Large and moderate deviations for matching problems and empirical discrepancies

Ayalvadi Ganesh* Neil O'Connell†

March 8, 2006

ABSTRACT: We study the two-sample matching problem and its connections with the Monge-Kantorovich problem of optimal transportation of mass. We exploit this connection to obtain moderate and large deviation principles. For the classical problem on the unit square we present a conjecture which, if true, yields an explicit formula for the rate function.

KEYWORDS: Matching problems, empirical processes, empirical discrepencies, large deviations, marriage lemma, Monge-Kantorovich problem.

1 Introduction

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

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

where S_n denotes the set of permutations of $\{1, \ldots, n\}$. This is the canonical two-sample matching problem. Ajtai, Komlòs and Tusnàdy [1] prove the

*Microsoft Research Labs, Cambridge, UK

†Department of Mathematics/Boole Centre for Research in Informatics, University College Cork, Ireland. Research supported in part by Science Foundation Ireland Grant Number SFI04/RP1/I512.

following: 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 [12] and Talagrand [13], 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 [9] 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. Here

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

where $\mathcal{F}$ is taken to be the set of Lipschitz continuous functions on the unit square with Lipschitz constant 1 and $\mu(f)$ denotes $\int f d\mu$ (see, for example, [13]). The norm is defined on the set of signed measure μ such that $\mu\big([0,1]^2\big) = 0$. 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 [3]). In particular, Dudley [4] 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

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

This should be compared with (1.1).

The outline of the rest of the paper is as follows. In Section 2, we give some background on the Monge-Kantorovich problem and make the connections between matching and empirical discrepancy explicit. In Section 3, we use these connections to obtain large and moderate deviation principles for the sequences T_n^1 and T_n^∞. We conclude with a conjecture which, if true, leads to explicit formulae for the corresponding rate functions.

2 The Monge-Kantorovich Problem

In 1781, Monge [10] 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 [6, 7]. Suppose that μ and ν are Borel probability measures on a compact metric space (E, d) and $\Pi(\mu, \nu)$ 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

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

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 [11] for a monumental survey of the literature. In particular, it was shown by Kantorovich and Rubinshtein [8] that ρ_1 metrises the weak topology on $M_1(E)$, the space of probability measures on E. A related metric, which is also associated with the Monge-Kantorovich problem, is defined by

$$\rho_\infty(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)\}} \sup\{\text{supp } \pi\circ d^{-1}\}, \tag{2.6}$$

where supp $\pi\circ d^{-1}$ denotes the support of the probability measure $\pi\circ d^{-1}$, which is bounded since E is compact. Equivalently,

$$\rho_\infty(\mu,\nu) = \inf\{\epsilon>0:\ \mu(A)\le\nu(A^\epsilon), A\in\mathcal{B}(E)\}, \tag{2.7}$$

where $A^\epsilon = \{x:\ d(x,A)<\epsilon\}$ and $\mathcal{B}(E)$ is the Borel σ-algebra on E. The topology generated by ρ_∞ is finer than the weak topology. However, it can be shown (see Lemma 3.3 below) that, in the special case where E is the unit square in $\mathbb{R}^2$ and λ is Lebesgue measure on E, the function $\rho_\infty(\cdot,\lambda)$ is weakly continuous.

The matching problem is related to the Monge-Kantorovich problem by the following lemma.

Lemma 2.1 *For $x,y\in E^n$, set $l_n=\frac{1}{n}\sum_{i=1}^n\delta_{x_i}$ and $m_n=\frac{1}{n}\sum_{i=1}^n\delta_{y_i}$. We have the following identities:*

$$\rho_1(l_n,m_n) = \inf_{\sigma\in S_n}\frac{1}{n}\sum_{i=1}^n d(x_i,y_{\sigma(i)})$$

$$\rho_\infty(l_n,m_n) = \inf_{\sigma\in S_n}\max_{i\le n} d(x_i,y_{\sigma(i)}).$$

Proof. For $\sigma\in S_n$, set

$$\pi^\sigma = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_{\sigma(i)})}.$$

