The Note on The Two-Sampling Matching Problem

Ayalvadi Ganesh, Neil O'Connell
Basic Research Institute in the Mathematical Sciences
HP Laboratories Bristol
HPL-BRIMS-98-11
May, 1998

empirical processes,
discrepancy,
large deviations

In this short note we tie up some loose ends regarding the two-sample matching problem and its connections with the Monge-Kantorovich problem of optimal transportation of mass. By making this connection explicit, we immediately obtain moderate and large deviation principles.

A note on the two-sample matching problem

Ayalvadi Ganesh and Neil O'Connell
BRIMS, Hewlett-Packard Labs, Bristol

May 12, 1998

ABSTRACT: In this short note we tie up some loose ends regarding the two-sample matching problem and its connections with the Monge-Kantorovich problem of optimal transportation of mass. By making this connection explicit, we immediately obtain moderate and large deviation principles.

KEYWORDS: empirical processes, discrepancy, large deviations

1 Introduction

Let X_i and Y_i be independent random variables, uniformly distributed on the unit square, and consider the random quantity

$$T_n^1 = \inf_{\sigma \in S_n} \sum_{i=1}^{n} |X_i - Y_{\sigma(i)}|.$$

This is the canonical two-sample matching problem. Ajtai, Komlòs and Tusnàdy [1] prove the following 'law of the (uniterated) logarithm' for the sequence T_n^1: there exists $K > 0$ such that

$$\frac{1}{K}(n \log n)^{1/2} < T_n^1 < K(n \log n)^{1/2}, \tag{1.1}$$

with probability $1 - o(1)$. Refinements and extensions of this result have been obtained by Shor [10] and Talagrand [11], among others. It is still an open problem to determine if $(n \log n)^{-1/2} T_n^1$ actually converges, even in expectation. A related problem is to determine the asymptotics of

$$T_n^\infty = \inf_{\sigma \in S_n} \max_{i \leq n} |X_i - Y_{\sigma(i)}|.$$

Leighton and Shor [7] obtained the following analogue of (1.1): there exists $K > 0$ such that

$$\frac{1}{K} n^{-1/2} (\log n)^{3/4} < T_n^\infty < K n^{-1/2} (\log n)^{3/4}, \tag{1.2}$$

with probability $1 - o(1)$. Concentration inequalities for these problems have also been obtained. One of the main tools is the connection with empirical processes. By 'duality', or generalisations of the marriage lemma, the random variable T_n^1 can be related to the 'empirical discrepancies'

$$D_n(X) = \|L_n - \lambda\|_{\mathcal{F}}$$

and

$$D_n(Y) = \|M_n - \lambda\|_{\mathcal{F}},$$

where $L_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, $M_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$, λ is Lebesgue measure on the unit square, and

$$\|\mu\|_{\mathcal{F}} = \sup\{|\mu(f)| : f \in \mathcal{F}\}, \tag{1.3}$$

for a signed measure μ, where $\mathcal{F}$ is taken to be the set of Lipschitz continuous functions on the unit square with Lipschitz constant 1 (see, for example, [11]). The asymptotics of such measures of empirical discrepancy have been studied extensively in the empirical processes literature. One motivation is the fact that on the space of probability measures the metric defined by $\beta(\mu, \nu) = \|\mu - \nu\|_{\mathcal{F}}$ generates the weak topology (see, for example, Dudley [2]). In particular, Dudley [3] obtains mean rates of convergence for $\beta(L_n, \lambda)$ (actually, he considers a much more general setting); he obtains, in this case, the estimate

$$E\beta(L_n, \lambda) \leq cn^{-1/2}(1 + \log n).$$

While all of the above reflects a deep understanding of the two-sample matching problem and its connection with the theory of empirical processes, it seems to ignore a huge body of literature, which dates back to 1781. Moreover, while the connections which exist have clearly been exploited, they can be made more explicit. This is the object of the present note, and leads to a better analytic understanding of the problem. In particular, we can easily infer large and moderate deviations behaviour.

2 The Monge-Kantorovich Problem

In 1781, Monge [8] formulated the following problem: *Split two equally large volumes into infinitely small particles and then associate them with each other so that the sum of products of these paths of the particles to a volume is least. Along what paths must the particles be transported and what is the smallest transportation cost?* This problem was first made precise and studied by Kantorovich [4, 5]. Suppose that μ and ν are Borel probability measures on a compact metric space (E, d) and $¶(\mu, \nu)$ is the space of all Borel probability measures π on $E \times E$ with fixed marginals $\mu(\cdot) = \pi(\cdot \times E)$ and $\nu = \pi(E \times \cdot)$. Kantorovich defined the metric

$$\rho_1(\mu, \nu) = \inf\{\int_{E\times E} d(x, y)\pi(dx,\ dy) :\ \pi \in ¶(\mu, \nu)\} \tag{2.4}$$

and proved that

$$\rho_1(\mu,\nu) = \|\mu-\nu\|_{\mathcal{F}}, \tag{2.5}$$

where $\|\cdot\|_{\mathcal{F}}$ is defined by (1.3). The properties of ρ_1 and its relatives have since been studied extensively: see Rachev [9] for a monumental survey of the literature. In particular, it was shown by Kantorovich and Rubinshtein [6] that ρ_1 metrises the weak topology on $M_1(E)$, the space of probability measures on E.

