A Large Deviations Heuristic Made Precise

Neil O'Connell
Basic Research Institute in the Mathematical Sciences
HP Laboratories Bristol
HPL-BRIMS-98-10
May, 1998

Sanov's theorem, Monge, Kantorovich, Ornstein, Azuma's inequality, extended contraction principle, bin-packing

Sanov's theorem states that the sequence of empirical measures associated with a sequence of iid random variables satisfies the large deviation principle (LDP) in the weak topology with rate function given by a relative entropy. We present a derivative which allows one establish LDP's for symmetric functions of many iid random variables under the condition that (i) a law of large numbers holds whatever the underlying distribution and (ii) the functions are uniformly Lipschitz. This heuristic (of the title) is that the LDP follows from (i) provided the functions are "sufficiently smooth". As an application, we obtain large deviations results for the stochastic bin-packing problem.

1 The heuristic

Let $X_1, X_2, \ldots$ be a sequence of independent random variables taking values in a Polish space (E, d) with common law μ. Suppose that, for each n, $f_n : E^n \to \mathbb{R}$ is symmetric and measurable and that (for any underlying distribution μ) we have a strong law of large numbers:

$$f_n(X_1, \ldots, X_n) \xrightarrow{a.s.} f(\mu).$$

Then, provided the f_n are sufficiently smooth, we might hope to deduce the existence of a large deviation principle for the sequence $f_n(X_1, \ldots, X_n)$ with rate function given by

$$J(y) = \inf\{H(\nu|\mu) : \ f(\nu) = y\}, \tag{1}$$

where $H(\cdot|\mu)$ is the relative entropy function:

$$H(\nu|\mu) = \begin{cases} \int_E d\nu \log \frac{d\nu}{d\mu} & \nu \ll \mu \\ \infty & \text{otherwise} \end{cases}$$

This is a heuristic application of the (extended) contraction principle. It is useful because, put simply, laws of large numbers are easier to prove than large deviation principles.

For example, consider the stochastic bin-packing problem. Here E is the unit interval and, for $x \in [0,1]^n$, $nf_n(x_1, \ldots, x_n)$ is the smallest number of unit-sized bins required to pack n objects of respective sizes $x_1, \ldots, x_n$. It is well-known, and easy to prove, that if $X_1, X_2, \ldots$ is a sequence of independent random variables taking values in $[0,1]$ with common law μ, then

$$f_n(X_1, \ldots, X_n) \xrightarrow{a.s.} c(\mu),$$

for some (finite) $c(\mu)$ called the 'packing constant'. According to the heuristic, the large deviation principle follows with rate function given by

$$J(y) = \inf\{H(\nu|\mu) : \ c(\nu) = y\}.$$

We shall see later that this statement is (almost) correct.

In this short note we present a sufficient condition on the f_n which is both simple to check and justifies the above heuristic. We make no attempt to prove the best possible result; the aim of this note is to give the reader an understanding of where the heuristic comes from and feeling for the 'type' of condition under which we can expect it to hold.

2 The main result

Let (E,d) be a bounded Polish space. Denote by $M_1(E)$ the space of Borel probability measures on E, endowed with the weak topology, and by $M_1^\mu(E)$ the subspace of probability measures which are absolutely continuous with respect to μ. Suppose that, for each n, $f_n : E^n \to \mathbb{R}$ is symmetric and Borel measurable and satisfies the Lipschitz condition

$$|f_n(x) - f_n(y)| \leq \frac{K}{n} \sum_{i=1}^{n} d(x_i, y_i) \tag{2}$$

for all $x, y \in E^n$, where K is a fixed constant (independent of n). We prove the following.

Theorem 1 *If there exists a mapping $f : M_1^\mu(E) \to \mathbb{R}$ such that for each $\nu \in M_1^\mu(E)$,*

$$\lim_n \int_{\mathbb{R}} x(\nu^{\times n} \circ f_n^{-1})(dx) = f(\nu),$$

then f is weakly continuous (in fact ρ-Lipschitz, where ρ is the MKO metric defined by 3 below) and the sequence $\mu^{\times n} \circ f_n^{-1}$ satisfies the LDP in $\mathbb{R}$ with good rate function given by

$$J(y) = \inf\{H(\nu|\mu) : \ f(\nu) = y\}.$$

We will begin by presenting the four main ingredients of the proof.

1. Sanov's theorem. If X_n are independent random variables taking values in a Polish space (E,d), with common law μ, and we set

$$L_n = \sum_{i=1}^{n} \delta_{X_i},$$

then the sequence $\mathcal{L}(L_n)$ satisfies the LDP in $M_1(E)$ with good convex rate function $H(\cdot|\mu)$.

