tag{3.18}$$

Proof : The martingale naturally associated to (3.16) is

$$M_t \stackrel{\text{def}}{=} h(Y_t) + \int_0^t \dot{f}(Y_s)\,\mathrm{d}s + \sum_{i=1}^{N_t - 1} \ddot{f}(X_i, X_{i+1}). \tag{3.19}$$

Equation (3.19) can be translated in differential form:

$$\mathrm{d}M_t = \mathrm{d}h(Y_t) + \dot{f}(Y_t)\,\mathrm{d}t + \ddot{f}(Y_t, Y_{t+\mathrm{d}t}),$$

with the convention $\ddot{f}(x, x) = 0$. It yields

$$\mathbb{E}\left[\frac{\mathrm{d}M_t}{\mathrm{d}t}\Big|\mathcal{F}_t\right] = Gh(Y_t) + \dot{f}(Y_t) + \frac{\ddot{g}_{Y_t}}{\dot{g}_{Y_t}} \ddot{f}_1(Y_t). \tag{3.20}$$

For the expectation (3.20) to be null, one must choose (using (3.13))

$$h \stackrel{\text{def}}{=} (I - P)^{-1} \ddot{f}_1 - G^{-1} \dot{f}.$$

Then, one checks that

$$\mathrm{d}\langle M \rangle_t = \big(\mathrm{d}h(Y_t)\big)^2 + 2\,\mathrm{d}h(Y_t) \ddot{f}(Y_t, Y_{t+\mathrm{d}t}) + \ddot{f}^2(Y_t, Y_{t+\mathrm{d}t}).$$

Hence

$$\begin{aligned} \mathbb{E}\left[\frac{\mathrm{d}\langle M \rangle_t}{\mathrm{d}t}\Big|Y_t = x\right] &= Gh^2(x) - 2h(x)Gh(x) \\ &\quad + \sum_{y \neq x} G_{xy} \Big(2\big(h(y) - h(x)\big) \ddot{f}(x, y) + \ddot{f}^2(x, y)\Big). \end{aligned}$$

Since

$$\langle M\rangle_t = \int_0^t \mathbb{E}\left[\frac{\mathrm{d}\langle M\rangle_s}{\mathrm{d}s}\middle| Y_s\right] \mathrm{d}s,$$

the ergodic theorem yields, after simplification using the form of h,

$$\frac{\langle M\rangle_t}{t} \stackrel{t\to\infty}{\longrightarrow} 2\sum_x h(x)\big(\dot{g}_x\dot{f}(x) + \ddot{g}_x\ddot{f}_2(x)\big) + \sum_{x,y}\ddot{g}_{xy}\ddot{f}^2(x,y).$$

This is precisely the variance expressed in (3.17); it is simplified since the last two lines of (3.17) are equal. ■

3.7 Conclusion

It goes without saying that all remarks of Section 2 can be applied in the continuous time framework. One can developp other calculations (e.g. similar to Appendix B) and other remarks, but the present section is aimed at showing shortly that the validity of the formula $\sigma^2 = \langle\nabla^2 H^{-1}f, f\rangle$ is luckily to be wide.

Appendix A From markov derivatives to i.i.d. derivatives

A nice algebraic exercise is to find the i.i.d. derivatives solely from the Markov equations. Indeed, if we define

$$\begin{array}{rcl} A : \mathcal{M}_1(E) & \mapsto & \mathcal{M}_s(E^2) \\ \nu & \to & \nu\otimes\nu, \end{array}$$

we can express the i.i.d. entropy from the Markov entropy:

$$H(\nu\|\mu) = H(A(\nu)\|\mu),$$

where μ in the right-hand side means the transition matrix $P(x, \mathrm{d}y)$ equal to $\mu(\mathrm{d}y)$. In the sequel, we shall denote identically both entropy functions.
The derivatives of A are easily calculated: for all $\eta,\zeta \in \mathcal{T}(\nu)$,

$$\begin{array}{rcl} \nabla A.\eta & = & \nu\otimes\eta + \eta\otimes\nu, \\ \nabla^2 A.\eta.\zeta & = & \zeta\otimes\eta + \eta\otimes\zeta. \end{array}$$

Using the formulæ:

$$\nabla(H\circ A).\eta = \nabla H\circ A.(\nabla A.\eta), \tag{A.1}$$
$$\nabla^2(H\circ A).\eta.\zeta = \nabla^2 H\circ A.(\nabla A.\eta).(\nabla A.\zeta)+\nabla H\circ A.(\nabla^2 A.\eta.\zeta). \tag{A.2}$$

and the expression of the derivatives of H (see (2.4) for ∇H and (2.8) for $\nabla^2 H$), we get

$$\begin{aligned}\nabla(H\circ A(\nu)).\eta &= \int \log\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(y)\nu(\mathrm{d}x)\eta(\mathrm{d}y)+\int\log\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(y)\eta(\mathrm{d}x)\nu(\mathrm{d}y),\\ &= \int\log\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(y)\eta(\mathrm{d}y),\end{aligned}$$

because η is centered (hence $\int \mathrm{d}\eta = 0$). This is the expression, identical to equation (1.12).
The second summand in the right-hand side of (A.2) is easily proved to be null. The first summand yields

