Sanov's theorem in the Wasserstein distance: A necessary and sufficient condition

Ran Wang, Xinyu Wang, Liming Wu

School of Mathematics and Statistics, Wuhan University, Wuhan 430072, PR China
Laboratoire de Mathématiques Appliquées, CNRS-UMR6620, Université Blaise Pascal, 63177 Aubière, France
Institute of Applied Mathematics, Chinese Academy of Sciences, 100190 Beijing, PR China

ARTICLE INFO

Article history:
Received 24 August 2009
Received in revised form 3 December 2009
Accepted 3 December 2009
Available online 16 December 2009

MSC:
60F10

ABSTRACT

Let \((X_n)_{n \geq 1}\) be a sequence of i.i.d. r.v.'s with values in a Polish space \((E, d)\) of law \(\mu\). Consider the empirical measures \(L_n = \frac{1}{n} \sum_{k=1}^{n} \delta_{X_k}, n \geq 1\). Our purpose is to generalize Sanov's theorem about the large deviation principle of \(L_n\) from the weak convergence topology to the stronger Wasserstein metric \(W_p\). We show that \(L_n\) satisfies the large deviation principle in the Wasserstein metric \(W_p\) (\(p \in [1, +\infty)\)) if and only if \(\int_E e^{\lambda(x_0, x)} d\mu(x) < +\infty\) for all \(\lambda > 0\), and for some \(x_0 \in E\).

© 2009 Elsevier B.V. All rights reserved.

1. Introduction and main results

Let \((E, d)\) be a complete metric and separable (say Polish) space equipped with its Borel \(\sigma\)-field \(\mathcal{E}\). We denote the set of all probability measures on \(E\) by \(\mathcal{M}_1(E)\).

1.1. Large deviation principle and Wasserstein distance

Large deviation principle (Dembo and Zeitouni, 1998). A function \(I : E \to [0, +\infty]\) is called a rate function, if it is lower semicontinuous, i.e. for any \(L \geq 0\), the level set \([I \leq L]\) is closed. It is called a good rate function, if it is inf-compact, i.e. for any \(L \geq 0\), the level set \([I \leq L]\) is compact.

Given a family of probability measures \((\mu_n)_{n \in \mathbb{N}}\) on \((E, \mathcal{E})\), a speed \(\lambda(n)\) which is a sequence of positive numbers tending to \(+\infty\), and a rate function \(I\) on \(E\), we recall that

(i) \((\mu_n)\) satisfies the Lower bound of Large Deviation with speed \(\lambda(n)\) and with rate function \(I\), if for any open subset \(G\) of \(E\):

\[
\liminf_{n \to +\infty} \frac{1}{\lambda(n)} \log \mu_n(G) \geq \inf_{x \in G} I(x).
\]

(ii) \((\mu_n)\) satisfies the good Upper bound of Large Deviation with speed \(\lambda(n)\) and with rate function \(I\), if \(I\) is inf-compact and for any closed subset \(F\) of \(E\):

\[
\limsup_{n \to +\infty} \frac{1}{\lambda(n)} \log \mu_n(F) \leq \inf_{x \in F} I(x).
\]
(iii) \((\mu_n)\) satisfies the **Large Deviation Principle** (in short: LDP) with speed \(\lambda(n)\) and with rate function \(I\), if the corresponding lower bound of large deviation and good upper bound of large deviation hold both.

Sanov’s theorem. Let \((X_n)_{n \geq 1}\) be a sequence of i.i.d.r.v.’s defined on a probability space \((\Omega, \mathcal{F}, P)\) with values in \(E\), of common law \(\mu\). Consider the empirical measures

\[ L_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}, \quad n \geq 1, \]

which are random elements of \(M_1(E)\). The well-known Sanov’s theorem tells us that \(P(L_n \in \cdot)\) satisfies, as \(n \to +\infty\), the LDP on \(M_1(E)\) equipped with the weak convergence topology \(\sigma(M_1(E),C_b(E))\), with speed \(n\) and with the rate function given by

\[ H(\nu \mid \mu) = \begin{cases} \int_E \frac{d\nu}{d\mu} \log \frac{d\nu}{d\mu} \, d\mu, & \text{if } \nu \ll \mu; \\ +\infty, & \text{otherwise}, \end{cases} \quad (1.1) \]

which is the relative entropy of \(\nu\) with respect to \(\mu\) (in short: w.r.t.) \(\mu\). Here \(C_b(E)\) denotes the space of bounded and continuous real functions on \(E\). We denote the weak convergence topology by “\(w\)”.

