PANORAMAS ET SYNTHÈSES 23

ON CRAMÉR'S THEORY IN INFINITE DIMENSIONS

Raphaël Cerf

Société Mathématique de France 2007
Publié avec le concours du Centre National de la Recherche Scientifique
et du Ministère de la Culture et de la Communication

R. Cerf
Université Paris Sud, Mathématique, Bâtiment 425, 91405 Orsay Cedex–France.
E-mail : rcerf@math.u-psud.fr

2000 *Mathematics Subject Classification.* — 60F10, 46A03, 49J35.

***Key words and phrases.* —** Cramér's theory, large deviations, topological vector spaces, Minimax problems, Fenchel–Legendre transform, Mosco convergence.

ON CRAMÉR'S THEORY IN INFINITE DIMENSIONS

Raphaël Cerf

***Abstract.* —** This text is a self–contained account of Cramér's theory in infinite dimensions. Our point of view is slightly different from the classical texts of Azencott, Bahadur and Zabell, Dembo and Zeitouni, Deuschel and Stroock. We have been trying to understand the relevance of the topological hypotheses necessary to carry out the core of the theory. We have also drawn some inspiration from the analogy between the large deviation proofs in statistical mechanics and for i.i.d. random variables.

***Résumé* (La théorie de Cramér en dimension infinie). —** Ce texte est un exposé autonome de la théorie de Cramér en dimension infinie. Le point de vue est légèrement différent des textes classiques d'Azencott, de Bahadur et Zabell, de Dembo et Zeitouni, et de Deuschel et Stroock. Nous avons essayé de comprendre la pertinence des hypothèses topologiques nécessaires pour faire fonctionner le cœur de la théorie. Nous avons également exploité l'analogie entre les preuves de grandes déviations en mécanique statistique et pour des variables aléatoires i.i.d.

CONTENTS

ACKNOWLEDGEMENTS

I warmly thank Alano Ancona for his guidance in the realm of topological vector spaces. I thank Ismael Bailleul, Cathy Maugis, Pierre Petit and an anonymous referee for their many comments on this text. I thank all the students of Orsay whom I taught large deviations.

Raphaël Cerf,

Orsay, June 2007.

CHAPTER 1

INTRODUCTION

One of the most famous results in probability theory is the law of large numbers. For $n \geq 1$, let $X_1, \ldots, X_n$ be n independent identically distributed real–valued random variables with mean m, and let $\overline{S}_n$ be their empirical mean, given by

$$\overline{S}_n = \frac{1}{n}(X_1 + \cdots + X_n)\,.$$

The weak law of large numbers asserts the convergence in probability of $\overline{S}_n$ towards m:

$$\forall \varepsilon > 0 \qquad \lim_{n\to\infty} P\big(\big|\overline{S}_n - m\big| > \varepsilon\big) \;=\; 0\,.$$

One starts usually to prove the law of large numbers for square integrable or even bounded random variables. In this situation, the weak law of large numbers is a straightforward consequence of the Bienaymé–Tchebytcheff inequality:

$$\forall \varepsilon > 0 \qquad P\big(\big|\overline{S}_n - m\big| > \varepsilon\big) \;\leq\; \frac{1}{\varepsilon^2}\,\mathrm{var}\big(\overline{S}_n\big) \;=\; \frac{\sigma^2}{n\varepsilon^2}\,,$$

where $\sigma^2 = E\big((X_1 - m)^2\big)$ is the common variance of the random variables $X_1, \ldots, X_n$. This inequality controls at once the probability of all the values which are outside a neighborhood of the mean. Yet we would like to have a strategy to prove laws of large numbers which works in more complicated situations. We have mainly two directions of generalization in mind, the infinite dimensional setting and the case of strongly dependent random variables. Large deviations provide a unified language to attack these questions. Let us illustrate the strategy of a large deviation proof in the above case. For simplicity, we consider the case where the random variables take their values in $[0, 1]$ and their common expectation is $1/2$. For $n \geq 1$, the empirical mean $\overline{S}_n$ takes its values in the compact space $[0, 1]$. For any $x \in [0, 1]$, we will estimate the asymptotics of the probability that $\overline{S}_n$ belongs to a neighborhood of x. It will turn out that, except when x is equal to the mean $1/2$, the probability $P(x - \varepsilon < \overline{S}_n < x + \varepsilon)$ decays exponentially fast like $\exp -cn$ (with $c > 0$ when ε is sufficiently small) as n goes to ∞. Since $[0, 1]$ is compact and since $\overline{S}_n$ has to be somewhere in $[0, 1]$, it will