To see how this related to the matching problem, observe that, if we take E to be the unit square, then

$$\frac{1}{n}T_n^1 = \rho_1(L_n, M_n). \tag{2.6}$$

If we define

$$\rho_\infty(\mu,\nu) = \inf\{\text{ess-sup } \pi \circ d^{-1} :\ \pi \in \P(\mu,\nu)\}, \tag{2.7}$$

then ρ_∞ also metrises the weak topology and

$$T_n^\infty = \rho_\infty(L_n, M_n).$$

For ρ_∞, we have the identity

$$\rho_\infty(\mu,\nu) = \inf\{\epsilon > 0 :\ \mu(A) \le \nu(A^\epsilon), A \in \mathcal{B}(E)\}, \tag{2.8}$$

where $A^\epsilon = \{x :\ d(x,A) < \epsilon\}$ and $\mathcal{B}(E)$ is the Borel σ-algebra on E.

It is clear that identity (2.5) is the underlying force behind the work of Talagrand and Shor, but it is never made entirely explicit. In the next section we will use (more general versions of) the identities (2.5) and (2.6) to obtain large and moderate deviation results for the matching problem.

3 Large and moderate deviations results

Let $(X_n)_{n\ge1}$ be a sequence of random variables defined on a probability space $(\Omega,\mathcal{F},P)$, with values in a Hausdorff topological vector space E equipped with the Borel σ-algebra $\mathcal{E}$. Denote by $M_1(E)$ (resp. $M_b(E)$) the space of probability measures (resp., finite signed measures) on $(E,\mathcal{E})$. Let μ_n denote the law of X_n. We say that the sequence X_n (equivalently, μ_n) satisfies the *large deviation principle* (LDP) with rate function I, if for all $B \in \mathcal{E}$,

$$-\inf_{x\in B^\circ} I(x) \le \liminf_n \frac{1}{n}\log\mu_n(B) \le \limsup_n \frac{1}{n}\log\mu_n(B) \le -\inf_{x\in\bar{B}} I(x).$$

Here B° and $\bar{B}$ denote the interior and closure of B, respectively.

Let $(\lambda_n)_{n\geq 1}$ be an increasing, positive sequence such that

$$\lambda_n \to \infty \quad \text{and} \quad \frac{\lambda_n}{\sqrt{n}} \to 0.$$

We say the the sequence X_n satisfies the *moderate deviation principle* (MDP) with rate function I and speed λ_n^{-2}, if for all $B \in \mathcal{E}$,

$$\begin{aligned} -\inf_{x\in B^\circ} I(x) &\leq \liminf_n \frac{1}{\lambda_n^2} \log P\left(\frac{\sqrt{n}}{\lambda_n} X_n \in B\right) \\ &\leq \limsup_n \frac{1}{\lambda_n^2} \log P\left(\frac{\sqrt{n}}{\lambda_n} X_n \in B\right) \leq -\inf_{x\in \bar{B}} I(x). \end{aligned}$$

It doesn't complicate matters to consider an abstraction of the matching problem, so we shall do just that. Let X_i and Y_i be independent random variables taking values in a compact metric space (E, d) with common law μ, and consider the random quantities defined by

$$T_n^1 = \inf_{\sigma\in S_n} \sum_{i=1}^n d(X_i, Y_{\sigma(i)})$$

and

$$T_n^\infty = \inf_{\sigma\in S_n} \max_{i\leq n} d(X_i, Y_{\sigma(i)}).$$

Then

$$\frac{1}{n} T_n^1 = \rho_1(L_n, M_n)$$

and

$$T_n^\infty = \rho_\infty(L_n, M_n),$$

where ρ_1 and ρ_∞ are the metrics on $M_1(E)$ defined by (2.4) and (2.7). Since E is compact, both metrics induce the weak topology on $M_1(E)$. The following is thus an immediate consequence of Sanov's theorem and the contraction principle (taking products is continuous in the weak topology).