**Wasserstein distance.** Given \(p \in [1, +\infty)\) and two probability measures \(\mu\) and \(\nu\) on \(E\), we define the quantity

\[ W_p(\mu, \nu) = \inf_{\pi} \left( \int_E d(x, y)^p \, d\pi(x, y) \right)^{1/p} \quad (1.2) \]

where the inf is taken over all probability measures on the product space \(E \times E\) with marginal distributions \(\mu\) and \(\nu\) respectively [say *couplings* of \((\mu, \nu)\)]. This quantity is commonly referred to as \(L^p\)-**Wasserstein distance** between \(\mu\) and \(\nu\). When \(d\) is the trivial metric \(d(x, y) = I_{\{x \neq y\}}\), \(2W_1(\mu, \nu) = \|\mu - \nu\|_{TV} = \sup_{|y| \leq 1} \int f(\mu - \nu), \) the total variation of \(\mu - \nu\).

The **Wasserstein space of order** \(p\) is defined as

\[ M^p_1(E) = \left\{ \nu \in M_1(E); \int d^p(x, x_0) \nu(dx) < +\infty \right\} \]

where \(x_0\) is some fixed point of \(E\). This space does not depend on the choice of the point \(x_0\). \(W_p\) is a finite distance on \(M^p_1(E)\), and \((M^p_1(E), W_p)\) is a Polish space (see Villani, 2009, Theorem 6.18). It is well known that for \(\nu_n, \nu \in M^p_1(E)\),

\[ \lim_{n \to +\infty} W_p(\nu_n, \nu) = 0 \quad \text{if and only if} \quad \nu_n \overset{w}{\to} \nu \quad \text{weakly} \quad \text{and} \quad \nu_n \overset{P}{\to} \nu. \]

Thus the \(W_p\)-topology is (much) stronger than the \(w\)-topology.

The study on the Wasserstein metric is very active, due to recent great progresses on optimal transportation, transportation inequalities, etc. see Villani (2003, 2009).

1.2. **Main result**

The purpose of this paper is to generalize the Sanov’s theorem in the \(W_p\)-metric, i.e. to check the LDP of \(L_n\) on \((M^p_1(E), W_p)\). We find out that the strong exponential integrability of Cramér functional is a sufficient and necessary condition.

**Theorem 1.1.** Let \((X_n)_{n \geq 1}\) be a sequence of i.i.d.r.v.’s defined on a probability space \((\Omega, \mathcal{F}, P)\) with values in a Polish space \((E, d)\), of common law \(\mu\), and let \((L_n)_{n \geq 1}\) be their empirical measures, \(P \in [1, +\infty)\).

Then \(P(L_n \in \cdot)\) satisfies the LDP on \((M^p_1(E), W_p)\) with speed \(n\) and with some good rate function \(I\), if and only if

\[ \Lambda(\lambda) := \log \int_E \exp(\lambda d^p(x, x_0)) \mu(dx) < +\infty, \quad \forall \lambda > 0, \quad (1.3) \]

for some (thus for any) \(x_0 \in E\). In that case \(I(\nu) = H(\nu \mid \mu)\), \(\nu \in M^p_1(E)\).

For numerous known extensions of Sanov’s theorem, see Eichelsbacher and Schmock (1998) and Dembo and Zeitouni (1998) (for bibliographies). The main new point here is the necessity of the strong exponential integrability condition (1.3).

**Remark 1.2.** Consider a particular case: let \(E\) be countable and \(d\) be the trivial metric \(d(x, y) = I_{\{x \neq y\}}\). As (1.3) holds for \(p = 1\), then \(P(L_n \in \cdot)\) as \(n \to +\infty\), satisfies the LDP on \((M_1(E), \|\cdot\|_{TV})\) with speed \(n\) and with the rate function given by

\[ H(\nu \mid \mu). \]

This strong conclusion is **false** once if \(E\) is infinite but not countable and \(\mu\) is diffuse (i.e. \(\mu(x) = 0\) for any \(x \in E\)): because \(\|L_n - \mu\|_{TV} = 2\). (Theorem 1.1 does not apply since \((E, d)\) is not a Polish space.)