$$\begin{aligned}&\nabla^2(H\circ A(\nu)).\eta.\zeta\\ &= \int\frac{\mathrm{d}\nu\otimes\eta+\eta\otimes\nu}{\mathrm{d}\nu\otimes\nu}\,\mathrm{d}(\nu\otimes\zeta+\zeta\otimes\nu)-\int\frac{\mathrm{d}\nu+\eta}{\mathrm{d}\nu}\,\mathrm{d}\zeta\\ &= \int\frac{\mathrm{d}\eta}{\mathrm{d}\nu}(y)\zeta(\mathrm{d}y)+\int\frac{\mathrm{d}\eta}{\mathrm{d}\nu}(x)\zeta(\mathrm{d}x)-\int\frac{\mathrm{d}\eta}{\mathrm{d}\nu}\,\mathrm{d}\zeta\\ &= \int\frac{\mathrm{d}\eta}{\mathrm{d}\nu}\,\mathrm{d}\zeta.\end{aligned}$$

Once again the result is consistent with equation (1.17).

Appendix B Variance expression for stationary ergodic processes

Let $\{X_i,\ i\in\mathbb{Z}\}$ be a stationary ergodic process on a space E and f a real functional on E. It is known that, under mild hypothesis, the random variable

$$S_n \stackrel{\text{def}}{=} \sum_{i=1}^n f(X_i)\in\mathbb{R}$$

tends to a normal variable of law $\mathcal{N}(nM, n\sigma^2)$. More formally, this means that

$$\frac{S_n-nM}{\sqrt{n\sigma^2}} \stackrel{n\to\infty}{\longrightarrow} \mathcal{N}(0,1).$$

Since X is stationary, the mean M is simply expressed as

$$M = \mathbb{E}\left[f(X_0)\right].$$

We would like to get an expression for the limit variance σ^2. First, we assume that f is centered, i.e. $M = 0$, which is always possible by taking $f - M$ instead of f. Formally

$$\begin{aligned}
\sigma_n^2 &\stackrel{\text{def}}{=} \frac{1}{n}\mathbb{E}\left[\left(\sum_{i=1}^{n} f(X_i)\right)^2\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i,j=1}^{n} f(X_i)f(X_j)\right] \\
&= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\sum_{i=1-j}^{n-j} f(X_0)f(X_i)\right] \\
&\stackrel{n\to\infty}{\longrightarrow} \sigma^2 = \mathbb{E}\left[\sum_{i=-\infty}^{\infty} f(X_0)f(X_i)\right].
\end{aligned} \tag{B.1}$$

We only used the stationarity of the process for shifting the indices and the ergodicity for the convergence, assuming the existence of all quantities in the equations. An alternative expression is

$$\sigma^2 = \mathbb{E}\left[f(X_0)\left(-f(X_0) + 2\sum_{i=0}^{\infty} f(X_i)\right)\right]. \tag{B.2}$$

B.1 The i.i.d. case

Since the X_i are independant of law π, we find

$$\sigma^2 = \mathbb{E}\left[f(X_0)^2\right] = \int f^2 \,\mathrm{d}\pi.$$

B.2 The Markov case

We assume the X_i are markovian, with an irreducible transition matrix P and stationary measure π. The expression (B.2) yields

$$\begin{aligned}
\sigma^2 &= -\int f^2 \,\mathrm{d}\pi + 2\int f \sum_{i=0}^{\infty} P^{(i)} f \,\mathrm{d}\pi \\
&= -\int f^2 \,\mathrm{d}\pi + 2\int f (I-P)^{-1} f \,\mathrm{d}\pi.
\end{aligned}$$

Denoting the Dynkin's transform $(I-P)^{-1}f$ by h, i.e. the unique centered solution of the equation

$$h - Ph = f,$$

we get

$$\sigma^2 = \int f(-f+2h)\,\mathrm{d}\pi = \int (h-Ph)(h+Ph)\,\mathrm{d}\pi = \int h^2 - (Ph)^2\,\mathrm{d}\pi,$$

which is the usual expression of the variance, as obtained by martingales methods.

References

[1] Bourbaki, N. *Eléments de mathématique. XVIII. Première partie: Les structures fondamentales de l'analyse. Livre V: Espaces vectoriels topologiques. Chapitre III: Espaces d'applications linéaires continues. Chapitre IV: La dualité dans les espaces vectoriels topologiques. Chapitre V: Espaces hilbertiens.* Hermann & Cie, Paris, 1955. Actualités Sci. Ind., no. 1229.

[2] de La Fortelle, A. Large deviation principle for markov chains in continuous time. *Problemy Peredachi Informatsii 36*, 2 (2001), 120–140.

[3] Dembo, A., and Zeitouni, O. *Large deviations techniques and applications*, second ed. Springer-Verlag, New York, 1998.

[4] Fayolle, G., and de La Fortelle, A. Large deviation principle for markov chains in discrete time. Tech. Rep. 3791, INRIA, 1999.

[5] Loève, M. *Probability theory. I*, fourth ed. Springer-Verlag, New York, 1977. Graduate Texts in Mathematics, Vol. 45.

[6] Nummelin, E. *General irreducible Markov chains and nonnegative operators.* Cambridge University Press, Cambridge, 1984.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399