

HIERARCHICAL LOW-RANK STRUCTURE OF PARAMETERIZED DISTRIBUTIONS

JUN QIN AND LEXING YING

ABSTRACT. This note shows that the matrix forms of several one-parameter distribution families satisfy a hierarchical low-rank structure. Such families of distributions include binomial, Poisson, and χ^2 distributions. The proof is based on a uniform relative bound of a related divergence function. Numerical results are provided to confirm the theoretical findings.

1. INTRODUCTION

This note is concerned with the matrix or operator form $f(x, \lambda)$ of a one-parameter distribution family indexed by the parameter λ . Such objects have long been considered in Bayesian statistics [2, 3, 13]. More recently, these matrices have played an important role in estimating distributions of distributions (also called fingerprints) [11, 12] and computing functionals of unknown distributions from samples [7, 8, 14]. When solving these problems, the computation often requires solving linear systems and optimizations problems associated with these matrices and operators.

In this note, we prove that, for several one-parameter family of distributions, including binomial, Poisson, and χ^2 distributions, $f(x, \lambda)$ exhibits a hierarchical low-rank structure. Roughly speaking, when viewed as a two-dimensional array, the off-diagonal blocks of $f(x, \lambda)$ are numerically low-rank, i.e., for a fixed accuracy ϵ , the numerical rank is bounded by a poly-logarithmic function of $1/\epsilon$. Such a structure ensures optimal complexity while approximating these matrices or performing basic linear algebra operations such as matrix-vector multiplications. In order to demonstrate the existence of such low-rank approximations, we first prove a new relative bound for a related divergence function, which might be of independent interest.

Similar hierarchical low-rank properties have been demonstrated for integral kernels [1, 4–6, 9] related to partial differential equations. For those kernels, the difficulty comes from the singularity along the diagonal. For the problems considered in this note, the location of the singularity is often near the boundary of the matrix/operator and thus the proof technique is quite different.

The rest of the note is organized as follows. Section 2 proves a relative bound of a related divergence function. Section 3 discusses the hierarchical low-rank structure of the negative exponentials of the divergence functions. Finally in Section 4 extends this result to parameterized distributions, including the binomial, Poisson, and χ^2 squared distributions.

The work of L.Y. is partially supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through Advanced Computing (SciDAC) program and also by the National Science Foundation under award DMS-1818449.

2. A RELATIVE BOUND FOR A DIVERGENCE FUNCTION

Consider the divergence function

$$E(p||q) \equiv p \ln(p/q) - (p - q)$$

for $0 \leq p, q < \infty$, which is convex and positive away from $p = q$. Let us first focus on the square $(p, q) \in (1, 2) \times (0, 1)$.

Theorem 1. For any $M > 0$, define p_M and q_M as follows:

- $q_M < 1$ is the value such that $E(1||q_M) = \ln 1/q_M - (1 - q_M) = M$.
- $p_M = \min(2, p')$ where $p' > 1$ is the number such that $E(p'||1) = p' \ln p' - (p' - 1) = M$.

There exists a uniform constant $C > 0$ such that for any $M > 0$

$$\frac{E(p_M||q_M)}{M} < C.$$

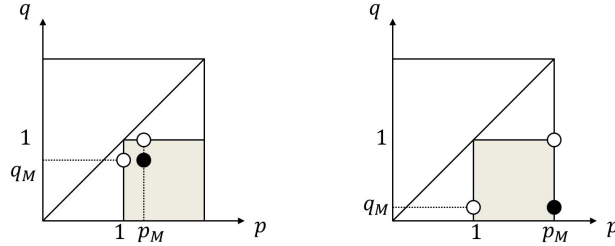


FIGURE 1. Locations of p_M and q_M for the case $(p, q) \in (1, 2) \times (0, 1)$. Left: small M . Right: large M .

Proof. Since the ratio $E(p_M||q_M)/M$ depends continuously on M , in order to show it is bounded by a uniform constant, it is sufficient to show that the ratio $E(p_M||q_M)/M$ has a finite limit as M goes to zero and to infinity.