**Remark 1.3.** In the hypothesis testing problem: \(H_0: \mu = \mu_0\) vs \(H_1: \mu \neq \mu_0\) (\(\mu_0\) is a given probability measure), when \(H_0\) is a very important hypothesis in practice (such as in medical testing), one may propose to **refuse** \(H_0\) only if \(W_p(L_n, \mu_0) > \delta\)
Lemma 2.1

By the proof of (1.4), we have

\[-c(\delta+) \leq \lim_{n \to +\infty} \left( \sup_{\delta} \inf_{n} \frac{1}{n} \log P_{0}(W_{p}(L_{n}, \mu_{0}) > \delta) \right) \leq -c(\delta) \tag{1.4}\]

where \(c(\delta) := \inf \{H(v|\mu_{0}) : W_{p}(v, \mu_{0}) \geq \delta\} \), and \(c(\delta+) := \lim_{\epsilon \to 0+} c(\delta + \epsilon)\). The constant \(c(\delta)\) can be estimated by means of the transportation inequality:

\[\alpha(W_{p}(v, \mu_{0})) \leq H(v|\mu_{0}), \quad \forall v \in M_{1}^{\delta}(E) \tag{1.5}\]

where \(\alpha\) is a nondecreasing left continuous function on \(\mathbb{R}^{+}\) with \(\alpha(0) = 0\). We have obviously \(c(\delta) \geq \alpha(\delta)\), which together with (1.4) give an asymptotic estimate of the probability of the first type error \(P_{0}(W_{p}(L_{n}, \mu_{0}) > \delta)\). See Gozlan and Léonard (2007) for several important characterizations of (1.5) in terms of concentration inequality (in relation with large deviations exactly), and Bolley et al. (2007) for concentration inequalities about \(P_{0}(W_{1}(L_{n}, \mu_{0}) > \delta)\) when \(E = \mathbb{R}^{d}\).

This paper is organized as follows. The next section is devoted to the proof of Theorem 1.1. We firstly prove the sufficiency by the technique of the exponential tightness, and then show the necessity by means of Cramér’s theorem and some analytical considerations, which is the main new contribution of this paper. In Section 3 we propose another proof of the necessity in Theorem 1.1 for \(p = 1\), which relates the LDP in the metric \(W_{1}\) with the large deviation of empirical processes, via Kantorovich–Rubinstein’s theorem. Two statistical applications are provided in the last section.

2. Proof of Theorem 1.1

2.1. Several lemmas

Lemma 2.1. Let \(Z\) be a real nonnegative random variable such that \(\Lambda(\lambda) = \log E e^{\lambda Z} < +\infty\) for all \(\lambda > 0\), and \(\Lambda^{\ast}\) be a semi-Legendre transform,

\[\Lambda^{\ast}(r) := \sup_{\lambda \geq 0} \{r\lambda - \log E(e^{\lambda Z})\}\]

for all \(r \geq 0\). Then \(\Lambda^{\ast}\) is a nondecreasing lower semicontinuous convex function, and

\[\frac{\Lambda^{\ast}(r)}{r} \rightarrow +\infty \text{ as } r \rightarrow +\infty.\]

Proof. Taking \(\lambda = 0\), we get \(\Lambda^{\ast}(r) \geq 0\). \(\Lambda^{\ast}(r)\) is a nondecreasing lower semicontinuous convex function because it is the supremum of a class of nondecreasing continuous affine functions. For any fixed numbers \(\lambda > 0\), and \(r > 0\),

\[\Lambda^{\ast}(r) \geq r\lambda - \Lambda(\lambda), \quad \frac{\Lambda^{\ast}(r)}{r} \geq \lambda - \frac{\Lambda(\lambda)}{r}.\]

As \(\Lambda(\lambda) < +\infty\) for any \(\lambda > 0\), \(\lim_{r \to +\infty} \frac{\Lambda^{\ast}(r)}{r} \geq \lambda\). The desired result follows, since \(\lambda\) is arbitrary and \(\Lambda^{\ast}\) is convex. \(\square\)

Lemma 2.2. Let \(Z\) be a real nonnegative random variable such that \(\Lambda(\lambda_{0}) = \log E e^{\lambda Z} < +\infty\) for some \(\lambda_{0} > 0\). Let \(\Lambda^{\ast}\) be a semi-Legendre transform as above, then for any \(0 \leq \delta < 1\),

\[E \exp[\delta \Lambda^{\ast}(Z)] \leq \frac{1}{1 - \delta} < +\infty.\] \(\tag{2.1}\)

Proof. By the proof of Lemma 2.1, \(\Lambda^{\ast}\) is a nondecreasing lower semicontinuous convex function, and \(\Lambda^{\ast}(0) = 0\), \(\lim_{r \to +\infty} \Lambda^{\ast}(r) = +\infty\). By Chebyshev’s inequality, for any \(r \geq 0\),

\[P(Z \geq r) \leq e^{-\Lambda^{\ast}(r)} E e^{\lambda Z}, \quad \forall \lambda \geq 0.\]

Taking infimum over \(\lambda \geq 0\), we get the so called Cramér’s inequality,

\[P(Z \geq r) \leq e^{-\Lambda^{\ast}(r)}, \quad \forall r \geq 0.\]

