Cramér's Theorem in Banach Spaces

Peter Eichelsbacher
Institut für Angewandte Mathematik
Universität Zürich
Winterthurer-Str. 190
CH-8057 Zürich

Abstract

In this note we point out that an application of a special contraction principle combined with a general representation of large deviation rates provides a new proof of the classical result of Cramér for Banach–valued random variables.

AMS Subject Classification (1992): 60F10

Keywords: large deviations $*$ rate function – representation $*$ contraction principle

1 Introduction

The asymptotic computation of small probabilities on an exponential scale is what large deviation theory analyses: consider a normal topological space X endowed with its Borel σ-field $\mathcal{B}(X)$ and let $(P_n)_n$ be a sequence of probability measures on X. A large deviation principle (LDP) characterizes the limiting behavior as $n \to \infty$ of the sequence $(P_n)_n$ in terms of a rate function. $(P_n)_n$ satisfies a LDP on X with rate function $I : X \to [0, \infty)$ if for all $A \in \mathcal{B}(X)$:

$$- \inf_{x \in A^o} I(x) \le \liminf_{n \to \infty} \frac{1}{n} \log P_n(A) \le \limsup_{n \to \infty} \frac{1}{n} \log P_n(A) \le - \inf_{x \in \bar{A}} I(x), \qquad (1)$$

where A^0 denotes the interior of A and $\bar{A}$ the closure of A. By definition a rate function has compact level sets $N_I(L) := \{x : I(x) \le L\}$ for all $L \in [0, \infty)$. After establishing the existence of a LDP we try to use one of various approaches for the identification of the resulting rate function. A very general observation is the following: If $(P_n)_n$ satisfies a LDP the rate function I can be represented as

$$I(x) = \sup\{f(x) - \Phi(f), f \in C_b(X)\}, \tag{2}$$

where $C_b(X)$ denotes the Banach space of all bounded continuous functions $f : X \to \mathbb{R}$ with the supremum norm and

$$\Phi(f) := \lim_{n\to\infty} \frac{1}{n} \log \int_X \exp(nf(x)) dP_n(x) \tag{3}$$

(see Dinwoodie [9]). We obtain that I takes a form analogous to a convex conjugate. Dinwoodie applied this identification to the empirical law of a Markov chain (see [10]).

The aim of this note is to improve a rate representation given in [12], where the class $C_b(X)$ is replaced by an arbitrary seperating class of the space of all continuous, real-valued functions. This provides a new proof of the classical result of Cramér for Banach–valued random variables.

2 Rate representation

First we describe two main tools in large deviation theory:
The first one is Varadhan's integral lemma which actually is one of the starting points for developing the large deviation theory. Suppose $(P_n)_n$ satisfies a LDP with rate I and let $f : X \to \mathbb{R}$ be continuous. Assume further the tail condition

$$\lim_{L\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log \int_{\{f(x)\geq L\}} \exp(nf(x)) dP_n(x) = -\infty,$$

then

$$\Phi(f) = \sup\{f(x) - I(x), x \in X\}.$$

The second is Bryc's inverse of Varadhan's Laplace method (see [5]) which he told the asymptotic value approach: Let $\mathcal{F}$ be an arbitrary separating class of $C(X)$, the space of all continuous, real-valued functions: $\mathcal{F}$ contains the constant functions, is closed under finite pointwise minima and separates points of X, e.g. for all $x, y \in X$ and for all $a, b \in \mathbb{R}$

there exists a $f \in \mathcal{F}$ such that $f(x) = a$ and $f(y) = b$. Bryc proves that if $\Phi(f)$ exists for each $f \in \mathcal{F}$ of such a class and if most of the probability mass of $(P_n)_n$ is concentrated on compact sets, the sequence $(P_n)_n$ satisfies a LDP with rate (2). The concentration property is defined by: let $(P_n)_n$ be an exponentially tight sequence on X, that is for every $L > 0$ there exists a compact set $K_L \subset X$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \log P_n(K_L^c) \leq -L,$$

where K_L^c denotes the complement of K_L.