When M goes to zero, the Taylor approximation of $E(p||q)$ near $p = 1$ and $q = 1$ is valid. The first order derivatives of $E(p||q)$ are

$$E_p = \ln p - \ln q, \quad E_q = -p/q + 1,$$

while the second order derivatives are

$$E_{pp} = 1/p, \quad E_{pq} = -1/q, \quad E_{qq} = p/q^2.$$

At the point $(p, q) = (1, 1)$,

$$E_p|_{(1,1)} = E_q|_{(1,1)} = 0, \quad E_{pq}|_{(1,1)} = E_{qq}|_{(1,1)} = 1, \quad E_{pp}|_{(1,1)} = -1.$$

Applying the definition of p_M and q_M shows that when M goes to zero

$$p_M = 1 + \sqrt{2M} + \text{h.o.t.} \quad q_M = 1 - \sqrt{2M} + \text{h.o.t.}$$

where h.o.t. stands for higher order terms (see Figure 1 (left)). Plugging them back to $E(p_M||q_M)$ and using the second order Taylor approximation shows

$$E(p_M||q_M) = 4M + \text{h.o.t.}$$

Therefore, when M goes to zero, the ratio $E(p_M||q_M)/M$ goes to 4.

When M goes to infinity, p_M goes to 2. From the definition, q_M satisfies

$$\ln(1/q_M) - (1 - q_M) = M.$$

Therefore, $q_M = e^{-(M+1)}(1 + \text{h.o.t.})$ (see Figure 1 (right)). Plugging them back to $E(p_M||q_M)$ shows that

$$E(p_M||q_M) = p_M \ln p_M/q_M - (p_M - q_M) = 2 \ln 2 + 2(M + 1) - 2 + \text{h.o.t.}$$

When M goes to infinity, the ratio $E(p_M||q_M)/M$ goes to 2.

Putting these two cases together proves the statement. □

Next, consider the square $(p, q) \in (0, 1) \times (1, 2)$.

Theorem 2. For any $M > 0$, now define p_M and q_M as follows:

- $q_M = \min(2, q')$ where $q' > 1$ satisfies $E(1||q') = \ln(1/q') - (1 - q') = M$.
- p_M is the minimum $p' \geq 0$ with $E(p'||1) = p' \ln p' - (p' - 1) \leq M$.

There exists a uniform constant $C > 0$ such that for any $M > 0$

$$\frac{E(p_M||q_M)}{M} < C.$$

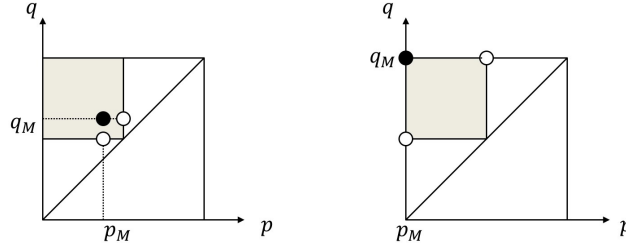


FIGURE 2. Locations of p_M and q_M for the case $(p, q) \in (0, 1) \times (1, 2)$. Left: small M . Right: large M .

Proof. Following the proof of the previous theorem, it is sufficient to show that the ratio has a limit as M goes to zero and infinity.

When M goes to zero, one can again use the second order Taylor expansion. Applying the definition of p_M and q_M , for sufficiently small M ,

$$p_M = 1 - \sqrt{2M} + \text{h.o.t.} \quad q_M = 1 + \sqrt{2M} + \text{h.o.t.}$$

(see Figure 2 (left)). Plugging them back to $E(p_M||q_M)$ and using again the Taylor approximation shows

$$E(p_M||q_M) = 4M + \text{h.o.t.}$$

Therefore, when M goes to zero, the ratio $E(p_M||q_M)/M$ goes to 4.