Let \(r_{0} = \sup \{r > 0 : \Lambda^{\ast}(r) < +\infty\}\). Then \(P(Z > r_{0}) = 0\) by the above inequality. Since \(\Lambda^{\ast}\) is a nondecreasing lower semicontinuous convex function, \(\Lambda^{\ast}\) is absolutely continuous on \([0, r_{0})\) and left continuous at \(r_{0}\). Then by Fubini’s theorem,

\[E \left[ e^{\delta \Lambda^{\ast}(Z)} \right] = E \left[ e^{\delta \Lambda^{\ast}(Z)} I_{[Z \leq r_{0}]} \right] = E \left[ \left( \int_{0}^{Z} e^{\delta \Lambda^{\ast}(t)} (\Lambda^{\ast}(t))' \, dt + 1 \right) I_{[Z \leq r_{0}]} \right].\]
\[
= 1 + \delta \int_0^{+\infty} e^{\delta s(t)} P(Z > t) \left( A^*(t) \right) dt
\leq 1 + \delta \int_0^{+\infty} e^{-(1-\delta) s(t)} \left( A^*(t) \right) dt
\leq 1 + \delta \int_0^{+\infty} e^{-(1-\delta) u} du = \frac{1}{1-\delta}. \quad \square
\]

**Remark 2.3.** Under the condition of Lemma 2.2, similar upper bounds for \( E e^{\delta A^*(Z)} (A^*(r) := \sup_{\lambda \in \mathbb{R}} \{ r\lambda - \log \mathbb{E}(e^{\lambda X}) \} \) is the Legendre transform) are already known: for any \( 0 \leq \delta < 1 \), the upper bound \( 2/(1 - \delta) \) can be found in Dembo and Zeitouni (1998, Lemma 5.1.14), and the upper bound \( (1 + \delta)/(1 - \delta) \) can be found in Gozlan (2005, Lemma 7.17).

### 2.2. Proof of the sufficiency in Theorem 1.1

By Sanov’s theorem, \( P(L_n \in \cdot) \), as \( n \to +\infty \), satisfies the LDP on \((M_1(E), w)\), with speed \( n \) and with rate function \( \nu \mapsto H(\nu|\mu) \). For the LDP of \( L_n \) on the \((M_1^p(E), W_p)\) with the rate function \( \nu \mapsto H(\nu|\mu) \), we have only to show that \( (L_n) \) is exponentially tight in the \( W_p \)-topology (see Wu et al., 1997, Theorem 1.3.6). At first by Sanov’s theorem, for any \( L > 0 \), there exists a compact subset \( K \) of \((M_1(E) \) in the \( w \)-topology, such that
\[
\limsup_{n \to +\infty} \frac{1}{n} \log P(L_n \in K^c) < -L.
\]

For fixed point \( x_0 \in E \), set
\[
A_L := \left\{ \nu \in M_1(E) \mid \int A^*(d\nu(x_0)) d\nu(x) \leq L \right\}, \quad K := A_L \cap K.
\]

Let us to show that \( K \) is \( W_p \)-compact. At first \( K \) is \( w \)-compact because \( \nu \mapsto \int A^*(d\nu(x_0)) d\nu(x) \) is lower semicontinuous on \((M_1(E), w)\). We need to verify that \( K \) has uniformly integrable \( p \)-moments. Let \( C(N) = \frac{\Lambda^*(N)}{x} \), \( N > 0 \). Then \( C(N) \not\to +\infty \) as \( N \to +\infty \) by Lemma 2.1. For each \( \nu \in K \),
\[
\int_{d^p(x_0) \geq N} d^p(x, x_0) d\nu(x) \leq \frac{1}{C(N)} \int_{d^p(x_0) \geq N} \frac{A^*(d^p(x_0))}{d^p(x_0)} d^p(x, x_0) d\nu(x)
\]
\[
\leq \frac{1}{C(N)} \int \Lambda^*(d^p(x_0)) d\nu(x)
\]
\[
\leq \frac{L}{C(N)} \to 0, \quad \text{as } N \to +\infty.
\]

Hence \( \{d^p(x_0), \nu \in K\} \) is uniformly integrable, and thus \( K \) is \( W_p \)-compact. Letting \( a \vee b = \max(a, b) \), we have
\[
\limsup_{n \to +\infty} \frac{1}{n} \log P(L_n \in K^c) = \limsup_{n \to +\infty} \frac{1}{n} \log P(L_n \in K^c \cup A_L^c)
\]
\[
= \limsup_{n \to +\infty} \frac{1}{n} \log P(L_n \in K^c) \vee \limsup_{n \to +\infty} \frac{1}{n} \log P(L_n \in A_L^c)
\]
\[
\leq (-L) \vee \limsup_{n \to +\infty} \frac{1}{n} \log P(L_n \in A_L^c).
\]