be close to $1/2$ with probability going exponentially fast to 1 as n goes to ∞. Let us be more precise. For any interval U included in $[0,1]$, for $n, m \geq 1$, using successively the convexity of U and the fact that $X_1, \ldots, X_{n+m}$ are independent and identically distributed, we have

$$\begin{aligned} P(\overline{S}_{n+m} \in U) \;&\geq\; P\Big(\frac{1}{n}\big(X_1 + \cdots + X_n\big) \in U,\, \frac{1}{m}\big(X_{n+1} + \cdots + X_{n+m}\big) \in U\Big) \\ &\geq\; P\Big(\frac{1}{n}\big(X_1 + \cdots + X_n\big) \in U\Big) P\Big(\frac{1}{m}\big(X_{n+1} + \cdots + X_{n+m}\big) \in U\Big) \\ &=\; P(\overline{S}_n \in U)\, P(\overline{S}_m \in U)\,. \end{aligned}$$

Therefore the sequence $-\ln P(\overline{S}_n \in U)$, $n \geq 1$, is subadditive. By the famous subadditive lemma, the limit

$$\lim_{n\to\infty} \; -\frac{1}{n} \ln P\big(\overline{S}_n \in U\big)$$

exists for any interval U included in $[0,1]$. The Cramér transform of the common law of $X_1, \ldots, X_n$ is the map $I : [0,1] \to [0, +\infty]$ defined by

$$\forall x \in [0,1] \qquad I(x) \;=\; \sup_{\varepsilon > 0} \; \lim_{n\to\infty} \; -\frac{1}{n} \ln P(x - \varepsilon < \overline{S}_n < x + \varepsilon)\,.$$

A remarkable feature of Cramér's theory is that I can be explicitly identified: it is the Fenchel–Legendre transform of the Log–Laplace of the law of X_1. The map I is the rate function governing the large deviations of the empirical mean $\overline{S}_n$. The constant c characterizing the exponential decay of $P(x - \varepsilon < \overline{S}_n < x + \varepsilon)$ is equal to $I(x)$; roughly speaking, the large deviations principle says that, for E a subset of $[0,1]$,

$$P\big(\overline{S}_n \in E\big) \underset{n\to\infty}{\overset{\ln}{\sim}} \exp\Big(- n \inf\big\{\, I(x) : x \in E \,\big\}\Big)\,.$$

Large deviations provide a bridge between probability theory and the calculus of variations: the estimation of the probability $P\big(\overline{S}_n \in E\big)$ boils down to the study of the variational problem $\inf\big\{\, I(x) : x \in E \,\big\}$. The hard part of the work is now to prove that I vanishes only at $1/2$. Indeed, if this was proved, we would have then $\inf_{[0,1]\setminus U} I > 0$ for any neighborhood U of $1/2$, whence, by the previous principle,

$$\lim_{n\to\infty} \; P\big(\overline{S}_n \in U\big) \;=\; 1\,.$$

This way, we recover not only the weak law of large numbers, but we obtain also the correct speed of decay of the probability $P\big(\overline{S}_n \notin U\big)$, which is of order $\exp -cn$ (with $c > 0$). The understanding of the structure of the minima of I requires probabilistic estimates specific to the model under study. The scheme of proof we have described is quite robust and can be considerably generalized in the two directions alluded before. It is a quite natural desire to try to generalize Cramér's theory in infinite dimensions. Beyond the mere pleasure of getting abstract and elegant formulations, such generalizations do have very useful applications. Let us mention two of them.

Cramér's theorem in a separable Banach space. — This sounds still quite abstract. Yet this result yields for instance a large deviation principle for the Minkowski average of random sets in finite dimensional spaces [**Cer99**], or for random functions [**Ter06**]. This result yields also a large deviation principle for Gaussian measures and in particular Schilder's theorem for the Brownian motion (see chapter 22). Schilder's theorem is the very starting point for the beautiful Freidlin–Wentzell theory of random perturbations of dynamical systems [**FW84**].

Sanov's theorem. — A central problem in statistics is to estimate an unknown law μ from the observation of an i.i.d. n–sample $X_1, \dots, X_n$. A natural try is to consider the empirical measure M_n defined by

$$M_n = \frac{1}{n}(\delta_{X_1} + \cdots + \delta_{X_n})\,.$$

Even if the law μ is defined on a finite dimensional space, the above measure belongs to the infinite dimensional space $\mathcal{M}(E)$ of the signed measures on E. Sanov's theorem is a large deviation principle for the random measure M_n, which holds in great generality. This large deviation principle can be obtained as an application of the general Cramér theory.

However, the most challenging direction is to leave the independent framework. To introduce stochastic dependence, we start with i.i.d. random variables and we define a new joint law with the help of a density factor.