Moreover, if for each $f \in \mathcal{F}$ there is $0 < a < 1$ and a measurable function F such that $f \leq aF$ and

$$\sup_n \left[\int_X \exp(nF(x))dP_n(x)\right]^{\frac{1}{n}} < \infty, \tag{4}$$

he verified the identification

$$I(x) = \sup\{f(x) - \Phi(f), f \in \mathcal{F}\}. \tag{5}$$

So combining these tools we get a representation as we strive for: apply for every $f \in \mathcal{F}$ Varadhan's lemma by using the additional assumption (4) and conclude as Bryc does.

But this kind of proof seems to be a diversion, becauce the inverse lemma is an approach to large deviations whereas we will start with a LDP. Of course, Dinwoodie proved (2) in a more elegant fashion. He required Varadhan's integral lemma only. In a direct proof of (5) we can dropp the additional assumption (4), which is shown in [12]. Furthermore we give a strengthening of this result since we get rid of the requirement that $\Phi(f) < \infty$ for all $f \in \mathcal{F}$.

Theorem 2.1 *Suppose that the sequence $(P_n)_n$ satisfies a LDP on X, a normal space, with rate function I. If $(P_n)_n$ is exponentially tight the rate function I is given by*

$$I(x) = \sup\{f(x) - \Phi(f); f \in \mathcal{F}\}. \tag{6}$$

for every separating class $\mathcal{F}$.

Remark 2.2 *If for example X is Polish or locally compact, the exponential tightness condition is a conclusion of the LDP-property of $(P_n)_n$.*

Proof: See [12]. Replace $\mathcal{F}$ by $\mathcal{F}_b$, the class of all bounded above functions in $\mathcal{F}$. Since $\Phi(f \wedge L) \leq \Phi(f)$ for every constant $L < \infty$, it follows

$$\sup_{f\in\mathcal{F}_b}\{f(x) - \Phi(f)\} = \sup_{f\in\mathcal{F}}\{f(x) - \Phi(f)\}.$$

Thus for $f \in \mathcal{F}_b$ Varadhan's integral lemma can be applied and thus the proof in [12] works with this little modification. For the sake of completeness we give the proof:
With $f \in \mathcal{F}_b$ we have

$$\Phi(f) = \sup_{x\in X}\{f(x) - I(x)\} = -\inf_{x\in X}\{I(x) - f(x)\}.$$

An immediate consequence is that

$$I(x) \geq \sup_{f\in\mathcal{F}_b}\{f(x) - \Phi(f)\}.$$

Thus the Theorem is proved if we can verify the converse inequality $I(x) \leq \sup_{f\in\mathcal{F}_b}\{f(x) - \Phi(f)\}$. Without loss of generality we can assume $I(x) > 0$, let $\alpha < I(x)$ and choose $\delta > 0$ such that $I(y) > \alpha$ for each $y \in B_{x,2\delta} := \{y : d(y,x) \leq 2\delta\}$ (I is semi–continuous). For a fixed $N \geq 0$ let g_N be a continuous bounded function defined on X such that $g_N(y) = 0$ in $B_{x,\delta}$, $g_N(y) = -N$ for all $y : d(y,x) \geq 2\delta$ and $-N \leq g_N(y) \leq 0$ for $\delta < d(y,x) < 2\delta$. Then by a simple well known calculation using the upper bound in LDP for closed sets and Varadhan's integral lemma we have:

$$\Phi(g_N) \leq -\min\{I(B_{x,2\delta}), N\}. \tag{7}$$

Since $(P_n)_n$ is exponentially tight we can choose for a fixed $M > 0$ a compact set $\hat{K}$ such that $\sup_n \frac{1}{n}\log P_n(K^c) \leq -M$. Denote by $K := \hat{K} \cup \{x\}$. For all measurable functions $h \leq M$ we get by [5]

$$\lim_{n\to\infty}\left(\frac{1}{n}\log\int \exp(nh)dP_n - \frac{1}{n}\log\int_K \exp(nh)dP_n\right) = 0. \tag{8}$$