Furthermore
\[
P(L_n \in A_L^c) = P \left( \frac{1}{n} \sum_{k=1}^{n} A^*(d^p(X_k, x_0)) > L \right)
\]
\[
\leq \exp \left( -\frac{NL}{2} \right) \mathbb{E} \exp \left\{ \frac{1}{2} \sum_{k=1}^{n} A^*(d^p(X_k, x_0)) \right\}
\]
\[
= \exp \left( -\frac{NL}{2} \right) \left( \mathbb{E} \exp \left\{ \frac{1}{2} A^*(d^p(X_1, x_0)) \right\} \right)^n
\]
\[
\leq \exp \left( -\frac{NL}{2} + 2n \right),
\]

where the last inequality follows from Lemma 2.2. So we get the desired exponential tightness in \((M_1^p(E), W_p)\).
2.3. Proof of the necessity in Theorem 1.1

Let \( a \wedge b := \min\{a, b\} \). For each \( R > 0 \), consider the mappings from \((M^p_1(E), W_p)\) to \( \mathbb{R} \):

\[
\Phi(v) := \int d^p(x, x_0) \, d\nu(x), \quad \Phi_R(v) := \int \{d^p(x, x_0) \wedge R\} \, d\nu(x).
\]

We divide the proof into two steps.

Step 1. \( \Phi_R \) is continuous with respect to \( \nu \)-topology, and thus it is continuous with respect to the stronger \( W_p \)-topology. On any non-empty compact subset \( K \) in \((M^p_1(E), W_p)\), since \( \{d^p(x, x_0), \nu \in K\} \) is uniformly integrable, we have

\[
\Phi_R(v) \to \Phi(v) \text{ uniformly over } K.
\]

Thus \( \Phi \) is also continuous in \((M^p_1(E), W_p)\). By the contraction principle and the assumed LDP of \( L_n \) in \((M^p_1(E), W_p)\),

\[
\mathbb{P}(\{\Phi(L_n) \in \cdot\}) = \mathbb{P}\left(\frac{1}{n} \sum^n_{k=1} d^p(X_k, x_0) \in \cdot\right)
\]

satisfies the LDP on \( \mathbb{R} \). Then by Cramér’s theorem (see Wu et al., 1997, Proposition 2.1.1), there is some \( \delta > 0 \) such that

\[
\Lambda(\delta) = \log \int \exp \left( -\frac{1}{n} \sum^n_{k=1} d^p(X_k, x_0) \right) \, d\mu(x) < +\infty.
\]

Recall that this (weak) exponential integrability condition implies that for each \( \nu \in M_1(E) \) such that \( H(\nu|\mu) < +\infty \),

\[
\delta \int d^p(x, x_0) \, d\nu(x) \leq H(\nu|\mu) + \Lambda(\delta) < +\infty,
\]

(the Donsker–Varadhan’s variational formula for entropy, see Wu et al., 1997, Proposition 3.1.3). We have

\[
\{\nu \in M_1(E); H(\nu|\mu) < +\infty\} \subset M^p_1(E).
\]

Step 2. By the contraction principle, and the assumed LDP of \( L_n \) on \((M^p_1(E), W_p)\), we have that \( \nu \to I(\nu) \) is inf-compact on \((M^p_1(E), W_p)\), and for any \( \delta > 0 \),

\[
\limsup_{n \to +\infty} \frac{1}{n} \log \mathbb{P}(\{\Phi(L_n) - \Phi_R(L_n) \geq \delta\}) \leq - \inf \left\{I(\nu); \nu \in M^p_1(E), \Phi(\nu) - \Phi_R(\nu) \geq \delta\right\}.
\]

The right hand side of (2.2) tends to \(-\infty\) as \( R \to +\infty \), because for any positive number \( L, [I \leq L] \) is a compact set in \((M^p_1(E), W_p)\),

\[
[I \leq L] \cap \{\nu; \nu \in M^p_1(E), \Phi(\nu) - \Phi_R(\nu) \geq \delta\} = \emptyset,
\]

for all \( R > 0 \) large enough by the uniform convergence of \( \Phi_R \) to \( \Phi \) on \([I \leq L] \).