1.1. The Ising Curie–Weiss model

Let us consider the state space $\Omega_n = \{-1, +1\}^n$. We define the following random variables on Ω_n: for $\omega = (\omega_1, \dots, \omega_n) \in \Omega_n$, we set

$$\forall i \in \{\,1, \dots, n\,\} \qquad X_i(\omega) = \omega_i$$

and we consider the empirical mean

$$\overline{S}_n = \frac{1}{n}(X_1 + \cdots + X_n)\,.$$

Let $T > 0$. We define a probability measure $\mu_{n,T}$ on Ω_n by

$$\begin{aligned} \forall \omega \in \Omega_n \qquad \mu_{n,T}(\omega) &= \frac{1}{Z_{n,T}} \exp\Big(-\frac{1}{nT} \sum_{1 \le i,j \le n} X_i(\omega) X_j(\omega) \Big) \\ &= \frac{1}{Z_{n,T}} \exp\Big(-\frac{n}{T} \overline{S}_n(\omega)^2 \Big)\,. \end{aligned}$$

Here $Z_{n,T}$ is the normalizing factor which ensures that $\mu_{n,T}(\Omega_n) = 1$. The measure $\mu_{n,+\infty}$ corresponding to $T = +\infty$ is simply the symmetric Bernoulli product measure on Ω_n, or equivalently the uniform law on Ω_n. The density of $\mu_{n,T}$ with respect to

$\mu_{n,+\infty}$ is a function of $\overline{S}_n$. By the classical Cramér theorem, the law of $\overline{S}_n$ under $\mu_{n,+\infty}$ satisfies a large deviation principle governed by the rate function I given by

$$\forall x \in [-1,1] \qquad I(x) \,=\, -\frac{1-x}{2}\ln(1-x) - \frac{1+x}{2}\ln(1+x)\,.$$

Varadhan's lemma implies then that the law of $\overline{S}_n$ under $\mu_{n,T}$ satisfies a large deviation principle governed by the rate function J given by

$$\forall x \in [-1,1] \qquad J(x) \,=\, -\frac{x^2}{T} + I(x) - \inf_{y\in[-1,1]}\Big(-\frac{y^2}{T} + I(y)\Big)\,.$$

Now, there exists a critical value $T_c \in]0,+\infty[$ such that:
• For $T \geq T_c$, the function J has a unique global minimum at $m^* = 0$.
• For $T < T_c$, the function J has two global minima at $-m^*$ and m^*, where $m^* > 0$.
The large deviation principle implies that

$$\forall \varepsilon > 0 \qquad \lim_{n\to\infty} P\big(\big||\overline{S}_n| - m^*\big| > \varepsilon\big) \,=\, 0\,.$$

Whenever $T \geq T_c$ and $m^* = 0$, the law of $\overline{S}_n$ converges towards δ_0, the Dirac mass at 0, and we have a weak law of large numbers very similar to the i.i.d. case. Let us look more closely at the case $T < T_c$. The system being symmetric under sign reversal, we conclude that

$$\forall \varepsilon > 0 \qquad \lim_{n\to\infty} P\big(\big|\overline{S}_n - m^*\big| \leq \varepsilon\big) \,=\, \lim_{n\to\infty} P\big(\big|\overline{S}_n + m^*\big| \leq \varepsilon\big) \,=\, \frac{1}{2}\,.$$

Therefore the law of $\overline{S}_n$ converges towards $\frac{1}{2}(\delta_{-m^*} + \delta_{m^*})$. This is a fundamentally new type of law of large numbers, where the limit is random.

1.2. The nearest–neighbour Ising model

To incorporate a more complex dependence between the random variables, we put a geometric structure by attaching the random variables to the d–dimensional lattice $\mathbb{Z}^d$. Let $\Lambda \subset \mathbb{Z}^d$ be a cubic box. A configuration in Λ is a map $\sigma : \Lambda \to \{-1,+1\}$. The energy or Hamiltonian $H^+_\Lambda(\sigma)$ of the configuration σ in Λ with plus boundary conditions is

$$H^+_\Lambda(\sigma) \,=\, -\frac{1}{2}\sum_{x,y\in\Lambda,|x-y|=1}\sigma(x)\sigma(y) \,-\sum_{x\in\Lambda,y\notin\Lambda,|x-y|=1}\sigma(x)\,.$$