2. The extended contraction principle. Let $\mathcal{X}$ be a Hausdorff topological space, equipped with its Borel σ-algebra, and let μ_n be a sequence of probability measures on $\mathcal{X}$. Let $\mathcal{Y}$ be another Hausdorff topological space. The usual contraction principle states that, if the sequence μ_n satisfies the

LDP in $\mathcal{X}$ with good rate function $I : \mathcal{X} \to \mathbb{R}_+$ and $f : \mathcal{X} \to \mathcal{Y}$ is a continuous mapping, then the sequence $\mu_n \circ f^{-1}$ satisfies the LDP in $\mathcal{Y}$ with good rate function given by

$$J(y) = \inf\{I(x) : \ f(x) = y\}.$$

The extended contraction principle applies to the case where we have, for each n, a mapping $f_n : \mathcal{X} \to \mathcal{Y}$ and wish to obtain an LDP for the sequence $\mu_n \circ f_n^{-1}$. There are a number of statements in the literature, dating back to the seminal paper of Varadhan [10] (see also [1] and [8]) which are roughly equivalent to the following. For completeness, we have included a short proof in the appendix.

Theorem 2 *Assume that $\mathcal{X}$ is a metric space. Suppose that for each n, supp $\mu_n \subset \mathcal{X}_n \subset \mathcal{X}$, $f_n : \mathcal{X}_n \to \mathcal{Y}$ is continuous, and the sequence μ_n satisfies the LDP in $\mathcal{X}$ with good rate function I with effective domain contained in $\mathcal{X}_\infty \subset \mathcal{X}$. Suppose also that there exists a continuous mapping $f : \mathcal{X}_\infty \to \mathcal{Y}$ such that whenever $x_n \in \mathcal{X}_n$ and $x_n \to x \in \mathcal{X}_\infty$, we have $f_n(x_n) \to f(x)$. Then the sequence $\mu_n \circ f^{-1}$ satisfies the LDP in $\mathcal{Y}$ with good rate function given by*

$$J(y) = \inf\{I(x) : \ f(x) = y\}.$$

In the context of Sanov's theorem, the extended contraction principle can be stated as follows. Suppose that, for each n, $f_n : E^n \to \mathcal{Y}$ is symmetric and Borel measurable and there exists a continuous mapping $f : M_1^\mu(E) \to \mathcal{Y}$ such that whenever, for $x \in E^\infty$,

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \ \xrightarrow{w} \ \nu \in M_1^\mu(E)$$

we have $f_n(x_1, \ldots, x_n) \to f(\nu)$. Then the sequence $\mu^{\times n} \circ f_n^{-1}$ satisfies the LDP in $\mathcal{Y}$ with good rate function given by

$$J(y) = \inf\{H(\nu|\mu) : \ f(\nu) = y\}.$$

3. The Monge-Kantorovich-Ornstein (MKO) distance. Let (E, d) be a metric space. For $\pi \in M_1(E^2)$, denote by π_1 and π_2 the respective marginals of π in $M_1(E)$. The MKO distance between two probability measures $\mu, \nu \in M_1(E)$ is defined by

$$\rho(\mu, \nu) = \inf\left\{\int_{E^2} d(x, y)\pi(dx, dy) : \ \pi \in M_1(E^2), \ \pi_1 = \mu, \ p_2 = \nu\right\}. \quad (3)$$

This measure of distance was first introduced in 1781 by Monge [6] in studying the most efficient way of transporting soil. It was later developed in a measure-theoretic context by Kantorovich [5] and Ornstein [7], among others (see Rachev [9] for an extensive and fascinating survey). For our purposes it is sufficient to note that, if (E, d) is compact, then ρ metrises the weak topology on $M_1(E)$.

4. A concentration inequality. The final ingredient in the proof is the following elementary concentration inequality, which is an immediate consequence of the Azuma-Hoeffding inequality for martingale differences.

Lemma 3 *Let (E, d) be a bounded metric space and $f_n : E^n \to \mathbb{R}$ a symmetric (Borel measurable) function which satisfies the Lipschitz condition*

$$|f_n(x) - f_n(y)| \leq \frac{K}{n} \sum_{i=1}^{n} d(x_i, y_i)$$

for all $x, y \in E^n$, where K is a constant. Let $X_1, \ldots, X_n$ be independent random variables in E and write $X^n = (X_1, \ldots, X_n) \in E^n$. Then:

$$P\left(|f_n(X^n) - Ef_n(X^n)| > t\right) \leq 2 \exp -At^2 n$$

where $A = 2/K^2\delta^2$ and $\delta = \sup_{x,y \in E^n} d(x, y)$.