We will argue that any $\pi\in\Pi(l_n,m_n)$ can be written as

$$\pi = \sum_{\sigma\in S_n}\beta_\sigma\pi^\sigma,$$

for some collection of non-negative numbers β_σ with

$$\sum_{\sigma\in S_n}\beta_\sigma = 1.$$

From this it would follow that

$$\begin{aligned}
\rho_1(l_n,m_n) &= \inf\Big\{\int_{E\times E} d(x,y)\pi(dx,dy) :\ \pi\in\Pi(l_n,m_n)\Big\}\\
&= \inf\Big\{\sum_{\sigma\in S_n}\beta_\sigma\int_{E^2} d(x,y)\pi^\sigma(dx,dy) :\ \beta_\sigma>0,\ \sum_{\sigma\in S_n}\beta_\sigma = 1\Big\}\\
&= \inf_{\sigma\in S_n}\int_{E^2} d(x,y)\pi^\sigma(dx,dy)\\
&= \inf_{\sigma\in S_n}\frac{1}{n}\sum_{i=1}^n d(x_i,y_{\sigma(i)})
\end{aligned}$$

and

$$\begin{aligned}
\rho_\infty(l_n,m_n) &= \inf\big\{\sup\ \big\{\text{supp}\ \pi\circ d^{-1} :\ \pi\in\Pi(l_n,m_n)\big\}\big\}\\
&= \inf\Big\{\sup\Big\{\text{supp}\Big(\sum_{\sigma\in S_n}\beta_\sigma\pi^\sigma\circ d^{-1} :\ \beta_\sigma>0,\ \sum_{\sigma\in S_n}\beta_\sigma = 1\Big)\Big\}\Big\}\\
&= \inf_{\sigma\in S_n}\sup\{\text{supp}\ \pi^\sigma\circ d^{-1}\}\\
&= \inf_{\sigma\in S_n}\max_{i\le n} d(x_i,y_{\sigma(i)})
\end{aligned}$$

as required. The above claim is fact an immediate consequence of the Birkoff-von Neumann theorem (or the Krein-Milman theorem, of which it is a corollary), which states that any doubly stochastic matrix can be written as a convex combination of permutation matrices (see, for example, [2, Theorem 2.1.6]). (Recall that a non-negative matrix is *doubly stochastic* if each of its rows and columns sum to one; a *permutation matrix* of order n is a matrix of the form $1\{\sigma(i)=j\}$, for some $\sigma\in S_n$.) To see this, note that if $\pi\in\Pi(l_n,m_n)$ then

$$\pi = \frac{1}{n}\sum_{i=1}^n a_{ij}\delta_{(x_i,y_j)}$$

where $A=(a_{ij})$ is a doubly stochastic matrix. By the Birkoff-von Neumann theorem, there exist non-negative constants β_σ with $\sum_{\sigma\in S_n}\beta_\sigma = 1$, such

that

$$a_{ij} = \sum_{\sigma \in S_n} \beta_\sigma 1_{\sigma(i)=j}$$

for all $i, j \leq n$. It follows that

$$\pi = \sum_{\sigma \in S_n} \beta_\sigma \pi^\sigma,$$

as claimed. □

We thus have the following identities, where E is now taken to be the unit square:

$$T_n^1 = n\rho_1(L_n, M_n) = n\|L_n - M_n\|_{\mathcal{F}}$$

and

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

The first of these identities is the underlying force behind the work of Talagrand and Shor. In the next section we will use (more general versions of) these identities to obtain large and moderate deviation results for the matching problem.

3 Large and moderate deviations results

Let $(X_n)_{n\geq 1}$ 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 $B \in \mathcal{E}$,

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

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

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

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

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

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

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

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

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 (2.6). Since E is compact, ρ_1 induces the weak topology on $M_1(E)$. Note that the compactness of E is required. To see this, let X have a Pareto distribution on $\mathbb{R}$ with infinite mean and let X_n have the truncated distribution, namely that of $X \wedge n$. Clearly, X_n converges weakly to X but, for any coupling (X, X_n), we have

$$E[X - X_n] \ge E[(X - X_n)1(X > n)] = E[X1(X > n)] - n = \infty,$$

for all n. Hence $\rho_1(X, X_n)$ is not even finite, for any n.

Since ρ_1 induces the weak topology on $M_1(E)$, the following is an immediate consequence of Sanov's theorem and the contraction principle (taking products is continuous in the weak topology).

Theorem 3.1 *The sequence $P(T_n^1/n \in \cdot)$ satisfies the LDP in $\mathbb{R}$ with rate function*

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

Here, for probability measures μ and ν on E, $H(\nu|\mu)$ denotes the relative entropy or Kullback-Leibler divergence of ν with respect to μ. It is defined as $\int_E d\nu \log \frac{d\nu}{d\mu}$ if ν is absolutely continuous with respect to μ, and is defined to be infinite otherwise.

Due to the highly discontinuous nature of ρ_∞, we can only conjecture that the sequence $P(T_n^\infty \in \cdot)$ also satisfies the LDP, with rate function given by

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

We can, however, deduce the following 'grid-matching' version.

