Local minimax rates for closeness testing of discrete distributions

Joseph Lam-Weil [1] & Alexandra Carpentier [2] & Bharath K. Sriperumbudur [3]

[1] *Otto-von-Guericke University*
Universitaetspl. 2, 39106 Magdeburg, Germany
joseph.lam@ovgu.de
[2] *Otto-von-Guericke University*
Universitaetspl. 2, 39106 Magdeburg, Germany
alexandra.carpentier@ovgu.de
[3] *Pennsylvania State University*
Old Main, State College, PA 16801, USA
bks18@psu.edu

Résumé. Nous considérons le problème de comparaison de distributions entre deux échantillons dans un modèle de Poisson vectoriel. Ce modèle est connu pour être asymptotiquement équivalent à celui des distributions multinomiales. Le but est de distinguer si deux échantillons de données ont été tirés d'une même distribution inconnue ou si leurs distributions respectives sont séparées en norme L_1. Nous cherchons en particulier à adapter la vitesse de test à la forme des distributions inconnues. Ainsi, nous travaillons dans un cadre *minimax local*. A notre connaissance, nous fournissons la première vitesse minimax locale de test pour la distance de separation à des facteurs logarithmiques près, ainsi qu'un test qui l'atteint. En regard de la vitesse obtenue, le problème de test à deux échantillons est subtantiellement plus difficile que celui de test d'adéquation d'un seul échantillon à une loi connue dans de nombreux cas.

Mots-clés. Statistique mathématique.

Abstract. We consider the closeness testing (or two-sample testing) problem in the Poisson vector model — which is known to be asymptotically equivalent to the model of multinomial distributions. The goal is to distinguish whether two data samples are drawn from the same unspecified distribution, or whether their respective distributions are separated in L_1-norm. In this paper, we focus on adapting the rate to the shape of the underlying distributions, i.e. we consider *a local minimax setting*. We provide, to the best of our knowledge, the first local minimax rate for the separation distance up to logarithmic factors, together with a test that achieves it. In view of the rate, closeness testing turns out to be substantially harder than the related one-sample testing problem over a wide range of cases.

Keywords. Mathematical statistics.

The aim of this paper is to provide local minimax rates for the closeness testing (or two-sample testing) problem in the Poisson vector model. A related problem that has been thoroughly studied is the one-sample testing setting in the papers of Valiant and Valiant (2017), and Balakrishnan and Wasserman (2017a). While the two-sample problem has also been studied for example by Chan et al. (2014), and Diakonikolas and Kane (2016), as highlighted by Balakrishnan and Wasserman (2017b), we are not aware of a complete study of the local minimax rates as carried out in the one-sample setting in the papers of Valiant and Valiant (2017), and Balakrishnan and Wasserman (2017a). In this paper, we bridge this gap. In the following, we provide a formal setting of the question along with required notations.

1 Setting

For $n > 0$, define $\mathbf{P} = \{p \in (\mathbb{R}^+)^n, \sum_i p_i = 1\}$. Let $\|.\|_1$ denote the L_1-norm.

Let $p, q \in (\mathbb{R}^+)^n$, and $k \in \mathbb{N} \setminus \{0\}$. The data are obtained from the Poisson vector setting:

$$X_i \sim \mathcal{P}(kp_i), \quad Y_i \sim \mathcal{P}(kq_i), \tag{1}$$

where $\mathcal{P}$ is the Poisson distribution. In this paper, our goal is to test whether p and q are the same based on the data (X, Y), i.e. a closeness or two-sample testing problem. Note that when $p, q \in \mathbf{P}$, the Poisson vector setting is asymptotically equivalent to the setting where one receives k data from two multinomial distributions p, q - see the paper of Valiant and Valiant (2017). Therefore, our goal reduces to two-sample testing for multinomial distributions.