We are now ready to prove Theorem 1. For $x \in E^n$, set

$$l_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

and, for each permutation $\sigma \in S_n$, write $x \circ \sigma$ for the permuted sequence $(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) \in E^n$.

By hypothesis, we have that for all $x, y \in E^n$,

$$|f_n(x) - f_n(y)| \leq \frac{K}{n} \sum_{i=1}^{n} d(x_i, y_i).$$

The first key observation is that, since f_n is symmetric, this implies

$$\begin{aligned} |f_n(x) - f_n(y)| &\leq \inf_{\sigma \in S_n} \frac{K}{n} \sum_{i=1}^{n} d(x_i, y_{\sigma(i)}) \\ &\equiv K\rho(l_n(x), l_n(y)). \end{aligned}$$

If Y_i are iid with common law $\nu \in M_1^\mu(E)$ then, again by hypothesis, $Ef_n(Y^n) \to f(\nu)$. It follows, applying the concentration inequality (and Borel-Cantelli) that $f_n(Y^n) \xrightarrow{a.s.} f(\nu)$. We also have, by Sanov's theorem (and Borel-Cantelli), that $l_n(Y^n) \xrightarrow{a.s.} \nu$ (in the weak topology). In particular, there exists a sequence $y \in E^\infty$ such that $f_n(y^n) \to f(\nu)$ and $l_n(y^n) \xrightarrow{w} \nu$. It follows that, for any $x \in E^\infty$ with $l_n(x^n) \xrightarrow{w} \nu$ we have, by the continuity of ρ,

$$|f_n(x^n) - f_n(y^n)| \leq K\rho(l_n(x^n), l_n(y^n)) \to 0,$$

and so $f_n(x^n) \to f(\nu)$, as required. In order to apply the extended contraction principle, it remains to check that f is weakly continuous. For $\mu, \nu \in M_1^\mu(E)$ we can find sequences $x, y \in E^\infty$ such that $l_n(x^n) \xrightarrow{w} \mu$ and $l_n(y^n) \xrightarrow{w} \nu$. From the above,

$$\begin{aligned} |f(\mu) - f(\nu)| &= \lim_n |f_n(x^n) - f_n(y^n)| \\ &\leq \limsup_n K\rho(l_n(x^n), l_n(y^n)) \\ &= K\rho(\mu, \nu), \end{aligned}$$

as required. Note that we have established more, namely that f is ρ-Lipschitz. This completes the proof of the theorem.

3 Application to the stochastic bin-packing problem

The standard reference on bin-packing is the book of Coffman and Lueker [2]. For $x \in [0,1]^n$, let $nf_n(x_1, \ldots, x_n)$ be the smallest number of unit-sized bins required to pack n objects of respective sizes $x_1, \ldots, x_n$. It is well-known, and easy to prove, that if $X_1, X_2, \ldots$ is a sequence of independent random variables taking values in $[0,1]$ with common law μ, then

$$f_n(X_1, \ldots, X_n) \xrightarrow{a.s.} c(\mu),$$

for some (finite) $c(\mu)$ called the 'packing constant'. Mean convergence, which is all we need to apply the theorem, follows from a simple subadditivity argument.

Suppose for the moment that μ is supported on a finite subset E of $[0,1]$ and let $\epsilon > 0$ be the minimal separation distance between elements of E.

Then the f_n's satisfy the Lipschitz condition (2) on E^n with $K = 1/\epsilon$, and we can apply the theorem to get that the sequence $f_n(X^n)$ satisfies the LDP in $[0,1]$ with good rate function given by

$$J(y) = \inf\{H(\nu|\mu) : \ c(\nu) = y\}.$$

For general μ, we need to do some extra work. Let F denote the distribution function associated with μ and for each positive integer m, set

$$\mu_m^+ = \frac{1}{m}\sum_{j=1}^{m} F^{-1}(j/m),$$

and

$$\mu_m^- = \frac{1}{m}\sum_{j=0}^{m-1} F^{-1}(j/m).$$

Then, as Coffman and Lueker observe,

$$c(\mu) - 1/m \le c(\mu_m^-) \le c(\mu) \le c(\mu_m^+) \le c(\mu) + 1/m.$$

We also have the related fact that $\mu^{\times n} \circ f_n^{-1}$ is majorised (resp. minorised) by $(\mu_m^+)^{\times n} \circ f_n^{-1}$ (resp. $(\mu_m^-)^{\times n} \circ f_n^{-1}$). Combining these observations, we obtain the following 'pseudo-LDP' for the sequence $\mu^{\times n} \circ f_n^{-1}$: for $c(\mu) < q < 1$,

$$\limsup_n \frac{1}{n} \log \mu^{\times n}(f_n^{-1}[q,1]) \le -\liminf_m \inf\{H(\nu|\mu) : \ c(\nu) \ge q - 1/m\}$$

and

$$\liminf_n \frac{1}{n} \log \mu^{\times n}(f_n^{-1}[q,1]) \ge -\liminf_m \inf\{H(\nu|\mu) : \ c(\nu) \ge q + 1/m\};$$

similar bounds hold for deviations to the left of $c(\mu)$. If J is continuous and increasing (resp. decreasing) to the right (resp. to the left) of $c(\mu)$, the LDP holds with rate function J. To say more than that requires a careful analysis of the variational problem and is beyond the scope of this paper.