Let $T > 0$ be the temperature. The Ising Gibbs measure $\mu^+_{\Lambda,T}$ in Λ at temperature T with plus boundary conditions is given by

$$\forall \sigma \in \{-1,+1\}^\Lambda \qquad \mu^+_{\Lambda,T}(\sigma) = \frac{\exp -\dfrac{H^+_\Lambda(\sigma)}{T}}{\displaystyle\sum_{\eta\in\{-1,+1\}^\Lambda}\exp -\frac{H^+_\Lambda(\eta)}{T}}\,.$$

The most likely configurations are those having a small energy, i.e., those for which the contacts between the minuses and the pluses are reduced. Thus we have built a complex probability law with strong spatial correlations. For $n \geq 1$, let

$$\Lambda(n) \,=\, \mathbb{Z}^d \cap]-n/2, n/2]^d .$$

We consider the empirical average

$$\overline{M}(\Lambda(n)) \,=\, \frac{1}{|\Lambda(n)|} \sum_{x \in \Lambda(n)} \sigma(x) .$$

A subadditive argument shows that for any $T > 0$, the following limit exists:

$$\lim_{n \to \infty} \mu^+_{\Lambda(n),T}(\overline{M}(\Lambda(n))) \,=\, m^*(T) .$$

The quantity $m^*(T)$ is called the spontaneous magnetization at temperature T. This terminology stems from the fact that the Ising model was originally introduced as a model of ferromagnetism (under some adequate conditions, a magnet submitted to the influence of a magnetic field will remember the sign of the field even after it has disappeared). We say that there is a phase transition at temperature T if $m^*(T) > 0$. In any dimension $d \geq 2$, there exists a positive and finite critical temperature $T_c(d)$ such that the Ising model exhibits a phase transition for $T < T_c(d)$ and it does not for $T > T_c(d)$.

Let us now examine the large deviation behavior of $\overline{M}(\Lambda(n))$. For any $T > 0$ and $\alpha \in [-1, 1]$, a subadditive argument yields the existence of the limit

$$J(\alpha) \,=\, \lim_{n \to \infty} \, -\frac{1}{n^d} \, \ln \mu^+_{\Lambda(n),T}(\overline{M}(\Lambda(n)) \geq \alpha) .$$

The sequence of the laws of $\overline{M}(\Lambda(n))$ under $\mu^+_{\Lambda(n),T}$, $n \geq 1$, satisfies a large deviation principle with speed n^d and governed by the good rate function $x \in [-1, 1] \mapsto J(|x|)$. This rate function vanishes on $[-m^*, m^*]$ and it is positive on $[-1, -m^*[\cup]m^*, 1]$. In this context, the analog of the Log–Laplace in Cramér's theory is the pressure p, defined as the following subadditive limit:

$$\forall t \in \mathbb{R} \qquad p(t) \,=\, \lim_{n \to \infty} \, \frac{1}{n^d} \, \ln \mu^+_{\Lambda(n),T}\Big(\exp\big(t n^d \overline{M}(\Lambda(n))\big)\Big) .$$

Exactly as in Cramér's theory, the rate function I is the Fenchel–Legendre transform of p. The previous large deviation principle allows to conclude that

$$\forall \varepsilon > 0 \qquad \lim_{n \to \infty} \mu^+_{\Lambda(n),T}\big(|\overline{M}(\Lambda(n))| \geq m^* + \varepsilon\big) \,=\, 0 .$$

For $T > T_c$, we have $m^* = 0$, the rate function I has a unique global minimum at 0, and we obtain therefore a weak law of large numbers as in the i.i.d. case. The situation is much more complex for $T < T_c$. The behavior of $\overline{M}(\Lambda(n))$ in the interval $[-m^*, m^*]$ is correctly described by a surface large deviation principle. For $m < m^*$, close to m^*, one has

$$\lim_{n \to \infty} \, \frac{1}{n^{d-1}} \ln \mu^+_{\Lambda(n),T}\big(\overline{M}(\Lambda(n)) \leq m\big) = \, -c\Big(\frac{m^* - m}{2m^*}\Big)^{\frac{d-1}{d}} ,$$

where $c = c(d, T)$ is a positive constant depending on the dimension and the temperature. The proof of such a result is rather delicate, still subadditivity plays a key role in it. However, this is quite another story (see [**Cer06**]), and our main concern in this text is the study of sequences of i.i.d. random variables in the infinite dimensional setting.