Given $\pi \in \mathbf{P}$, we define $\mathcal{U}_\pi$ as the discrete multinomial distribution that takes value $\{\pi_i\}$ with probability $1/n$ for each $i \in \{1, \dots, n\}$. Then, for a fixed $\rho > 0$ and fixed unknown $\pi \in \mathbf{P}$, the closeness testing problem that we consider in our paper is given by:

$$H_0^{(1)}(\pi) : p = q, \;\; q \sim \mathcal{U}_\pi^{\otimes n}, \quad \text{versus} \quad H_1^{(1)}(\pi, \rho) : \|p-q\|_1 \geq \rho, \;\; q \sim \mathcal{U}_\pi^{\otimes n}, p \in (\mathbb{R}^+)^n. \tag{2}$$

With the definition in Equation (2), the vectors that are too close to q are removed from the alternative hypothesis. We want to find the minimax optimal ρ such that a test with non-trivial error exists, *dependent on* π. Intuitively, if π is, for example, the uniform distribution, the testing problem is more difficult (and the minimax optimal separation distance larger) than if π has just a few non-zero atoms. We want to capture this effect, as done in the paper of Valiant and Valiant (2017) for one-sample testing.

Before describing our problem as well as the related literature in more details, we define the generic notions of separation distance and minimax sample complexity. Given a test φ whose inputs are k i.i.d. data points $\{(X_i, Y_i)\}_{i \leq k}$ distributed as in Equation (1), the generic risk for a testing problem with hypotheses H_0, H_1 is defined as the sum of type I and type II error probabilities:

$$R(H_0, H_1, \varphi; \rho, k) = \sup_{p,q \in H_0} \mathbb{P}_{p,q}(\varphi(\{(X_i, Y_i)\}_{i \leq k}) = 1) + \sup_{p,q \in H_1(\rho)} \mathbb{P}_{p,q}(\varphi(\{(X_i, Y_i)\}_{i \leq k}) = 0).$$

Then, fixing some $\gamma \in (0,1)$, we say that a testing problem can be solved with error less than γ, if we can construct a uniformly γ-consistent test, that is, if there exists φ such that:

$$R(H_0, H_1, \varphi; \rho, k) \leq \gamma.$$

Now $\rho \mapsto R(H_0, H_1, \varphi; \rho, k)$ is non-increasing, and greater or equal to one when $\rho = 0$. Then, we define the separation distance for some fixed $\gamma \in (0,1)$:

$$\rho_\gamma(H_0, H_1, \varphi; k) = \inf\{\rho > 0 : R(H_0, H_1, \varphi; \rho, k) \leq \gamma\}.$$

A good test φ is characterized by a small separation distance. So we define the minimax separation distance, also known as local critical radius, as

$$\rho^*_\gamma(H_0, H_1; k) = \inf_\varphi \rho_\gamma(H_0, H_1, \varphi; k).$$

Besides, it is possible to consider either local minimax rates or global minimax rates. Here, the local minimax separation distance would be written as $\rho^*_\gamma(H_0^{(1)}(\pi), H_1^{(1)}(\pi); k)$. On the other hand, global minimax separation distance and sample complexity are weaker compared to their local counterparts. They are written as $\sup_\pi \rho^*_\gamma(H_0^{(1)}(\pi), H_1^{(1)}(\pi); k)$ and $\sup_\pi k^*_\gamma(H_0^{(1)}(\pi), H_1^{(1)}(\pi); \rho)$, respectively.

Additional notations In what follows, we also establish the following notations. For a vector $u \in \mathbb{R}^n$, let s be a permutation of $\{1, \ldots, n\}$ be such that $u_{s(1)} \geq u_{s(2)} \geq \ldots \geq u_{s(n)}$. We write $u_{(.)} := u_{s(.)}$. Set also $J_u = \min_{j \leq n} \left\{ j : u_{(j)} \leq \frac{1}{k} \right\}$. We also write for $\gamma > 0$ and for $(a_k)_k, (b_k)_k$ two real sequences that $a_k = O_\gamma(b_k)$ if there exist $c_\gamma > 0, C_\gamma > 0$ that depend only on γ and such that $c_\gamma b_k \leq a_k \leq C_\gamma b_k$ for any k. We write $\tilde{O}^k_\gamma$ for the same concept but where the quantities c_γ, C_γ can be dependent of a polylog(nk) to a power that depends on γ only.