Since $\mathcal{F}_b$ is a separating class too, there exists for any $\varepsilon > 0$ a finite collection $\{g_i\}_{1\leq i\leq n}$ of functions in $\mathcal{F}_b$ such that

$$\sup_{y\in K}|g_N(y) - \max_{i=1,\dots,n} g_i(y)| \leq \varepsilon. \tag{9}$$

It is well known that

$$\Phi(\max(g_1, \dots, g_n)) = \max(\Phi(g_1), \dots, \Phi(g_n)). \tag{10}$$

Define $\Phi_A(f) := \frac{1}{n} \log \int_A \exp(nf(x)) dP_n(x)$ for suitable, measurable $f : X \to \mathbb{R}$. Then for each j

$$\begin{aligned} \Phi(g_N) &\geq \limsup_{n\to\infty} \Phi_K(g_N) \geq \liminf_{n\to\infty} \Phi_K(g_N) \\ &\geq \liminf_{n\to\infty} \Phi_K(\max_i g_i) - \varepsilon = \max_i \Phi(g_i) - \varepsilon \geq \Phi(g_j) - \varepsilon. \end{aligned}$$

The first inequality stems from an elementary fact, see Exercise 2.1.24 in [8], furthermore we have used (9). The equality is a consequence of (8) and (10). Thus using (7) we obtain

$$\min\{I(B_{x,2\delta}), N\} \leq -\Phi(g_j) + \varepsilon \leq -\Phi(g_j) + g_j(x) + 2\varepsilon,$$

since $-g_j(x) \leq \varepsilon$, if we choose j such that $\max_i g_i(x) = g_j(x)$. Now if $I(B_{x,2\delta}) = \infty$ we get

$$N \leq g_j(x) - \Phi(g_j) + 2\varepsilon.$$

and hence

$$I(x) = \infty = \sup_{f \in \mathcal{F}_b} \{f(x) - \Phi(f)\}.$$

Otherwise choose $N > I(B_{x,2\delta})$ we obtain

$$0 < \alpha < I(B_{x,2\delta}) \leq g_j(x) - \Phi(g_j) + 2\varepsilon$$

which proves the result, since $\alpha > 0$ and $\varepsilon > 0$ were choosen arbitrarily. $\square$

Theorem 2.1 and a well known contraction principle are the main tools of a new proof of Cramér's Theorem for Banach-valued random variables: Let μ be a probability measure on a separable Banach space X ($\mu \in \mathcal{M}^1(X)$), such that

$$\int_X \exp(\sigma||x||)\mu(dx) < \infty \tag{11}$$

for all $\sigma > 0$. Let $X_1, \dots, X_n$ be a sequence of i.i.d. random variables taking values in X with each X_i distributed according to μ. Let μ_n be the distribution of $S_n := \frac{1}{n}(X_1 + \cdots + X_n)$ and $P := \otimes_{\mathbb{N}} \mu$.

Theorem 2.3 *Let assumption (11) hold. Then $(\mu_n)_n$ satisfies a LDP with rate*

$$\sup_{\lambda \in X^*} \{< \lambda, x > - \Lambda_\mu(\lambda)\} \tag{12}$$

(Cramér-rate), where X^* is the dual space of X namely the space of all continuous linear functionals of X and

$$\Lambda_\mu(\lambda) := \log \int_X \exp[< \lambda, x >] d\mu(x)$$

($< \lambda, x >$ denotes the value of $\lambda(x) \in \mathbb{R}$).

Remark 2.4 *Other approaches to Cramér's Theorem are well known. For example the combination of sub-additivity methods and the so called projective limit approach (discribed in [7], Chapter 6) traced back to Ruelle [14] and Lanford [13] and Dawson and Gärtner [6]. Donsker and Varadhan [11] used a type of contraction method related to the special subject of their interest. Their approach was also considered by Bolthausen [3, 4].*

3 Proof

Let (Y, d) be a metric space.

Definition 3.1 *For each $n, m \in \mathbb{N}$ let $(\Omega, \mathcal{B}_n, P_{n,m})$ be a probability space and let Y_n and $Y_{n,m}$ be Y-valued random variables distributed according to the joint law $P_{n,m}$ with marginals ν_n and $\nu_{n,m}$. $(Y_{n,m})_{n,m}$ are called exponentially good approximations of $(Y_n)_n$ if for every $\delta > 0$ the set $\{(Y_n, Y_{n,m}) \in A_\delta\}$ is $\mathcal{B}_n$ measurable and*

$$\lim_{m\to\infty} \lim_{n\to\infty} \frac{1}{n} \log P_{n,m}(A_\delta) = -\infty$$

where $A_\delta := \{(x, y) : d(x, y) > \delta\}$. The measures $(\nu_{n,m})_{n,m}$ are exponentially good approximations of $(\nu_n)_n$ if one can construct the spaces as above.