Theorem 3.2 *If E is the unit square and the points X_i are generated according to Lebesgue measure λ on E, then the sequence $P(\rho_\infty(L_n, \lambda) \in \cdot)$ satisfies the LDP in $\mathbb{R}$ with rate function*

$$I_\infty(x) = \inf\{H(\nu|\lambda) : \ \rho_\infty(\nu, \lambda) = x\}.$$

This is a simple consequence of Sanov's theorem, the contraction principle, and the next lemma.

Lemma 3.3 *If E is the unit square and λ is Lebesgue measure on E, then the function $\rho_\infty(\cdot, \lambda)$ is weakly continuous.*

Proof. Let ν_n be a sequence of probability measures on E converging weakly to ν. Define

$$d = \rho_\infty(\lambda, \nu), \quad d_1 = \liminf_{n\to\infty} \rho_\infty(\lambda, \nu_n), \quad d_2 = \limsup_{n\to\infty} \rho_\infty(\lambda, \nu_n).$$

Then we can find a subsequence $\nu_{n(k)}$ such that $\rho_\infty(\lambda, \nu_{n(k)}) \to d_1$ as $k \to \infty$. In other words, given $\epsilon > 0$, there is a K such that $\rho_\infty(\lambda, \nu_{n(k)}) < d_1 + \epsilon$ for all $k > K$. Hence,

$$\lambda(A) \le \nu_{n(k)}(A^{d_1+\epsilon}) \quad \forall A \in \mathcal{B}(E), \ \ k > K.$$

Now, using the fact that the $\nu_{n(k)}$ converge weakly to ν, we have

$$\lambda(A) \leq \limsup_{k\to\infty} \nu_{n(k)}(\overline{A^{d_1+\epsilon}}) \leq \nu(\overline{A^{d_1+\epsilon}}) \leq \nu(A^{d_1+2\epsilon}) \quad \forall A \in \mathcal{B}(E),$$

and so $\rho_\infty(\lambda,\nu) \leq d_1 + 2\epsilon$. Since $\epsilon > 0$ is arbitrary, we obtain

$$\rho_\infty(\lambda,\nu) \leq d_1 = \liminf_{n\to\infty} \rho_\infty(\lambda,\nu_n),$$

i.e., $\rho_\infty(\lambda,\cdot)$ is lower semicontinuous.

To show upper semicontinuity, we shall use the fact that the weak topology is metrised by Levy's metric,

$$D(\mu,\nu) = \inf\{\delta : \mu(F) \leq \nu(F^\delta) + \delta \ \forall \text{ closed } F \subseteq E\}.$$

Recall that $\lambda(A) \leq \nu(\overline{A^{d+\epsilon}})$ for any $A \in \mathcal{B}(E)$ and any $\epsilon > 0$, where $d = \rho_\infty(\lambda,\nu)$. Since ν_n converges weakly to ν, it follows that for any given $\epsilon > 0$, there is an N such that

$$\lambda(A^\epsilon) \leq \nu(\overline{A^{d+2\epsilon}}) \leq \nu_n(\overline{A^{d+2\epsilon+\epsilon^3}}) + \epsilon^3 \quad \forall \ A \in \mathcal{B}(E),\ n > N.$$

Suppose first that A^ϵ isn't the unit square, E. Then there is a point x in E such that $B(x,\epsilon/2)$, the circle of radius $\epsilon/2$ centred at x, is contained in A^ϵ but doesn't intersect A. Since λ is Lebesgue measure, it follows that

$$\lambda(A^\epsilon) \geq \lambda\big(A \cup B\big(x,\frac{\epsilon}{2}\big)\big) = \lambda(A) + \lambda\big(B\big(x,\frac{\epsilon}{2}\big)\big) = \lambda(A) + \frac{\pi}{4}\epsilon^2.$$

Thus, for ϵ sufficiently small,

$$\lambda(A) \leq \nu_n(A^{d+3\epsilon}) \quad \forall \ A \in \mathcal{B}(E),\ n > N.$$

The above inequality holds trivially, for all n, if $A^\epsilon = E$ since $\lambda(A) \leq 1 = \nu_n(E)$ by the fact that λ and ν_n are probability measures on E. Thus, we have shown that $\rho_\infty(\lambda,\nu_n) \leq d + 3\epsilon$ for all $n > N$, where $d = \rho_\infty(\lambda,\nu)$. But $\epsilon > 0$ is arbitrary, so we conclude that

$$\rho_\infty(\lambda,\nu) \geq \limsup_{n\to\infty} \rho_\infty(\lambda,\nu_n),$$

i.e., $\rho_\infty(\lambda, \cdot)$ is upper semicontinuous. This completes the proof of the lemma. $\square$