By the lower bound of large deviations in Cramér’s theorem without moment condition (Dembo and Zeitouni, 1998, Theorem 2.2.3), the left hand side of (2.2) is not less than

\[
\liminf_{n \to +\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum^n_{k=1} (d^p - d^p \wedge R)(X_k, x_0) > \delta\right) \geq - \inf_{r > \delta} A^*_R(r),
\]

where

\[
\Lambda_R(\lambda) = \log \int \exp \left\{\lambda(d^p - d^p \wedge R)(x, x_0)\right\} \, d\mu(x);
\]

\[
A^*_R(r) = \sup_{\lambda \in \mathbb{R}} \{\lambda r - \Lambda_R(\lambda)\}.
\]

For any \( \delta > 0 \), and for all \( R \) large enough, say \( R \geq R_\delta \) for some \( R_\delta > 0 \), we have

\[
\varepsilon(R) := \int (d^p - d^p \wedge R)(x, x_0) \, d\mu(x) < \delta.
\]

As \( A^*_R(\varepsilon(R)) = 0 \), \( A^*_R \) is nondecreasing on \([\varepsilon(R), +\infty)\), then for any \( \delta' > \delta \),

\[
\inf_{r > \delta} A^*_R(r) \leq A^*_R(\delta').
\]

Since the left hand side of (2.2) goes to \(-\infty\) as \( R \to +\infty \) (as proved before), we get

\[
A^*_R(\delta') \to +\infty, \quad \text{as } R \to +\infty.
\]
On the other hand, for $R \geq R_0$, since $\delta' > \delta > \varepsilon(R)$, we have for any $\lambda < 0$, 
\[ \lambda \delta' - A_R(\lambda) < \lambda \varepsilon(R) - A_R(\lambda) \leq \Lambda^*(\varepsilon(R)) = 0, \]
thus 
\[ A_R^*(\delta') = \sup_{\lambda \geq 0} \{ \lambda \delta' - A_R(\lambda) \}. \]
If on the contrary 
\[ \lambda_0 := \sup \{ \lambda \geq 0; A(\lambda) < +\infty \} < +\infty, \]
then for $R \geq R_0$, 
\[ A_R^*(\delta') = \sup_{\lambda \geq 0} \{ \lambda \delta' - A_R(\lambda) \} \]
\[ = \sup_{0 \leq \lambda \leq \lambda_0} \{ \lambda \delta' - A_R(\lambda) \} \]
\[ \leq \lambda_0 \delta' < +\infty. \]
This is in contradiction with (2.3). So $\lambda_0 = +\infty$ and the proof is completed.

3. Another proof of the necessity in Theorem 1.1 for $p = 1$

When $p = 1$, we have by Kantorovich–Rubinstein’s theorem (see Villani, 2003, Theorem 1.14)
\[ W_1(\mu, \nu) = \sup \left\{ \int \varphi \, d(v - \mu); \varphi : E \to \mathbb{R}, \| \varphi \|_{Lip} \leq 1 \right\}, \]
where $\| \varphi \|_{Lip} = \sup_{x \neq y} \frac{|\varphi(x) - \varphi(y)|}{d(x, y)}$ is the Lipschitzian seminorm. This result identifies the LDP estimation of $L_n$ on $(M^1_1(E), W_1)$ as that of $L_n(f)$ over the class of all Lipschitz functions on $E$, with Lipschitz constant not more than 1.

More precisely, given 
\[ \mathcal{F} := \{ \varphi : E \to \mathbb{R}, \| \varphi \|_{Lip} \leq 1, \varphi(x_0) = 0 \} \quad \text{for some fixed } x_0 \in E, \]
let $l_\infty(\mathcal{F})$ be the space of all bounded real functions on $\mathcal{F}$ with sup-norm $\| F \|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} | F(f) |$. Then $M^1_1(E) \subset l_\infty(\mathcal{F})$ by identification of $\psi \in M^1_1(E)$ with $F_\psi(f) = \psi(f)$, for any $f \in \mathcal{F}$. Kantorovich–Rubinstein’s theorem tells us that $W_1(\psi \cdot 1, \psi \cdot 1) = \| \psi \|_{\mathcal{F}}$, where $\psi \cdot 1$ is a closed subset of $(l_\infty(\mathcal{F}), \| \cdot \|_{\mathcal{F}})$.

By using of isoperimetric inequality of Talagrand, the third named author (Wu, 1994) obtained the necessary and sufficient conditions for LDP of $L_n(\cdot)$ on $l_\infty(\mathcal{F})$ under the strong exponential integrability condition. Arcones (2003) extended this result in the general case. Let $X$ be an $E$-valued random variable of law $\mu$.