The following Theorem provides the contraction principle which we will use in the proof:

Theorem 3.2 *(Dembo,Zeitouni [7]) Let $(P_n)_n$ be a sequence of probability measures on X that satisfies a LDP with rate function I. For $m = 1, 2, \ldots$, let $f_m : X \to Y$ be continuous functions. Assume there exists a measurable map $f : X \to Y$ such that*

$$\limsup_{m\to\infty} \sup\{\, d(f_m(x), f(x)) : x \in X, I(x) \le C \,\} = 0 \tag{13}$$

for all $C \in (0, \infty)$. Then any sequence of probability measures $(Q_n)_n$ for which $(P_n \circ f_m^{-1})_{n,m}$ are exponentially good approximations satisfies the LDP in Y with rate function $J(y) := \inf\{I(x) : x \in X, y = f(x)\}$.

Remarks 3.3

1. *Condition (13) implies that the contraction map f is continuous on the level sets of I.*
2. *Barbe [1] gave some refinements of the contraction principle. He pointed out that the exponentially good approximation property is sufficient to get the lower-bound in the LDP for $(P_n \circ f^{-1})_n$.*

Proof of Theorem 2.3: Define the empirical measure of the i.i.d random variables $X_1, \ldots, X_n$:

$$L_n(\omega, B) := \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j(\omega)}(B); \qquad B \in \mathcal{B}(X),$$

then $S_n = F(L_n)$ where $F : \mathcal{M}^1(X) \to X$ is given by $F(\varrho) := \int x d\varrho$. The distribution of L_n satisfies a LDP with rate $H(\cdot, \mu)$, the relative entropy, even if $\mathcal{M}^1(X)$ is equipped with the τ-topology that is the coarsest topology for which the map $Q \to Q(A), Q \in \mathcal{M}^1(X)$, is continuous for all $A \in \mathcal{B}(X)$. Here

$$H(\varrho, \mu) := \begin{cases} \int f \log f d\mu & : \quad \varrho << \mu, f := \frac{d\varrho}{d\mu} \\ +\infty & : \quad \text{otherwise.} \end{cases}$$

This is the so called Sanov Theorem. For a proof see for example [8], Theorem 3.2.21.
F is not continuous, so we try to apply Theorem 3.2 by a truncation argument:
Let $x_L = x$ if $||x|| \le L$, otherwise 0, and $F_L(\varrho) := \int x_L d\varrho$. F_L is continuous in the τ-topology and hence we will prove that F_L converges sufficiently quickly to F in the sense

of Theorem 3.2:

To check the continuity of F on the level sets of $H(\cdot,\mu)$ we use a well known result of Donsker and Varadhan [11], Lemma 5.1: If $(\varrho_n)_n$ is a sequence in $\mathcal{M}^1(X)$ such that $H(\varrho_n,\mu)\leq C$ for all n, then if condition (11) holds it follows:

$$\lim_{L\to\infty}\sup_n\int_{||x||\geq L}||x||d\varrho_n(x)=0.$$

For the sake of completeness we will give the proof of this result at the end of this section. Hence, (13) is checked: take any level set N of $H(\cdot,\mu)$ we have

$$\lim_{L\to\infty}\sup_{\varrho\in N}||\int(x-x_L)d\varrho||=0.$$

Now Theorem 3.2 told us that any sequence $(Q_n)_n$ for which $(P\circ F_L(L_n)^{-1})_n$ are exponentially good approximations satisfies a LDP with rate

$$I(x)=\inf_{\varrho:F(\varrho)=x}H(\varrho,\mu).$$

Thus we have to show that $(P\circ S_n^{-1})_n$ is an exponentially good approximation.