We are not able to explicitly solve the optimisation problem in Theorem 3.1 but we characterise the solution in the theorem below, following some definitions. Given a compact metric space (E, d) and a Borel probability measure μ on it, we denote by $\mathcal{F}_0$ the set of Lipschitz functions f on E, with Lipschitz constant 1, such that $\int_E f d\mu = 0$. For $\theta \in \mathbb{R}$, we define

$$\Lambda_f(\theta) = \log \mathbf{E}_\mu \left[e^{\theta f}\right] = \log \int_E e^{\theta f(x)} d\mu(x), \quad L_f(\theta) = \Lambda_f(\theta) + \Lambda_f(-\theta), \tag{3.8}$$

and for $x \in \mathbb{R}$, we set

$$\Lambda_f^*(x) = \sup_{\theta \in \mathbb{R}} \theta x - \Lambda_f(\theta), \quad L_f^*(x) = \sup_{\theta \in \mathbb{R}} \theta x - L_f(\theta).$$

Remark 3.4 *It is easily seen to be a consequence of the fact that E is compact and f Lipschitz that $\Lambda_f(\theta)$ and $L_f(\theta)$ are finite and continuously differentiable at all $\theta \in \mathbb{R}$, with*

$$\Lambda_f'(\theta) = \frac{\int_E f(x) e^{\theta f(x)} d\mu(x)}{\int_E e^{\theta f(x)} d\mu(x)} = \int_E f(x) e^{\theta f(x) - \Lambda_f(\theta)} d\mu(x), \tag{3.9}$$

$$L_f'(\theta) = \Lambda_f'(\theta) - \Lambda_f'(-\theta). \tag{3.10}$$

Moreover, $\Lambda_f'(0) = L_f'(0) = 0$ by the assumption that $f \in \mathcal{F}_0$.

Theorem 3.5 *In the context of Theorem 3.1,*

$$I_1(x) = \begin{cases} \inf_{f \in \mathcal{F}_0} L_f^*(x), & \text{if } x \geq 0, \\ +\infty, & \text{if } x < 0. \end{cases} \tag{3.11}$$

Note that if $x < 0$, then there do not exist probability measures ν_1 and ν_2 on (E, d) such that $\|\nu_1 - \nu_2\|_{\mathcal{F}} = x$, since $\|\cdot\|_{\mathcal{F}}$ is non-negative. It is now immediate from the definition of I_1 in Theorem 3.1 that $I_1(x) = \infty$. Next, if $x = 0$, the infimum in the definition of I_1 in Theorem 3.1 is zero and is achieved by $\nu_1 = \nu_2 = \mu$ (as the infimum clearly cannot be negative).

Moreover, it can easily be verified from the definitions that $L_f^*(0) = 0$ for any $f \in \mathcal{F}_0$, since $L_f'(0) = 0$ (see Remark 3.4) and $L_f(0) = 0$. Thus, $I_1(0) = \inf_{f\in\mathcal{F}_0} L_f^*(0) = 0$. Hence, it only remains to prove the theorem in the case $x > 0$, which we do following the next two lemmas.

Lemma 3.6 *For each $x > 0$, either $I_1(x) = \infty$ or there are Borel probability measures ν_1, ν_2 on (E, d) such that the infimum in the definition of I_1 in Theorem 3.1 is achieved at (ν_1, ν_2).*

Proof. Let $M_1(E)$ denote the set of Borel probability measures on E, equipped with the weak topology. $M_1(E)$ is compact since E was assumed to be compact and so $M_1(E) \times M_1(E)$ is compact in the product topology. $(\nu_1, \nu_2) \to \rho_1(\nu_1, \nu_2)$ is continuous on $M_1(E) \times M_1(E)$ since ρ_1 metrises the weak topology on $M_1(E)$. Hence,

$$\{(\nu_1, \nu_2) \in M_1(E) \times M_1(E) : \ \|\nu_1 - \nu_2\|_{\mathcal{F}} = x\},$$

being a closed subset of a compact set, is either empty or compact. Since $H(\cdot|\mu)$ is lower semicontinuous in the weak topology (see, for example, [5, Lemma 1.4.3]), it follows that the infimum in the definition of $I_1(x)$ in Theorem 3.1 is either infinite or is achieved by some ν_1, $\nu_2 \in M_1(E)$. □

Lemma 3.7 *Let $x > 0$, and let ν_1, ν_2 be probability measures on E such that $\rho_1(\nu_1, \nu_2) = x$. Then, there is an $f \in \mathcal{F}_0$ such that $\rho_1(\nu_1, \nu_2) = (\nu_1 - \nu_2)(f)$.*