When M goes to infinity, p_M goes to 0 and q_M goes to 2 (see Figure 2 (right)). Plugging them back to $E(p_M||q_M)$ shows that

$$E(p_M||q_M) = 2 + \text{h.o.t.}$$

Therefore, when M goes to infinity, the ratio $E(p_M||q_M)/M$ goes to 0.

Putting these two cases together proves the statement. □

Remark 1. Theorems 1 and 2 also hold for the dual divergence of E defined as

$$E^*(p||q) = q \ln(q/p) - (q - p)$$

for $0 < p, q < \infty$ by simply switching the roles of p and q .

3. HIERARCHICAL LOW-RANK STRUCTURE OF NEGATIVE EXPONENTIAL OF DIVERGENCE

3.1. Divergence $E(p||q)$. Consider now the negative exponential of the divergence $E(p||q)$

$$(1) \quad \exp(-nE(p||q) = \exp(-n(p \ln(p/q) - (p - q)))$$

for $0 \leq p, q < \infty$ and any $n > 0$.

We consider a hierarchical decomposition that partitions the domain $(p, q) \in (0, \infty)^2$ into non-overlapping squares in a multiscale way. For each level ℓ indexed by integers, introduce the blocks $B_{\ell, k}$ defined as follows for $k = 0, 1, \dots$,

$$B_{\ell, k} = \begin{cases} [k/2^\ell, (k+1)/2^\ell] \times [(k+1)/2^\ell, (k+2)/2^\ell], & \text{for } k \text{ even,} \\ [k/2^\ell, (k+1)/2^\ell] \times [(k-1)/2^\ell, k/2^\ell], & \text{for } k \text{ odd.} \end{cases}$$

An illustration of this partitioning is shown in Figure 3 (left).

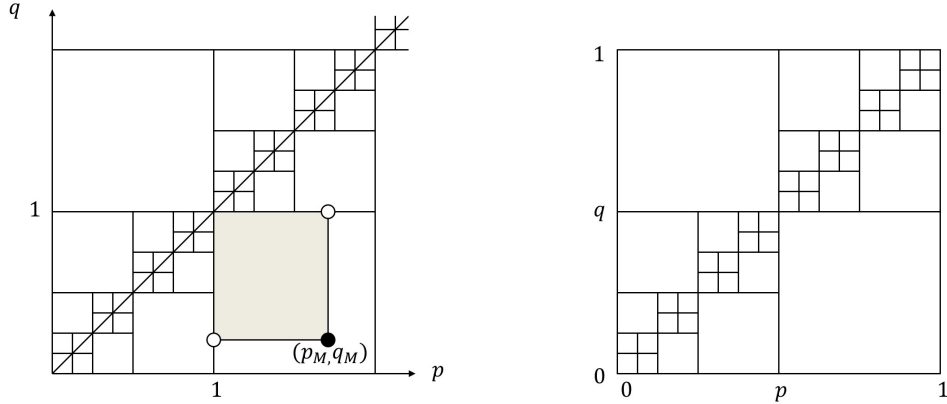


FIGURE 3. Left: Hierarchical decomposition for $\exp(-nE(p||q))$ for $(p, q) \in (0, \infty)^2$. Right: Hierarchical decomposition for $\exp(-nD(p||q))$ for $(p, q) \in (0, 1)^2$.

The main goal of this section is to prove the following theorem concerning the numerical rank of (1) restricted to each $B_{\ell, k}$.

Theorem 3. *For any $\epsilon > 0$, there exists a constant $T_\epsilon = O(\text{polylog}(1/\epsilon))$ such that for any $n > 0, \ell, k$ the restriction of $\exp(-nE(p||q))$ to $B_{\ell, k}$ has an $O(\epsilon)$ -accurate T_ϵ -term separated approximation. More precisely, there exists functions $\{\alpha_i(p)\}$ and $\{\beta_i(q)\}$ for $1 \leq i \leq T_\epsilon$ such that in $B_{\ell, k}$*

$$\exp(-nE(p||q)) = \sum_{i=1}^{T_\epsilon} \alpha_i(p) \beta_i(q) + O(\epsilon).$$

Proof. Consider first the blocks $B_{\ell, k}$ with k odd. These blocks are below the diagonal $p = q$. The top left corner of $B_{\ell, k}$ is (c, c) with $c = k/2^\ell$. Let us make two key observations.