We have to check that for every $\delta>0$:

$$\lim_{L\to\infty}\limsup_{n\to\infty}\frac{1}{n}\log P(||F_L(L_n)-F(L_n)||\geq\delta)=-\infty.$$

Now

$$||F_L(L_n)-F(L_n)||=\frac{1}{n}||Y_1+\cdots+Y_n||$$

with

$$Y_j=\begin{cases}X_j & : \ ||X_j||>L\\ 0 & : \ ||X_j||\leq L\end{cases}.$$

Thus for every $\delta>0$ and $\theta>0$:

$$P(||F_L(L_n)-F(L_n)||\geq\delta)\leq e^{-\theta\delta}E(\exp(\tfrac{\theta}{n}\textstyle\sum_{j=1}^n||Y_j||)).$$

The sum consists of independent summands and therefore, substituting θ by $n\theta$, we obtain

$$\limsup_{n\to\infty}\frac{1}{n}\log P(||F_L(L_n)-F(L_n)||\geq\delta)\leq-\theta\delta+\log E(\exp(\theta||Y_j||))$$

and by definition of Y_j and since the inequality is true for all $\theta>0$ we conclude that

$$\lim_{L\to\infty}\limsup_{n\to\infty}\frac{1}{n}\log P(||F_L(L_n)-F(L_n)||\geq\delta)=-\infty.$$

Therefore with Theorem 2.1 and Theorem 3.2 it follows that $(P \circ S_n^{-1})_n$ satisfies a LDP with rate

$$I(x) = \inf_{\varrho : F(\varrho) = x} H(\varrho, \mu) = \sup\{f(x) - \Phi(f), f \in \mathcal{F}\}.$$

Define the separating class

$$\mathcal{F} := \{g : g(x) = \min_i \{\lambda_i(x) + c_i\}; c_i \in \mathbb{R}, \lambda_i \in X^*, 1 \le i \le n, n \in \mathbb{N}\}.$$

Since the supremum in (12) is taken over a smaller set therefore

$$I(x) \ge \sup_{\lambda \in X^*} \{< \lambda, x > - \Lambda_\mu(\lambda)\}.$$

Checking the converse inequality we remark that

$$\Phi(\min_{i=1,\ldots,n} \lambda_i) = \inf\{\Phi(a_1\lambda_1 + \cdots + a_n\lambda_n) : a_i \ge 0, a_1 + \ldots + a_n = 1\}$$

as a consequence of Gateaux-differentiability of $\Phi(\lambda)$ for each $\lambda \in X^*$ (see [5]). Thus for each $g \in \mathcal{F}$ there are suitable numbers $a_j \ge 0, j = 1, \ldots, n$, $\sum_{j=1}^n a_j = 1$, such that

$$g(x) - \Phi(g) = g(x) - \Phi(\sum_{i=1}^n a_i\lambda_i) \le \sum_{i=1}^n a_i\lambda_i - \Phi(\sum_{i=1}^n a_i\lambda_i) \le (12),$$

and thus the Cramér rate is identified. □

Remark:

Of course the extended version of the contraction principle, Theorem 3.2, implies that all sequences which are exponentially good approximations of $(P \circ F_L(L_n)^{-1})_n$ satisfy a LDP with the same rate. Thus all sequences which are exponentially good approximations of $(P \circ S_n^{-1})_n$ satisfy a LDP with rate (12) too.

For example let $X_{n1}, \ldots, X_{nn}$ be i.i.d with distribiution ν_n and $\overline{\nu_n}$ denotes the distribution of $\frac{1}{n}(X_{1n} + \cdots + X_{nn})$. Baxter and Jain proved in [2]:

If $\nu_n \to \nu$ weakly and $\int \exp(\sigma||x||)d\nu_n$ is bounded in n for each $\sigma > 0$, then $(\overline{\nu_n})_n$ is an exponentially good approximation of $(\mu_n)_n$. Moreover the distribution of $\frac{1}{n}(\delta_{X_{1n}} + \cdots + \delta_{X_{nn}})$ satisfies a LDP with rate $H(\cdot, \nu)$; thus we can prove Cramér's result in the triangular array setting in two different ways.