2 Literature review

The particular problem of goodness-of-fit testing, also known as identity testing, or one-sample testing, consists in distinguishing whether the data are drawn from a specified distribution π, versus a composite alternative separated from the null in L_1-distance:

$$H_0^{(3)}(\pi) : p = \pi, \qquad \text{versus} \qquad H_1^{(3)}(\pi) : \|p - \pi\|_1 \geq \rho,\ p \in \mathbf{P}. \tag{3}$$

The distributions considered will be restricted to certain classes of distributions. Indeed, there exist no consistent test that can distinguish an arbitrary distribution π from the alternatives separated in L_1 (Barron, 1989; LeCam et al., 1973).

The global minimax rate is given by Paninski (2008) and tightened by Valiant and Valiant (2017) for the class of multinomial distributions over a support of size n with the

L_1 distance for Problem (3). The global minimax separation distance for this problem is $\sup_\pi \rho^*_\gamma(H_0^{(3)}(\pi), H_1^{(3)}(\pi); k) = O_\gamma(n^{1/4}/\sqrt{k})$. This rate is obtained by taking π as a uniform distribution, which is the most difficult distribution for one-sample testing.

From the observation that the rates might take values substantially different from that of the worst case, the concept of minimaxity has been refined in recent lines of research. One such refinement corresponds to local minimaxity, also known as instance-optimality. Thus the rate depends on π. Valiant and Valiant (2017), and Balakrishnan and Wasserman (2017) obtained the minimax rate in the local setting. We formulate their lower bound in the following way in the present paper: $\rho^*_\gamma(H_0^{(3)}(\pi), H_1^{(3)}(\pi); k) \geq O_\gamma\Big(\min_m \Big[\frac{\|(\pi_{(i)}^{2/3}\mathbf{1}\{2\leq i<m\})_i\|_1^{3/4}}{\sqrt{k}} \vee \frac{1}{k} \vee \|(\pi_{(i)}\mathbf{1}\{i \geq m\})_i\|_1\Big]\Big)$.

On the other hand, in the closeness testing setting, the global minimax rate has been identified by Chan et al. (2014), using the tools developed by Valiant (2011). It corresponds in particular to an upper bound for Problem (2): $\sup_\pi \rho^*_\gamma(H_0^{(3)}(\pi), H_1^{(3)}(\pi); k) \leq O_\gamma(\frac{n^{1/2}}{k^{3/4}} \vee \frac{n^{1/4}}{k^{1/2}})$. A very interesting message from Chan et al. (2014) is that there exists a substantial difference between identity testing and closeness testing, and that the latter is harder. It is interesting to note that while the uniform distribution is the most difficult distribution π to test in Problem (3), π in Problem (2) can be chosen in a different appropriate way which worsens the rate.

Now, as explained in the review of Balakrishnan and Wasserman (2017b), the definition of local minimaxity in closeness testing is more involved than in identity testing, and in fact, is an interesting open problem that we focus on in this paper. The difficulty arises from the fact that both distributions are unknown, although we would like the rates to depend on them. Indeed, Problem (2) is composite-composite. Now, the existence and the size of the gap for every π due to this adaptivity constraint are open questions. We remind that Chan et al. (2014) disclose such a gap, but only in the worst case of π.