1.3. Overview of the text

This text is a self–contained account of Cramér's theory in infinite dimensions. It is mainly based on the classical texts by Azencott [**Aze80**], Bahadur and Zabell [**BZ79**], Dembo and Zeitouni [**DZ98**], Deuschel and Stroock [**DS89**]. However the order of our presentation is slightly different. In fact, we focus on the infinite dimensional setting and we try to understand the relevance of the topological hypotheses necessary to carry out the core of the theory; to this end we make appeal to various tools of functional analysis in topological vector spaces. In particular, minimax results pop up at several key places throughout the text. Another motivation is to push further the analogy between the large deviation proofs in statistical mechanics and for i.i.d. random variables. The subadditive argument was initially imported by Lanford [**Lan73**] from statistical mechanics into Cramér's theory and it sheds a new light on the i.i.d. case. When performing this argument, we systematically use the Minkowski functional and this way we get rid of some topological hypotheses appearing in [**DS89, DZ98**]: namely we do not need to work in a Polish space.

Conversely, large deviation techniques constitute a precious guideline in the study of phase coexistence [**Cer00**]. In the percolation context, the usual large deviation upper bound could not be proven because of a lack of compactness, instead it was replaced by an enhanced upper bound. We provide here in the general setting of Cramér's theory a similar enhanced upper bound. When adapting this idea to the finite dimensional situation, we obtain an apparently new upper bound valid for any probability measure in $\mathbb{R}^d$. In the case of the real line, we have written a proof of Cramér's theorem relying entirely on subadditivity. Not only is it instructive, but we gather slightly more information than with the standard proof. Moreover, the proof of the volume large deviation principle for percolation and Ising models looks like a twin sister of it [**Cer06**].

To pursue further the similarity in the general case, we separate the issue of proving the existence of a rate function and the problem of its identification. In statistical mechanics, one is usually unable to provide an operational description of the rate function I (one notable exception is the two dimensional Ising model, thanks to Onsager's computation [**Mes04**]), while in the i.i.d. case with law μ the rate function I, called the Cramér transform of μ, coincides with the Fenchel–Legendre transform Λ^* of the Log–Laplace of μ. The inequality $I \geq \Lambda^*$ holds in full generality. In this text we

explore various conditions and mechanisms ensuring the converse inequality $I \leq \Lambda^*$ and thus the equality $I = \Lambda^*$.

• For discrete measures, the equality $I = \Lambda^*$ is a consequence of a rough version of Stirling's formula and some simple probabilistic estimates.

• For smooth measures in $\mathbb{R}^d$ having finite exponential moments of any order, the equality $I = \Lambda^*$ can be obtained with the help of calculus techniques.

>From these two previous simple cases, we perform a rather delicate density argument and we extend the equality $I = \Lambda^*$ to any Borel measure μ on a finite dimensional vector space. To this end, we examine the regularity of I and Λ^* as functions of the initial measure μ, a point of view which does not seem to have been exploited in a systematic fashion in previous expositions. Whenever the topology of the vector space can be suitably approached by finite dimensional topologies, this equality can be lifted to the infinite dimensional setting. This is the case for weak topologies.

By adding a geometric condition on the space, called convex regularity, we obtain an adequate upper bound on I and we prove again that $I = \Lambda^*$. Convex regularity occurs for instance in a separable Banach space or in the dual space of a Banach space. This leads to a natural generalization of the classical Cramér theorem in $\mathbb{R}^d$. If μ is a probability measure on the cylinder σ–field of the dual space of a Banach space and if its Log–Laplace is finite in a neighborhood of the origin, the empirical mean of an i.i.d. sample of μ satisfies the full large deviation principle.

We give several versions of Cramér's theorem in different contexts: in $\mathbb{R}$, in $\mathbb{R}^d$, in the weak topology, in the dual of a Banach space and finally in a separable Banach space, following the presentation of De Acosta [**DA85**] and Gao [**Gao97**]. As in the classical expositions, we apply this last theorem to the case of the Gaussian measures and this way we recover Schilder's theorem.

In the last part, we discuss the relationship between Cramér's theorem and Sanov's theorem. We first provide an autonomous derivation of Sanov's theorem, due to Groeneboom, Oosterhoff and Ruymgaart [**GOR79**]. We show how the most general Sanov theorem is implied by our version of Cramér's theorem in a dual space. In the classical exposition [**DS89**], Cramér's theorem was shown to imply Sanov's theorem for the weak topology. In [**DZ98**], Sanov's theorem for the stronger τ–topology is derived via an apparently different method involving projective limits. Hence the two approaches are reconciled here. We show also how Cramér's theorem can be recovered from Sanov's theorem in a compact situation.

A simple technique to get an upper bound on I is to condition the n sample to be in a fixed compact. With this technique, we prove again in the last chapter the equality $I = \Lambda^*$ for any probability measure μ on a separable Banach space. A key point is the upper semicontinuity of Λ^* as a function of the measure. The argument is a variant of the proof of the celebrated Mosco convergence theorem [**Mos71, Zab92b**].