Proof of Lemma 5.1 (Donsker and Varadhan [11])

For $\sigma > 0$ and $x \geq 0, y \geq 0$,

$$x\, y \leq \exp(\sigma x) + \frac{y}{\sigma} \log \frac{y}{\sigma}.$$

Let $f_n(x) := \frac{d\varrho_n}{d\mu}$, we obtain

$$\int_{||x|| \geq L} ||x|| \varrho_n(dx) \leq \int_{||x|| \geq L} \left(e^{\sigma ||x||} + \tfrac{f_n(x)}{\sigma} \log \tfrac{f_n(x)}{\sigma} \right) \mu(dx);$$

so if we take $\sigma \geq 1$ so that $\frac{\log \sigma}{\sigma} \geq 0$ we have

$$\sup_n \int_{||x|| \geq L} ||x|| \varrho_n(dx) \leq \sup_n \int_{||x|| \geq L} e^{\sigma ||x||} \mu(dx) + \frac{1}{\sigma} \sup_n \int f_n(x) |\log f_n(x)| \mu(dx).$$

By assumption (11) the first term on the right vanishes for all $\sigma > 0$. Since σ can be taken as large as we please, it suffices to prove

$$\sup_n \int f_n(x) |\log f_n(x)| \mu(dx) < \infty.$$

This can be proven by a tricky argument of Donsker and Varadhan:

$$\int f_n(x) |\log f_n(x)| \mu(dx) = \int f_n(x) \log f_n(x) \mu(dx) - 2 \int_{f_n \leq 1} f_n(x) \log f_n(x) \mu(dx).$$

Since $|\log x| \leq \frac{c}{\sqrt{x}}$ for $|x| \leq 1$ and by our assumption $H(\varrho_n, \mu) \leq C$ we obtain by Jensen's inequality

$$\int f_n(x) |\log f_n(x)| \mu(dx) \leq C + 2c \int \sqrt{f_n(x)} \mu(dx) \leq C + 2c$$

which gives the result.

References

[1] Barbe, P.: A review on some large deviation results; Pub.Inst.Stat.Univ.Paris XXXVIII, facs.1, 1994.

[2] Baxter, J.R.; Jain, N.C.: A comparison principle for large deviations; Proc.Amer.Math.Soc. 103,4, 1235–1240, 1988.

[3] Bolthausen, E.: On the probability of large deviations in Banach spaces; Ann.Probab. 12, 427–435, 1984.

[4] Bolthausen, E.: Laplace approximations for sums of independent random vectors; Prob.Th.Rel.Fields 72, 305–318, 1986.

[5] Bryc, W.: Large deviations by the asymptotic value method; Diffusion Processes and Related Problems in Analysis 1 (M. Pinsky ed.) Birkhäuser Boston 447–472, 1990.

[6] Dawson, D.A.; Gärtner, J.: Large deviations from McKean-Vlasov limit for weakly interacting diffusions; Stochastics 20, 247–308, 1987.

[7] Dembo, A.; Zeitouni, O.: Large Deviations Techniques and Applications; Jones and Bartlett Publishers Boston London, 1993.

[8] Deuschel, J.-D.; Stroock, D.W.: Large Deviations; Academic Press 137, 1989.

[9] Dinwoodie, I.H.: Identifying a large deviation rate function; Ann.Probab. 21, 216–231, 1993.

[10] Dinwoodie, I.H.: Occupation measures for chains of infinite order; Stoch.Proc.Appl. 50, 73–82, 1994.

[11] Donsker, M.D.; Varadhan, S.R.S.: Asymptotic evaluation of certain Markov process expectations for large time III; Commun.Pure Appl.Math.29, 389–461, 1976.

[12] Eichelsbacher, P.: Some aspects in large deviations, preprint 93-048, Sonderforschungsbereich 343, Diskrete Strukturen in der Mathematik, Universität Bielefeld, 1993.

[13] Lanford, O.E.: Entropy and equilibrium states in classical statistical mechanics; in A.Lenard, editor, Statistical Mechanics and Mathematical Problems, Lecture Notes in Physics 20, 1–113, Springer-Verlag, New York, 1973.

[14] Ruelle, D.: A variational formulation of equilibrium statistical mechanics and the Gibbs phase rule; Commun.Math.Physics 5, 324–329, 1967.