Proof. Since ν_1, ν_2 are probability distributions, $(\nu_1 - \nu_2)(f)$ is unchanged by adding a constant to f. Hence, by (2.5),

$$\rho_1(\nu_1, \nu_2) = \sup_{f\in\mathcal{F}} (\nu_1 - \nu_2)(f) = \sup_{f\in\mathcal{F}_0} (\nu_1 - \nu_2)(f).$$

But $\mathcal{F}_0$ is a compact subset of $L^\infty(E)$ by the Arzela-Ascoli theorem, since it is closed, bounded (due to the compactness of E) and equicontinuous. Moreover, $f \to (\nu_1 - \nu_2)(f)$ is continuous on $L^\infty(E)$. Hence, the supremum in the equation above is attained at some $f \in \mathcal{F}_0$, as claimed. □

Proof of Theorem 3.5. Recall that we need only prove the theorem for $x > 0$. We first show that $I_1(x) \geq inf_{f \in \mathcal{F}_0} L_f^*(x)$ for all $x > 0$. If $I_1(x) = \infty$, then the inequality holds trivially. Hence, suppose that $I_1(x)$ is finite. Then, by Lemma 3.6, the infimum in the definition of $I_1(x)$ in Theorem 3.1 is attained by some $\nu_1,\ \nu_2 \in M_1(E)$ which are absolutely continuous with respect to μ. Furthermore, by Lemma 3.7, there is an $f \in \mathcal{F}_0$ such that

$$\rho_1(\nu_1, \nu_2) = (\nu_1 - \nu_2)(f) = x. \tag{3.12}$$

Fix one such f. Since $\nu_1, \nu_2 \ll \mu$, the last equality in (3.12) implies that ess-sup f − ess-sup $(-f) \geq x$, where the essential suprema are with respect to μ.

Let Λ_f and L_f be defined as in (3.8), with $f \in \mathcal{F}_0$ as above. Then, $L_f'(0) = 0$ as noted in Remark 3.4, and

$$\lim_{\eta \to \infty} L_f'(\eta) = \lim_{\eta \to \infty} \Lambda_f'(\eta) - \Lambda_f'(-\eta) = \text{ess-sup } f - \text{ess-sup } (-f) \geq x.$$

Hence, given $0 < \epsilon < x$, it follows from the continuity of L_f' that we can find a $\theta \in \mathbb{R}_+$ such that

$$L_f'(\theta) = x - \epsilon, \text{ and so } L_f^*(x - \epsilon) = \theta(x - \epsilon) - L_f(\theta). \tag{3.13}$$

For θ as above, define λ_θ and $\lambda_{-\theta}$ by

$$\frac{d\lambda_\theta}{d\mu}(z) = e^{\theta f(z) - \Lambda_f(\theta)}, \quad \frac{d\lambda_{-\theta}}{d\mu}(z) = e^{-\theta f(z) - \Lambda_f(-\theta)}. \tag{3.14}$$

Clearly, λ_θ and $\lambda_{-\theta}$ are probability measures on (E, d), $\nu_1 \ll \lambda_\theta$ and $\nu_2 \ll \lambda_{-\theta}$ since λ_θ and $\lambda_{-\theta}$ are equivalent to μ. Thus, by the definition of ν_1 and

ν_2,

$$\begin{aligned}
I_1(x) &= H(\nu_1|\mu) + H(\nu_2|\mu) \;=\; \int_E d\nu_1 \log \frac{d\nu_1}{d\mu} + \int_E d\nu_2 \log \frac{d\nu_2}{d\mu} \\
&= \int_E d\nu_1 \log \frac{d\nu_1}{d\lambda_\theta} + \int_E d\nu_2 \log \frac{d\nu_2}{d\lambda_{-\theta}} + \int_E d\nu_1 \log \frac{d\lambda_\theta}{d\mu} + \int_E d\nu_2 \log \frac{d\lambda_{-\theta}}{d\mu} \\
&= H(\nu_1|\lambda_\theta) + H(\nu_2|\lambda_{-\theta}) + \int_E (\theta f(z) - \Lambda_f(\theta)) d\nu_1(z) + \int_E (-\theta f(z) - \Lambda_f(-\theta)) d\nu_2(z) \\
&\geq \theta(\nu_1 - \nu_2)(f) - \Lambda_f(\theta) - \Lambda_f(-\theta) \;=\; \theta x - L_f(\theta) \\
&\geq L_f^*(x - \epsilon).
\end{aligned}$$