- It is sufficient to prove the theorem for the restriction of $\exp(-nE(p||q))$ to $(c, 2c) \times (0, c)$ as the later contains $B_{\ell,k}$.
- The second observation is that, as the statement is uniform in n , it is sufficient to scale the box $(c, 2c) \times (0, c)$ to $(1, 2) \times (0, 1)$ by scaling the value of n accordingly.

Based on these two observations, it is sufficient to consider the box $(1, 2) \times (0, 1)$ for any $\epsilon > 0$ and any $n > 0$.

For fixed $\epsilon > 0$ and $n > 0$, define $M = \frac{1}{n} \ln \frac{1}{\epsilon}$. Applying Theorem 1 along with the definition of p_M and q_M gives

$$E(p_M||q_M) \leq CM = C \frac{1}{n} \ln \frac{1}{\epsilon}$$

and by monotonicity

$$E(p||q) \leq CM = C \frac{1}{n} \ln \frac{1}{\epsilon}, \quad \forall (p, q) \in [1, p_M] \times [q_M, 1].$$

In order to construct a separated approximation of $\exp(-n(p \ln(p/q) - (p - q)))$, we resort to the polynomial expansion for $(p, q) \in [1, p_M] \times [q_M, 1]$.

In order for this, consider the function $\exp(-x)$ in $x \in [0, L]$ for some $L > 0$. Using the Lagrange interpolation at the Chebyshev grids in $[0, L]$ and the uniform bound of the derivatives of $\exp(-x)$ (see for example Theorem 8.7 of [10]), we know that there exists a degree $d = O(\ln L + \ln(1/\epsilon))$ polynomial $h_d(x)$ such that

$$\exp(-x) - h_d(x) = O(\epsilon).$$

Plugging $x = nE(p, q)$ for $(p, q) \in [1, p_M] \times [q_M, 1]$ with the bound $L = n \cdot C \frac{1}{n} \ln \frac{1}{\epsilon} = C \ln(1/\epsilon)$, one arrives at

$$\exp(-nE(p||q)) - h_d(nE(p||q)) = O(\epsilon),$$

with $d = O(\ln(1/\epsilon))$. As $E(p||q) = p \ln p - p \ln q - p + q$, by expanding the polynomial $h_d(\cdot)$, we obtain a $O(\text{polylog}(1/\epsilon))$ -term separated approximation to $\exp(-nE(p||q))$ for $(p, q) \in [1, p_M] \times [q_M, 1]$. The individual terms define the functions $\{\alpha_i(p)\}$ for $p \in [1, p_M]$ and $\{\beta_i(q)\}$ for $q \in [q_M, 1]$, respectively.

For any point $(p, q) \in (1, 2) \times (0, 1)$ but outside $[1, p_M] \times [q_M, 1]$, as $\exp(-nE(p||q)) \leq \epsilon$, one can simply approximate it by zero without introducing an error larger than ϵ . In terms of the functions $\alpha_i(p)$ and $\beta_i(q)$, we simply define $\alpha_i(p)$ to be zero in $[p_M, 2]$ and $\beta_i(q)$ to be zero in $q \in [0, q_M]$, respectively.

Next, we consider the blocks $B_{\ell,k}$ with k even. These are the blocks above the diagonal $p = q$. The above argument goes through except that Theorem 2 is invoked. □

Remark 2. The same theorem is true for

$$\exp(-nE^*(p||q)) \equiv \exp(-n(q \ln(q/p) - (q - p))),$$

for $0 < p, q < \infty$ by switching the roles of p and q .