Theorem 3.1 *The sequences $P(T_n^1/n \in \cdot)$ and $P(T_n^\infty \in \cdot)$ both satisfy the LDP in $\mathbb{R}$ with respective rate functions*

$$I_1(x) = \inf\{H(\nu_1|\mu) + H(\nu_2|\mu) : \ \rho_1(\nu_1, \nu_2) = x\}$$

and

$$I_\infty(x) = \inf\{H(\nu_1|\mu) + H(\nu_2|\mu) : \ \rho_\infty(\nu_1, \nu_2) = x\}.$$

Now suppose E is the unit square and that μ denotes Lebesgue measure on E. Consider the variational problem,

$$\inf\{H(\nu_1|\mu) + H(\nu_2|\mu) : \ \rho_\infty(\nu_1, \nu_2) = x\},$$

where ρ_∞ is defined by (2.8). If $0 \leq x \leq 1/2$, then it is not hard to see from the convexity of $H(\cdot|\mu)$ that a solution (it is not unique) of the variational problem is as follows: $\nu_1 \equiv \mu$, ν_2 is identically zero on the circle of radius x centred at the mid-point of the square, and is proportional to Lebesgue measure outside this circle, with the constant of proportionality chosen so that ν_2 is a probability measure. Thus, a simple calculation yields that

$$I_\infty(x) = -\log(1 - \pi x^2), \quad 0 \leq x \leq 1/2.$$

It is not as straightforward to obtain the moderate deviation principles for T_n^1/n and T_n^∞ from Sanov's theorem because $\rho_1(\cdot,\cdot)$ and $\rho_\infty(\cdot,\cdot)$ are not continuous functions on the space of signed measures. We shall use results of Wu [12], who derives conditions for the MDP to hold uniformly over a class of functions, to obtain an MDP for T_n^1/n. The MDP for T_n^∞ remains an open problem.

Denote by $\mathcal{F}$ the space of Lipschitz functions f on the unit square, with Lipschitz constant 1 and such that $0 \leq f \leq 2$. Let $d_2(f,g) = (\int (f-g)^2 d\mu)^{1/2}$ denote the L_2 metric on $\mathcal{F}$, where μ is Lebesgue measure on the unit square. It is not hard to see that $(\mathcal{F}, d_2)$ is totally bounded. Denote by $\ell_\infty(\mathcal{F})$ the space of bounded real functions on $\mathcal{F}$, and equip it with the sup norm. Every signed measure $\nu \in M_b(E)$ corresponds to an element $\nu^{\mathcal{F}} \in \ell_\infty(\mathcal{F})$ given by $\nu^{\mathcal{F}}(f) = \nu(f) := \int f d\nu$ for all $f \in \mathcal{F}$.

It is suggested by (1.1), (2.5) and (2.6), and has been shown by Talagrand [11, Theorem 4.1] that, for any positive sequence λ_n, we have

$$\frac{\lambda_n}{\sqrt{\log n}} \to \infty \ \Rightarrow \ \frac{\sqrt{n}}{\lambda_n}\rho_1(L_n, \mu) \xrightarrow{p} 0 \ \Rightarrow \ \frac{\sqrt{n}}{\lambda_n}(L_n - \mu)^{\mathcal{F}} \xrightarrow{p} 0, . \tag{3.9}$$

where $\xrightarrow{p}$ denotes convergence in probability and μ denotes Lebesgue measure on the unit square. Now the following theorem is an easy consequence of [12, Theorem 2].

Theorem 3.2 *For any positive sequence λ_n such that*

$$\frac{\lambda_n}{\sqrt{n}} \to 0 \quad \textit{and} \frac{\lambda_n}{\sqrt{\log n}} \to \infty,$$

the sequence $P(\frac{T_n^1}{\sqrt{n}\lambda_n} \in \cdot)$ satisfies the MDP in $\mathbb{R}$ with speed λ_n^{-2} and rate function

$$J_1(x) = \inf\left\{\frac{1}{2}\int \left[(\frac{d\nu_1}{d\mu})^2 + (\frac{d\nu_2}{d\mu})^2\right] d\mu : \ \rho_1(\nu_1, \nu_2) = x\right\}.$$

References

[1] M. Ajtai, J. Komlós and G. Tusnády, On optimal matchings, *Combinatorica*, 4 (1984) 259–264.

[2] R. M. Dudley, Distances of probability measures and random variables, *Ann. Math. Stat.*, 39 (1968) 1563–1572.

[3] R. M. Dudley, The speed of mean Glivenko-Cantelli convergence, *Ann. Math. Stat.*, 40 (1969) 40–50.

[4] L. V. Kantorovich, On the transfer of masses, *Dokl. Acad. Nauk. SSSR*, 37 (1942) 227–229.

[5] L. V. Kantorovich, On a problem of Monge, *Uspekhi Mat. Nauk.*, 3 (1948) 225–226.

[6] L. V. Kantorovich and G. Sh. Rubinshtein, On a function space and some extremum problems, *Dokl. Acad. Nauk. SSSR*, 115 (1957) 1058–1061.

[7] T. Leighton and P.W.Shor, Tight bounds for minimax grid matching with applications to the average case analysis of algorithms, *Combinatorica*, 9 (1989) 161–187.

[8] 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) 666–704.

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

[10] P. W. Shor, *Random planar matching and bin packing*, Ph.D. thesis, M.I.T., 1985.

[11] M. Talagrand, Matching theorems and empirical discrepancy computations using majorizing measures, *J. Amer. Math. Soc.*, 7 (1994) 455–537.

[12] Liming Wu, Large deviations, moderate deviations and LIL for empirical processes, *Ann. Prob.*, 22 (1994) 17-27.