Here, the first inequality follows from the non-negativity of relative entropy and the fact that ν_1 and ν_2 are probability measures, the following equality from (3.12), and the last inequality from (3.13) and the non-negativity of θ and ϵ. Since the above holds for arbitrarily small $\epsilon > 0$, it follows from the lower semicontinuity of L_f^* (it is the supremum of continuous functions) that $L_f^*(x) \leq I(x)$. Hence, also

$$\inf_{g \in \mathcal{F}_0} L_g^*(x) \leq I_1(x). \tag{3.15}$$

Next, we prove the converse inequality, i.e., that $L_f^*(x) \geq I_1(x)$ for all $f \in \mathcal{F}_0$. If f is such that ess-sup f − ess-sup $(-f) < x$, then it can readily be verified that $L_f^*(x) = \infty$, and so the claimed inequality holds trivially. Suppose next that $f \in \mathcal{F}_0$ is such that ess-sup f − ess-sup $(-f) \geq x$. We proceed as before: for arbitrary $\epsilon \in (0, x)$, we choose θ as in (3.13) and define λ_θ and $\lambda_{-\theta}$ as in (3.14). Then,

$$\begin{aligned}
(\lambda_\theta - \lambda_{-\theta})(f) &= e^{-\Lambda_f(\theta)} \int_E f e^{\theta f} d\mu - e^{-\Lambda_f(-\theta)} \int_E f e^{-\theta f} d\mu \\
&= \Lambda_f'(\theta) - \Lambda_f'(-\theta) \;=\; L_f'(\theta) \;=\; x - \epsilon.
\end{aligned} \tag{3.16}$$

Hence, $\rho_1(\lambda_\theta, \lambda_{-\theta}) \geq x - \epsilon$, and $\alpha := (x-\epsilon)/\rho_1(\lambda_\theta, \lambda_{-\theta}) \in (0, 1]$. (It is strictly bigger than zero because we assumed that $\epsilon < x$, and $\rho_1(\lambda_\theta, \lambda_{-\theta})$ is finite by the assumption that E is compact.) Now, if we define $\tilde{\lambda}_\theta = \alpha\lambda_\theta + (1 - \alpha)\mu$

and $\tilde{\lambda}_{-\theta} = \alpha\lambda_{-\theta} + (1-\alpha)\mu$, then it is easy to see that $\rho_1(\tilde{\lambda}_\theta, \tilde{\lambda}_{-\theta}) = x - \epsilon$. Moreover, by the convexity of $H(\cdot|\mu)$, $H(\tilde{\lambda}_\theta|\mu) \le H(\lambda_\theta|\mu)$ and $H(\tilde{\lambda}_{-\theta}|\mu) \le H(\lambda_{-\theta}|\mu)$. Thus, we have

$$\begin{aligned} I_1(x-\epsilon) &\le H(\tilde{\lambda}_\theta|\mu) + H(\tilde{\lambda}_{-\theta}|\mu) \;\le\; H(\lambda_\theta|\mu) + H(\lambda_{-\theta}|\mu) \\ &= \int_E (\theta f - \Lambda_f(\theta))d\lambda_\theta + \int_E (-\theta f - \Lambda_f(-\theta))d\lambda_{-\theta} \\ &= \theta(\lambda_\theta - \lambda_{-\theta})(f) - \Lambda_f(\theta) - \Lambda_f(-\theta) \\ &= \theta(x-\epsilon) - L_f(\theta) \;=\; L_f^*(x-\epsilon). \end{aligned} \tag{3.17}$$

Next, we note that $L_f^*(\cdot)$ is non-decreasing on $\mathbb{R}_+$. This is immediate from the fact that L_f^* is non-negative (since $L_f(0) = 0$) and convex (since it is the supremum of linear functions), and that $L_f^*(0) = 0$ (since $L_f'(0) = 0$). Therefore, we obtain from (3.17) that $I_1(x-\epsilon) \le L_f^*(x)$ for all $\epsilon \in (0, x)$. But I_1 is lower semicontinuous since it is a rate function. Hence, letting ϵ decrease to zero, we get $I_1(x) \le L_f^*(x)$ for all $f \in \mathcal{F}_0$. Combining this with the converse inequality established in (3.15) completes the proof of the theorem. □

It is not as straightforward to obtain moderate deviation principles from Sanov's theorem because $\rho_1(\cdot,\cdot)$ and, in the case where E is the unit square and the points X_i, Y_i are generated according to Lebesgue measure λ on E, $\rho_\infty(\cdot, \lambda)$, are not continuous functions on the space of signed measures. We shall use results of Wu [14], 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 $\rho_\infty(L_n, \lambda)$ remains an open problem.