Remark 3. The same theorem is true for

$$\exp(-nE(1 - p||1 - q))$$

for $-\infty < p, q < 1$ with a similar hierarchical partitioning of the domain $-\infty < p, q < 1$.

3.2. Kullback-Leibler divergence. The Kullback-Leibler (KL) divergence of two Bernoulli distributions with parameters $p, q \in [0, 1]$ is defined as

$$D(p||q) \equiv p \ln(p/q) + (1-p) \ln((1-p)/(1-q)).$$

This section proves the hierarchical low-rank property for

$$\exp(-nD(p||q) = \exp(-n(p \ln(p/q) + (1-p) \ln((1-p)/(1-q))))$$

with $0 < p, q < 1$. For the domain $(p, q) \in [0, 1] \times [0, 1]$, the hierarchical decomposition needs to be restricted to

$$\ell \geq 1, \quad k = 0, 1, \dots, 2^\ell - 1.$$

An illustration of this partitioning is shown in Figure 3 (right).

Theorem 4. *For any $\epsilon > 0$, there exists a constant $S_\epsilon = O(\text{polylog}(1/\epsilon))$ such that for any $n > 0, \ell \geq 1, k = 0, 1, \dots, 2^\ell - 1$, the restriction of $\exp(-nD(p||q))$ to $B_{\ell,k}$ has an $O(\epsilon)$ -accurate S_ϵ -term separated approximation. More precisely, there exists functions $\{\alpha_i(p)\}$ and $\{\beta_i(q)\}$ for $1 \leq i \leq S_\epsilon$ such that in $B_{\ell,k}$*

$$\exp(-nD(p||q)) = \sum_{i=1}^{S_\epsilon} \alpha_i(p) \beta_i(q) + O(\epsilon).$$

Proof. The proof is based on a simple observation: $D(p||q) = E(p||q) + E(1-p||1-q)$, which implies

$$\exp(-nD(p||q)) = \exp(-nE(p||q)) \exp(-nE(1-p||1-q)).$$

From Theorem 3 and the remarks right after, the following two estimates hold for each $B_{\ell,k}$.

$$\begin{aligned} \exp(-nE(p||q)) &= \sum_{i=1}^{T_\epsilon} \alpha_i(p) \beta_i(q) + O(\epsilon), \\ \exp(-nE(1-p||1-q)) &= \sum_{j=1}^{T_\epsilon} \alpha'_j(p) \beta'_j(q) + O(\epsilon), \end{aligned}$$

Taking the product of these two expansions and using the fact that each expansion is bounded by $1 + O(\epsilon)$ results in

$$\exp(-nD(p||q)) = \sum_{i,j=1}^{T_\epsilon} (\alpha_i(p) \alpha'_j(p)) (\beta_i(q) \beta'_j(q)) + O(\epsilon).$$

Noticing that T_ϵ^2 is still of order $O(\text{polylog}(1/\epsilon))$, setting $S_\epsilon = T_\epsilon^2$ completes the proof. \square

4. PARAMETERIZED DISTRIBUTIONS

In this section, we apply the theorems in Section 3 to demonstrate the hierarchical low-rank property for three commonly-encountered distribution families.

4.1. Binomial distribution. The binomial distribution with parameter $q \in [0, 1]$ and n trials is

$$f(k, q) = \binom{n}{k} q^k (1 - q)^{n-k},$$

for $k \in \{0, \dots, n\}$. By introducing $p = k/n$, we can rewrite the binomial distribution in the form

$$f(p, q) = \binom{n}{np} q^{np} (1 - q)^{n-np}$$

with $p = 0, \frac{1}{n}, \dots, 1$. Applying the Stirling formula to the factorials results in

$$f(p, q) = c_{n,p} \frac{q^{np} (1 - q)^{n-np}}{p^{np} (1 - p)^{n-np}} = c_{n,p} \exp \left(-n \left(p \ln \frac{p}{q} + (1 - p) \ln \frac{1 - p}{1 - q} \right) \right),$$

where $c_{n,p} \approx \frac{1}{\sqrt{2\pi n}} \frac{1}{\sqrt{p(1-p)}}$ except at $p = 0$ and $p = 1$.