4 Concluding remarks

The main result of this paper can be extended in many directions. The concentration inequality holds for Banach-space-valued functions, but not for

unbounded d. In the case of unbounded d, Sanov's theorem can be extended to hold in the corresponding MKO topology under a moment condition such as

$$E \exp \delta d(X_1, x) < \infty,$$

for some $\delta > 0$ and $x \in E$; however, since we cannot appeal to a concentration inequality we must assume that $f_n(X^n)$ converges almost surely to $f(\mu)$ in the hypothesis.

An appealing feature of Theorem 1 is that we deduce the LDP under the type of condition which is usually associated with concentration inequalities. The particular Lipschitz condition we assume is, in a sense, the most naive in this class; it would be interesting to see if the LDP can be obtained under more sophisticated (milder!) conditions that have been developed in that area.

Finally, a word of caution. In many problems of combinatorial optimisation, such as the traveling salesman and longest increasing subsequence problems, the functionals of interest are highly discontinuous and the heuristic discussed in this paper breaks down. This is because such functionals depend on much finer properties of the empirical measure than those which are asymptotically captured in the weak topology. There is a recent paper by Deuschel and Zeitouni [4] which beautifully illustrates this point for the longest increasing subsequence problem.

Appendix: Proof of Theorem 2

The simplest way to prove Theorem 2 is as follows. Denote by $\mathbb{N}^*$ the extended natural numbers and equip $\mathbb{N}^*$ with the metric

$$h(n,m) = \left| \frac{1}{n} - \frac{1}{m} \right|,$$

with the convention that $1/\infty = 0$. Then (trivially) the sequence (μ_n, n) satisfies the LDP in $\mathcal{X} \times \mathbb{N}^*$, equipped with the product topology, with good rate function given by

$$I_e(x,n) = \begin{cases} I(x) & n = \infty \\ \infty & \text{otherwise} \end{cases}$$

We can restrict this LDP to the (measurable) subspace

$$\bigcup_{n\in\mathbb{N}^*} \mathcal{X}_n \times \{n\},$$

as the effective domain of I_e lies in this subspace (see, for example, [3, Lemma 4.1.5]). The statement of the theorem now follows by applying the usual contraction principle to the mapping

$$F : \bigcup_{n\in\mathbb{N}^*} \mathcal{X}_n \times \{n\} \to \mathcal{Y}$$

defined by

$$F(x,n) = \begin{cases} f_n(x) & x \in \mathcal{X}_n,\ n < \infty \\ f(x) & n = \infty. \end{cases}$$

Note that we have used the fact that since $\mathcal{X}$ is a metric space, we can check continuity of F using sequences. ∎

References

[1] N.R. Chaganty. Large deviations for joint distributions and statistical applications. Technical report TR93-2, Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA (1993).

[2] E.G. Coffman Jr. and George S. Lueker. *Probabilistic Analysis of Packing and Partitioning Algorithms.* Wiley, 1991.

[3] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications.* Jones and Bartlett, London, 1992.

[4] J.-D.Deuschel and O.Zeitouni. On increasing subsequences of i.i.d. samples. To appear.

[5] L.V. Kantorovich. On the transfer of masses. *Dokl. Akad. Nauk. SSSR* 37(7-8): 227-229, 1942.

[6] G. Monge. Mémoire sur la théorie des déblais et des remblais. Histoire de l'Académie des sciences de Paris, avec les Mémoires de mathématique et de physique pour la même année, 1781, pp. 666-704.

[7] D. Ornstein. An application of ergodic theory to probability theory. *Ann. Prob.* 1(1):43–65, 1973.

[8] A. Puhalskii. Large deviation analysis of the single server queue. *Queueing Systems* 21 (1995) 5–66.

[9] S.T Rachev. The Monge-Kantorovich mass transference problem and its stochastic applications. *Theor. Probab. Appl.* 29(4):647–676, 1984.

[10] S.R.S. Varadhan. Asymptotic probabilities and differential equations. *Comm. Pure Appl. Math.* 19:261–286, 1966.