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

It is suggested by (1.1), (2.5) and the identity $T_n^1/n = \rho_1(L_n, M_n)$, and has been shown by Talagrand [13, Theorem 4.1] that, for any positive sequence a_n, we have

$$\frac{a_n}{\sqrt{\log n}} \to \infty \;\Rightarrow\; \frac{\sqrt{n}}{a_n}\rho_1(L_n,\lambda) \xrightarrow{p} 0 \;\Rightarrow\; \frac{\sqrt{n}}{a_n}(L_n-\lambda)^{\hat{\mathcal{F}}} \xrightarrow{p} 0, \qquad (3.18)$$

where $\xrightarrow{p}$ denotes convergence in probability and λ denotes Lebesgue measure on the unit square. Now the following theorem is an easy consequence of [14, Theorem 2].

Theorem 3.8 *For any positive sequence a_n such that*

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

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

$$J_1(x) = \inf\left\{\frac{1}{2}\int\left[(\frac{d\nu_1}{d\lambda})^2 + (\frac{d\nu_2}{d\lambda})^2\right]d\lambda : \; \nu_1(E) = \nu_2(E) = 0, \nu_1, \nu_2 \ll \lambda, \rho_1(\nu_1,\nu_2) = x\right\}.$$

The solution to the above variational problem is as follows.

Lemma 3.9 *In the context of Theorem 3.8,*

$$J_1(x) = \frac{x^2}{4\sup_{f\in\mathcal{F}_0}\int_E f^2 d\lambda}.$$

Proof. Let the supremum in the denominator above be attained at $f \in \mathcal{F}_0$ (recall that $\mathcal{F}_0$ is a compact subset of $L^\infty(E)$ and that $g \to \int_E g^2 d\lambda$ is continuous on $L^\infty(E)$). Fix $x > 0$ and let the signed measures λ_1, λ_2 on E be given by

$$\frac{d\lambda_1}{d\lambda}(z) = \frac{xf(z)}{2\int f^2 d\lambda}, \qquad \frac{d\lambda_2}{d\lambda}(z) = \frac{-xf(z)}{2\int f^2 d\lambda}.$$

It follows from the definition of $\mathcal{F}_0$ that $\lambda_1(E) = \lambda_2(E) = 0$. We also have

$$(\lambda_1 - \lambda_2)(f) \;=\; \frac{2x\int f^2 d\lambda}{2\int f^2 d\lambda} \;=\; x,$$

and so $\rho_1(\lambda_1, \lambda_2) = \|\lambda_1 - \lambda_2\|_{\mathcal{F}} \geq x$. Also,

$$\frac{1}{2}\int \Big[(\frac{d\lambda_1}{d\lambda})^2 + (\frac{d\lambda_2}{d\lambda})^2\Big] d\lambda \;=\; \frac{x^2}{2}\frac{2\int f^2 d\lambda}{(2\int f^2 d\lambda)^2} \;=\; \frac{x^2}{4\int f^2 d\lambda}.$$

Now, if $\rho_1(\lambda_1, \lambda_2) = y > x$, then defining $\tilde{\lambda}_i = x\lambda_i/y$, $i = 1, 2$, we have $\rho_1(\tilde{\lambda}_1, \tilde{\lambda}_2) = x$, and

$$\begin{aligned}\frac{1}{2}\int \Big[(\frac{d\tilde{\lambda}_1}{d\lambda})^2 + (\frac{d\tilde{\lambda}_2}{d\lambda})^2\Big] d\lambda &= \frac{x^2}{2y^2}\int \Big[(\frac{d\lambda_1}{d\lambda})^2 + (\frac{d\lambda_2}{d\lambda})^2\Big] d\lambda \\ &< \frac{1}{2}\int \Big[(\frac{d\lambda_1}{d\lambda})^2 + (\frac{d\lambda_2}{d\lambda})^2\Big] d\lambda.\end{aligned}$$

Thus, we have from Theorem 3.8 and the definition of f that

$$J_1(x) \leq \frac{x^2}{4\int f^2 d\lambda} = \frac{x^2}{4\sup_{g\in\mathcal{F}_0}\int g^2 d\lambda}. \tag{3.19}$$