Applying Theorem 4 to this case shows that the binomial distribution $f(p, q)$ for $p = 0, 1/n, \dots, 1$ and $q \in [0, 1]$ has the hierarchical low-rank property. Here the two points $p = 0$ and 1 can be treated separately without affecting the rank estimates. Figure 4 plots the numerical rank of different blocks for a specific choice of n and ϵ (left) and its dependence on ϵ (right). Note that the rank is bounded by 10 even for $\epsilon = 10^{-9}$ and the dependence of the rank on $\ln(1/\epsilon)$ is linear.

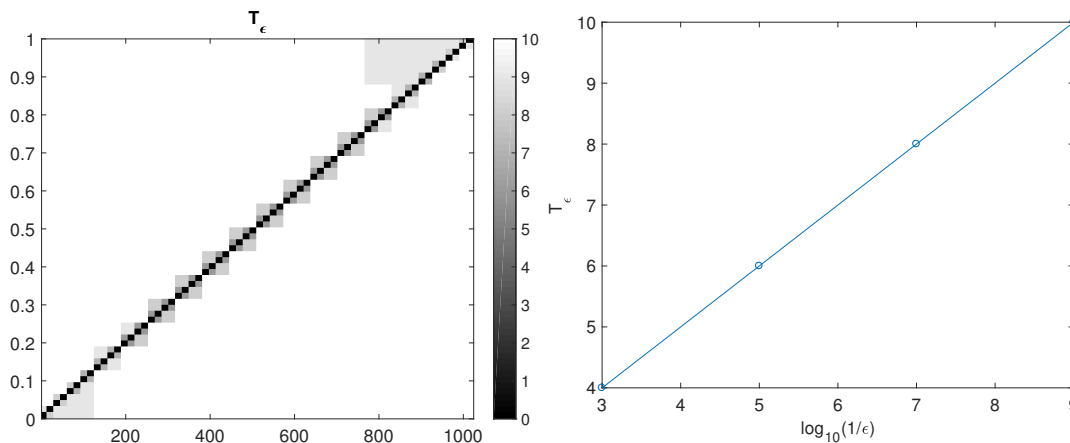


FIGURE 4. Binomial distribution. Left: the numerical rank T_ϵ of different blocks with $n = 2^{10}$ and $\epsilon = 10^{-9}$. Right: the maximum of the numerical ranks T_ϵ as a function of ϵ with $n = 2^{10}$.

4.2. Poisson distribution. The Poisson distribution with parameter $\lambda > 0$ is

$$f(k, \lambda) = e^{-\lambda} \frac{\lambda^k}{k!}$$

for $k \in \{0, 1, \dots\}$. Applying the Stirling formula to $k!$ gives for $k > 0$

$$f(k, \lambda) \approx \frac{1}{\sqrt{2\pi k}} \exp(-(k \log(k/\lambda) - (k - \lambda))).$$

By identifying $p = k$ and $q = \lambda$, this is the negative exponential of the divergence $E(p||q)$ with $n = 1$ in Section 3.1, modulus the term $\frac{1}{\sqrt{2\pi k}}$.

Applying Theorem 3 shows that the Poisson distribution $f(k, \lambda)$ for $k = 0, 1, \dots$ and $\lambda > 0$ exhibits the hierarchical low-rank property. Figure 5 shows the numerical rank of different blocks for a specific choice of ϵ (left) and its dependence on ϵ (right). Note that the rank is bounded by 10 even for $\epsilon = 10^{-9}$ and there is strong evidence that the dependence of the rank on $\ln(1/\epsilon)$ is linear.