Lemma 3.1 (Arcones, 2003, Theorem 2.7). Given a class $\mathcal{F}$ of real measurable functions on $E$ such that $\sup_{f \in \mathcal{F}} | f(X) | < +\infty$ a.s., then the following properties (a) and (b) are equivalent:

(a) \[ \left\{ \frac{1}{n} \sum_{k=1}^{n} f(X_k); f \in \mathcal{F} \right\} \]
satisfies the large deviation principle on $l_\infty(\mathcal{F})$ with speed $n$ and with a good rate function.

(b) the following conditions hold:

(b.1) $(\mathcal{F}, \rho)$ is totally bounded, where $\rho(f, g) = \mathbb{E} \{ | f(X) - g(X) | \}$.

(b.2) There exists $\lambda_0 > 0$, such that 
\[ \mathbb{E} \{ \exp(\lambda_0 F(X)) \} < +\infty, \]
where $F(x) = \sup_{f \in \mathcal{F}} | f(x) |$.

(b.3) For each $\lambda > 0$, there exists some $\eta > 0$, such that 
\[ \mathbb{E} \{ \exp(\lambda F^\eta(X)) \} < +\infty, \]
where $F^\eta(x) = \sup_{\rho(f, g) \leq \eta} | f(x) - g(x) |$.

(b.4) $\sup_{f \in \mathcal{F}} | \frac{1}{n} \sum_{k=1}^{n} (f(X_k) - \mathbb{E} [ f(X_k) ] ) | \to 0$ in probability.

Proof (The Necessity in Theorem 1.1 for $p = 1$). Since $(M^1_1(E), W_1)$ is a closed subset of $(l_\infty(\mathcal{F}), \| \cdot \|_{\mathcal{F}})$, and $L_n$ satisfies LDP on $(M^1_1(E), W_1)$, we have that $\mathbb{P}(L_n \in \cdot)$ satisfies the LDP on $(l_\infty(\mathcal{F}), \| \cdot \|_{\mathcal{F}})$.

By (b.3), for each $\lambda > 0$, there exists some $\eta > 0$, such that 
\[ \mathbb{E} \{ \exp(\lambda F^\eta(X)) \} < +\infty, \]
where $F^\eta(x) = \sup_{\rho(f, g) \leq \eta} | f(x) - g(x) |$. 

Proof (The Sufficiency in Theorem 1.1 for $p = 1$). Since $(M^1_1(E), W_1)$ is a closed subset of $(l_\infty(\mathcal{F}), \| \cdot \|_{\mathcal{F}})$, and $L_n$ satisfies LDP on $(M^1_1(E), W_1)$, we have that $\mathbb{P}(L_n \in \cdot)$ satisfies the LDP on $(l_\infty(\mathcal{F}), \| \cdot \|_{\mathcal{F}})$.
For any positive integer number $N$, let $f(x) = d(x, x_0)$, $g(x) = d(x, x_0) \wedge N$, then $f$, $g \in \mathcal{F}$, and $\mathbb{E}|f(X)| < +\infty$ by (b.2). We can choose $N$ large enough so that $\rho(f, g) < \eta$. Thus
\[
\mathbb{E}[\exp(\lambda d(X, x_0)) I_{d(x, x_0) > N}] = \mathbb{E}[\exp(\lambda (|f(X) - g(X)| + N)) I_{d(x, x_0) > N}]
\leq \mathbb{E}[\exp(\lambda F^n(X))] \cdot e^{3N} < +\infty,
\]
which follows
\[
\mathbb{E}[\exp(\lambda d(X, x_0))] < +\infty.
\]
The proof is completed. □

4. Two examples

Let us give two statistical applications.

Example 4.1. Let $(X_n)_{n \geq 1}$ be real valued i.i.d.r.v.’s with common distribution function $F(\mu(dx) = dF(x)$ such that \(\int_\mathbb{R} |x| \mu(dx) < +\infty\). Let $F_n^*(x) = \frac{1}{n} \sum_{k=1}^n I_{[-\infty,x]}(X_k)$, $n \geq 1$ be the empirical distribution functions. Since
\[
W_1(v_1, v_2) = \int_\mathbb{R} |F_{v_1}(x) - F_{v_2}(x)| dx, \quad v_1, v_2 \in M_1(\mathbb{R})
\]
(see Villani, 2003, Page 75), where $F_n(x) := \nu(-\infty, x]$ is the distribution function of $\nu$, then $\nu \rightarrow F_n(x) - F(x)$ is isometric from $\mathcal{M}_1(\mathbb{R})$, $W_1$ to $L^1(\mathbb{R}, dx)$. By Theorem 1.1, $\mathbb{P}(F_n^* - F \in \cdot)$ satisfies the LDP on $L^1(\mathbb{R}, dx)$ if and only if $\int_\mathbb{R} e^{\lambda dF(x)} < +\infty$, $\forall \lambda > 0$. In that case the rate function is given by
\[
I(h) = \begin{cases} H(v|\mu), & \text{if } h = F_n - F \text{ (dx - a.e.)} \\ +\infty, & \text{otherwise}. \end{cases}
\]
In particular, for any $\delta > 0$, letting
\[
c(\delta) := \inf \left\{ H(v|\mu); \int_\mathbb{R} |F_n(x) - F(x)| dx \geq \delta \right\} = \inf \left\{ H(v|\mu); \ W_1(v, \mu) \geq \delta \right\}.
\]
then $c(\delta) > 0$ by the inf-compactness of $I$ and
\[
- \lim_{\epsilon \rightarrow 0^+} c(\delta + \epsilon) \leq \lim_{n \rightarrow +\infty} \left( \sup_{x} \inf_{\nu} \frac{1}{n} \log \mathbb{P} \left( \int_\mathbb{R} |F_n^*(x) - F(x)| dx > \delta \right) \right) \leq -c(\delta).
\]
For example, it is known that (Liu, 2009)