Next, suppose ν_1, ν_2 belong to the set over which the infimum in the definition of J_1 is taken. Then, $\nu_1, \nu_2 \ll \lambda$. Let $g_1 = d\nu_1/d\lambda$ and $g_2 = d\nu_2/d\lambda$ be the corresponding density functions. We have for arbitrary $f \in \mathcal{F}_0$ that

$$\begin{aligned}\big[(\nu_1 - \nu_2)(f)\big]^2 = \Big(\int f(g_1 - g_2) d\lambda\Big)^2 \;&\leq\; \int f^2 d\lambda \int (g_1 - g_2)^2 d\lambda \\ &\leq 2\int f^2 d\lambda \int (g_1^2 + g_2^2) d\lambda,\end{aligned}$$

where the first inequality is the Cauchy-Schwarz inequality, and the second follows from the fact that $2|g_1 g_2| \leq g_1^2 + g_2^2$ pointwise. Since this inequality holds for all $f \in \mathcal{F}_0$, taking the supremum over $\mathcal{F}_0$ yields

$$\rho_1(\nu_1, \nu_2)^2 \leq 2\sup_{f\in\mathcal{F}_0}\int f^2 d\lambda \int \Big[(\frac{d\nu_1}{d\lambda})^2 + (\frac{d\nu_2}{d\lambda})^2\Big] d\lambda.$$

But $\rho_1(\nu_1, \nu_2) = x$. Since the above inequality holds for all ν_1, ν_2 in the set over which the infimum in the definition of J_1 is taken, it follows that

$$J_1(x) \geq \frac{x^2}{4\sup_{f\in\mathcal{F}_0}\int f^2 d\lambda}. \tag{3.20}$$

Combining this with (3.19) yields the claim of the lemma. □

4 A conjecture

Consider the case where E is the unit square, centred at the origin. By Theorem 3.9, the rate function governing the MDP for T_n^1 is given by $J_1(x) = x^2/4c$, where

$$c = \sup_{f \in \mathcal{F}_0} \int_E f^2 d\lambda. \tag{4.21}$$

We conjecture that, in (4.21), the supremum is achieved at $f = \varphi \in \mathcal{F}_0$, where

$$\varphi(x, y) = \frac{x + y}{\sqrt{2}}, \quad -\frac{1}{2} \leq x, y \leq \frac{1}{2}.$$

A simple calculation yields that $\int_E \varphi^2 d\lambda = 1/12$ and so our conjecture is that

$$J_1(x) = 3x^2.$$

Clearly, taking $f = \varphi$ yields the upper bound $J_1(x) \leq 3x^2$; an easy lower bound is obtained by noting that, for any $g \in \mathcal{F}_0$,

$$\begin{aligned} \int_{x \in E} g(x)^2 d\lambda(x) &= \frac{1}{2} \int_{x,y \in E} [g(x) - g(y)]^2 d\lambda(x) d\lambda(y) \\ &\leq \frac{1}{2} \int_{x,y \in E} \|x - y\|^2 d\lambda(x) d\lambda(y) \;=\; \frac{1}{6}, \end{aligned}$$

which implies $J_1(x) \geq 3x^2/2$.

It is also possible in this case to compute the rate function I_1 governing the LDP, but again we need to base this on a conjecture. This time, the conjecture is that, for all $\theta \in \mathbb{R}$, the supremum of $L_f(\theta)$ over $f \in \mathcal{F}_0$ is achieved at $f = \varphi$. If this is true, then

$$I_1(x) = L_\varphi^*(x)$$

for $x > 0$. To see this, note first that $I_1 \leq L_\varphi^*$, by Theorem 3.5. On the other hand, assuming the above conjecture, we have

$$\begin{aligned} I_1(x) &= \inf_{f \in \mathcal{F}_0} \sup_{\theta \in \mathbb{R}} \theta x - L_f(\theta) \\ &\geq \sup_{\theta \in \mathbb{R}} \inf_{f \in \mathcal{F}_0} \theta x - L_f(\theta) \\ &= \sup_{\theta \in \mathbb{R}} \theta x - L_\varphi(\theta), \end{aligned}$$

as claimed. It is an elementary calculation to show that

$$L_\varphi(\theta) = 2\log\left[\frac{4}{\theta^2}\left(\cosh\left(\frac{\theta}{4\sqrt{2}}\right) - 1\right)\right].$$

Note that both of the above conjectures are true if:

Conjecture. Let E be the unit square centred at the origin. For all positive integers k, the supremum of the $(2k)^{th}$ cumulant $\Lambda_f^{(2k)}(0)$ over $f \in \mathcal{F}_0$ is achieved at $f = \varphi$.

References

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

[2] R.B.Bapat and T.E.S.Raghavan. *Non-negative matrices and Applications.* Cambridge University Press, 1997.

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

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

[5] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, John Wiley &Sons, 1997.

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

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

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

[9] 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.

[10] 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.

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

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

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

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