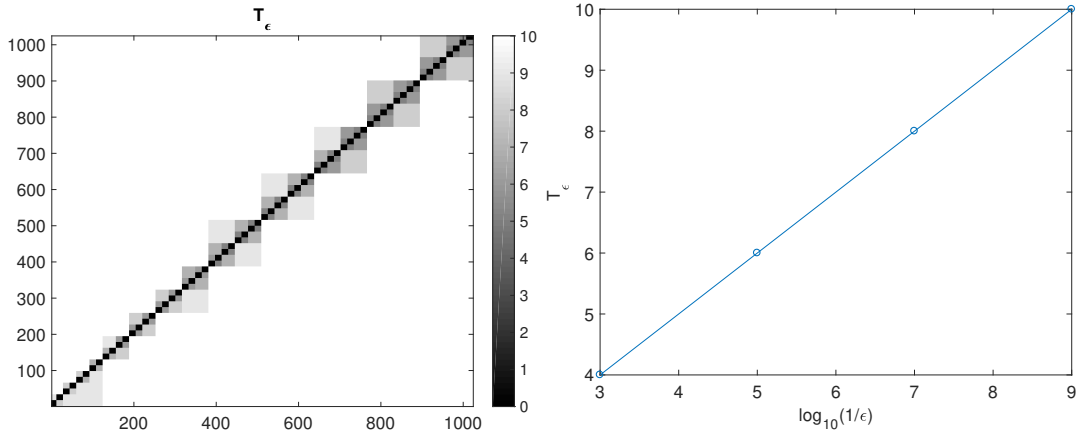


FIGURE 5. Poisson distribution. Left: numerical rank T_ϵ of different blocks with $k, \lambda \leq 2^{10}$ and $\epsilon = 10^{-9}$. Right: the maximum of the numerical ranks T_ϵ as a function of ϵ .

4.3. χ^2 **distribution.** The χ^2 distribution, parameterized by integer $k \geq 1$ is

$$f(x, k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}.$$

for $x > 0$. Applying again the Stirling formula shows that

$$f(x, k) \approx \frac{1}{2\sqrt{2\pi}\left(\frac{k}{2} - 1\right)} \exp\left(-\left(\frac{x}{2} - \left(\frac{k}{2} - 1\right) + \left(\frac{k}{2} - 1\right) \ln\left(\frac{k/2 - 1}{x/2}\right)\right)\right).$$

By identifying $x/2 = p$ and $k/2 - 1 = q$, this is

$$\exp(-(q \ln(q/p) - (q - p)))$$

modulus the factor $\frac{1}{2\sqrt{2\pi}(k/2-1)}$.

Applying the remark after Theorem 3 shows that the χ^2 distribution exhibits the hierarchical low-rank property. Figure 6 plots the numerical rank of different blocks for a specific choice of ϵ (left) and its dependence on ϵ . Again, the rank remains small even for $\epsilon = 10^{-9}$ and the dependence of the rank on $\ln(1/\epsilon)$ is linear.

5. DISCUSSIONS

The hierarchical low-rank property has significant numerical implications for these distribution families. Naive approaches for representing the matrix form of these distributions would require $O(n^2)$ numbers. Even by thresholding small entries, it would still need at least