(i) if $\nu$ is the Gaussian measure with mean $m$ and variance $\sigma^2$, $c(\delta) = \frac{\delta^2}{2\sigma^2}$,

(ii) if $\nu$ is the Poisson distribution of parameter $\lambda > 0$, $c(\delta) = (\delta + \lambda) \log \frac{\delta + \lambda}{\lambda} - \delta$.

See Villani (2009) for estimates of $c(\delta)$ by means of transportation inequalities, and the third named author (Wu, 1994) for large deviations of $F_n^*$ in the sup-norm.

Example 4.2 (U-statistics). Let $(X_k)$ be as in Theorem 1.1 and $u : E^N \rightarrow \mathbb{R}$ a continuous function ($N \geq 1$). The quantity
\[
U_N := \int \cdots \int u(y_1, \ldots, y_N) L_n(dy_1) \cdots L_n(dy_N) = \frac{1}{n^N} \sum_{1 \leq k_1, \ldots, k_N \leq n} u(X_{k_1}, \ldots, X_{k_N})
\]
is a so called U-statistic. Assume that for some $p \in [1, +\infty)$ fixed, our strong exponential integrability condition (1.3) holds, and for some constant $C > 0$,
\[
|u(y_1, \ldots, y_N)| \leq C \left( 1 + \sum_{i=1}^N d^p(y_i, x_0) \right), \quad \forall (y_1, \ldots, y_N) \in E^N.
\]
Then $\mathbb{P}(U_n \in \cdot)$ satisfies the LDP on $\mathbb{R}$ with speed $n$ and with the rate function given by
\[
I(r) = \inf \left\{ H(v|\mu); \Phi(v) := \int \cdots \int u(y_1, \ldots, y_N) v(dy_1) \cdots v(dy_N) = r \right\}.
\]
In fact by Theorem 1.1 and the contraction principle it is enough to show that $\nu \rightarrow \Phi(v)$ is continuous in the metric $W_p$. To this purpose note that if an $E^N$-valued random variable $Y = (Y_1, \ldots, Y_N)$ is of law $\nu_{\otimes N}$ (product measure), then
\[
\Phi(v) = \int \cdots \int u(y_1, \ldots, y_N) v(dy_1) \cdots v(dy_N) = \mathbb{E}u(Y).
\]
Let $\nu_n \rightarrow \nu$ on $(M^p(E), W_p)$. By Skorokhod’s lemma we can find, on some probability space $(\Omega, \mathcal{F}, P)$, a sequence of random variables $Y_n = (Y_{n,1}, \ldots, Y_{n,N})$ of law $\nu_n^{\otimes N}$ such that $Y_n \rightarrow Y$, a.s. in the product space $E^N$, where the law of $Y$ is $\nu^{\otimes N}$. Noting that $u(Y_n) \rightarrow u(Y)$ by the continuity of $u$, and

$$|u(Y_n)| \leq C \left(1 + \sum_{i=1}^{N} d_p(Y_{N,i}, x_0)\right), \quad n \geq 1$$

are uniformly integrable, we get $E u(Y_n) \rightarrow E u(Y)$, and thus $\Phi(\nu_n) \rightarrow \Phi(\nu)$.

Similar results have been established in Eichelsbacher and Löwe (1998).

Acknowledgements

The authors are very grateful to the editor and associated editor for their comments. We are particularly indebted to the referee for his attentive corrections and helpful comments. The first named author also thanks Professor F.Q. Gao, Z.L. Zhang and N. Yao for the useful discussions.

References