Diakonikolas and Kane (2016) construct a test which leads to an upper bound on the local minimax separation rate: $\rho^*_\gamma(H_0^{(3)}(\pi), H_1^{(3)}(\pi); k) \leq \tilde{O}^k_\gamma\Big(\frac{\|\mathbf{1}\{\pi<1/k\}\|_1^{1/2}\|\pi^2\mathbf{1}\{\pi<1/k\}\|_1^{1/4}}{\sqrt{k}} \vee \frac{\|\pi^{2/3}\|_1^{3/4}}{\sqrt{k}}\Big)$. Their bound matches the global minimax rate obtained by Chan et al. (2014) for some choice of π and m. However no matching lower bound is provided in the local case.

3 Contributions

The following are the major contributions of this work:

- We provide a lower bound on the local minimax separation distance for Problem (2), $\rho^*_\gamma(H_0^{(1)}(\pi), H_1^{(1)}(\pi); k)$ — see Equation (4) for $u > 2.001$.
- We propose a test that nearly reaches the obtained lower bound. This represents an upper bound on $\rho^*_\gamma(H_0^{(2)}(\pi), H_1^{(2)}(\pi); k)$ and $\rho^*_\gamma(H_0^{(1)}(\pi), H_1^{(1)}(\pi); k)$ — see Equation (4) for $u = 1$. So the test is almost local minimax near-optimal for Problem (2) which is related to closeness testing. An important feature of this test is that it does not need π as a parameter although it adapts to it.
- We point out the similarities and differences in regimes with local minimax identity testing.

More precisely we prove:

$$\rho^*_\gamma(H_0^{(1)}(\pi), H_1^{(1)}(\pi); k) = \tilde{O}^k_\gamma \Bigg\{ \min_{I \geq J_\pi} \left[\frac{\sqrt{I}}{k} \vee \left(\sqrt{\frac{I}{k}} \|\pi^2 \exp(-uk\pi)\|_1^{1/4} \right) \vee \|(\pi_{(i)} \mathbf{1}\{i \geq I\})_i\|_1 \right] \\ \vee \frac{\left\| (\pi_{(i)}^{2/3} \mathbf{1}\{i \leq J_\pi\})_i \right\|_1^{3/4}}{\sqrt{k}} \vee \sqrt{\frac{1}{k}} \Bigg\}, \tag{4}$$

where J_π and $\pi_{(.)}$ are defined in Section 1, $u = 2.001$ for the lower bound and $u = 1$ for the upper bound.

Bibliography

Balakrishnan, S. and Wasserman, L. (2017a). Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *arXiv preprint arXiv:1706.10003.*

Balakrishnan, S. and Wasserman, L. (2017b). Hypothesis testing for high-dimensional multinomials: A selective review. *arXiv preprint arXiv:1712.06120.*

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8, 577-606.

Barron, A. R. (1989). Uniformly powerful goodness of fit tests. *The Annals of Statistics*, pp. 107-124.

Chan, S.-O., Diakonikolas, I., Valiant, P., and Valiant, G. (2014). Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1193-1203. SIAM.

Diakonikolas, I. and Kane, D. M. (2016). A new approach for testing properties of discrete distributions. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 685-694. IEEE.

Goldreich, O., Goldwasser, S., and Ron, D. (1998). Property testing and its connection to learning and approximation. *Journal of the ACM (JACM)*, 45, 653-750.

Ingster, Y. and Suslina, I. A. (2012). *Nonparametric goodnessof-fit testing under Gaussian models*, 169. Springer Science & Business Media.

Ingster, Y. I. and Suslina, I. A. (1998). Minimax detection of a signal for Besov bodies and balls. *Problemy Peredachi Informatsii*, 34, 56-68.

LeCam, L. et al. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1, 38-53.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses.* Springer Science & Business Media.

Neyman, J. and Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A*, 231, 289-337.

Paninski, L. (2008). A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54, 4750-4755.

Rubinfeld, R. and Sudan, M. (1996). Robust characterizations of polynomials with applications to program testing. *SIAM Journal on Computing*, 25, 252-271.

Valiant, G. and Valiant, P. (2017). An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46, 429-455.

Valiant, P. (2011). Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40, 1927-1968.