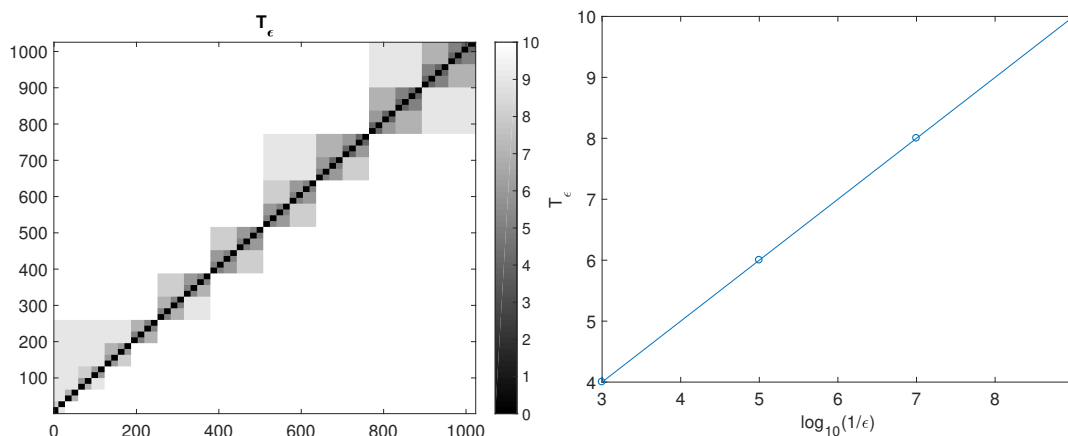


FIGURE 6. χ^2 distribution. Left: numerical rank T_ϵ of different blocks with $x, k \lesssim 2^{10}$ and $\epsilon = 10^{-9}$. Right: the maximum of the numerical ranks T_ϵ as a function of ϵ .

$O(n^{3/2})$ storage space for most of these distributions. The hierarchical low-rank property proved here allows for storing the matrix with no more than $O(n \log n \text{ polylog}(1/\epsilon))$ entries. By combining the low-rank property with thresholding, this can potentially be brought down to $O(n \text{ polylog}(1/\epsilon))$.

The theorems proved here show an $O(\text{polylog}(1/\epsilon))$ upper bound for the numerical ranks. However, the numerical results suggest that the actual dependence on $\log(1/\epsilon)$ seems to be linear. An immediate direction for future work is to obtain sharper bounds for the rank growth.

REFERENCES

- [1] Mario Bebendorf and Wolfgang Hackbusch, *Existence of ϵ -matrix approximants to the inverse fe-matrix of elliptic operators with l -coefficients*, Numerische Mathematik **95** (2003), no. 1, 1–28.
- [2] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin, *Bayesian data analysis*, Chapman and Hall/CRC, 2013.
- [3] Subhashis Ghosal and Aad Van der Vaart, *Fundamentals of nonparametric bayesian inference*, Vol. 44, Cambridge University Press, 2017.
- [4] Leslie Greengard and John Strain, *The fast gauss transform*, SIAM Journal on Scientific and Statistical Computing **12** (1991), no. 1, 79–94.
- [5] Leslie F Greengard, *The rapid evaluation of potential fields in particle systems, acm distinguished dissertation 1987*, MIT Press, 1988.
- [6] Wolfgang Hackbusch, *Hierarchical matrices: algorithms and analysis*, Vol. 49, Springer, 2015.
- [7] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman, *Minimax estimation of functionals of discrete distributions*, IEEE Transactions on Information Theory **61** (2015), no. 5, 2835–2885.
- [8] Liam Paninski, *Estimation of entropy and mutual information*, Neural computation **15** (2003), no. 6, 1191–1253.
- [9] Vladimir Rokhlin, *Rapid solution of integral equations of classical potential theory*, Journal of computational physics **60** (1985), no. 2, 187–207.
- [10] Endre Süli and David F Mayers, *An introduction to numerical analysis*, Cambridge university press, 2003.
- [11] Kevin Tian, Weihao Kong, and Gregory Valiant, *Learning populations of parameters*, Advances in neural information processing systems, 2017, pp. 5778–5787.
- [12] Paul Valiant and Gregory Valiant, *Estimating the unseen: improved estimators for entropy and other properties*, Advances in neural information processing systems, 2013, pp. 2157–2165.

- [13] Larry Wasserman, *All of statistics: a concise course in statistical inference*, Springer Science & Business Media, 2013.
- [14] Yihong Wu and Pengkun Yang, *Minimax rates of entropy estimation on large alphabets via best polynomial approximation*, IEEE Transactions on Information Theory **62** (2016), no. 6, 3702–3720.

(Jun Qin) TARGET CORPORATION, SUNNYVALE, CA, 94086
E-mail address: `jun.qin@target.com`

(Lexing Ying) DEPARTMENT OF MATHEMATICS AND ICME, STANFORD UNIVERSITY, STANFORD, CA 94305
E-mail address: `lexing@stanford.edu`