A. Proof of Theorem 1

Proof. We derived in the main text that $r(\mathbf{A}_{Q_\Theta}) \leq d+1$. In addition, Eckart-Young-Mirsky theorem gives:

$$\|\mathbf{A}_{P^*} - \mathbf{B}\|_F^2 \geq \sqrt{\sigma_{d+2}^2 + \ldots + \sigma_M^2},$$

$$\forall \mathbf{B} \in \mathbb{R}^{M \times N} \; s.t. \; r(\mathbf{B}) \leq d+1$$

Thus, our result follows for $\mathbf{B} = \mathbf{A}_{Q_\Theta}$. □

B. Proof of Theorem 2

Proof. i) Using the non-negativity property of the KL divergence, one derives:

$$KL(R\|Q) = H(R,Q) - H(R) \geq 0 \tag{16}$$

for any probability distribution R. The result follows easily by taking $R = P^*$.

ii) Let $Q_\mathbf{h}(x_i) \propto \exp(\langle \mathbf{w}_i, \mathbf{h}\rangle)$. Then, for any probability distribution R, it is straightforward to derive that

$$H(R, Q_\mathbf{h}) = -\langle \mathbb{E}_R[\mathbf{w}], \mathbf{h}\rangle + \log Z^{(\mathbf{h})} \tag{17}$$

Moreover, if $R \in \mathcal{P}^*$ is any distribution satisfying the d-dimensional linear constraints, one derives from eq. (17) that

$$H(P^*, Q_\mathbf{h}) = H(R, Q_\mathbf{h}), \; \forall R \in \mathcal{P}^* \tag{18}$$

combining eqs. (16) and (18), we get:

$$H(P^*, Q_\mathbf{h}) \geq H(R), \; \forall R \in \mathcal{P}^* \tag{19}$$

thus

$$H(P^*, Q_\mathbf{h}) \geq max_{R\in\mathcal{P}^*} H(R) \tag{20}$$

which, since $Q_\mathbf{h}$ is arbitrary in the above exponential family, implies that

$$min_\mathbf{h} H(P^*, Q_\mathbf{h}) \geq max_{R\in\mathcal{P}^*} H(R) \tag{21}$$

We are only left with proving the reverse, namely that $min_\mathbf{h} H(P^*, Q_\mathbf{h}) \leq max_{R\in\mathcal{P}^*} H(R)$. We use the standard derivations for the Maximum Entropy Principle, namely we form the Lagrangian:

$$L(\boldsymbol{\lambda}, \beta, \mathbf{h}) := H(R) + \beta\left(\sum_{i=1}^{M} R(x_i) - 1\right) + \\ + \langle \boldsymbol{\lambda}, \mathbb{E}_R[\mathbf{w}] - \mathbb{E}_{P^*}[\mathbf{w}]\rangle \tag{22}$$

Setting its derivatives to 0, one gets that the optimal $R^* = \arg\max_{R\in\mathcal{P}^*} H(R)$ has the form

$$R^*(x_i) \propto \exp(\langle \mathbf{w}_i, \boldsymbol{\lambda}^*\rangle) \tag{23}$$

for some $\boldsymbol{\lambda}^* \in \mathbb{R}^d$ that is chosen by solving the d-linear system $\mathbb{E}_{R^*}[\mathbf{w}] - \mathbb{E}_{P^*}[\mathbf{w}] = 0$. One can observe that $Q_{\boldsymbol{\lambda}^*} = R^*$, getting

$$min_\mathbf{h} H(P^*, Q_\mathbf{h}) \leq H(P^*, Q_{\boldsymbol{\lambda}^*}) = H(P^*, R^*)$$

Finally, using eq. (18), we get:

$$H(P^*, R^*) = H(R^*, R^*) = H(R^*) = max_{R\in\mathcal{P}^*} H(R)$$

which concludes the proof. □

C. Proof of Theorem 3

Proof. Since $f(\mathbf{A})$ has rank at least K, there exists at least one submatrix $\mathbf{M} \in \mathbb{R}^{K\times K}$ of $\mathbf{A}$ such that $\det(f(\mathbf{M})) \neq 0$. Let $b_1 < b_2 < \ldots < b_T$ be all the distinct values of $\mathbf{M}$. Denote by $\epsilon = \frac{1}{4}\min_{i>1}|b_i - b_{i-1}|$. We first prove the following lemmas.

Lemma 8. *Let $P \in \mathbb{R}[X_1, \ldots, X_T]$ be a multivariate polynomial with real coefficients. Assume there exist infinite sets $S_1, \ldots, S_T$ such that P vanishes on all the points of $S_1 \times S_2 \times \ldots \times S_T$. Then P vanishes on any point of $\mathbb{R}^T$.*

Proof. We prove this by induction over T. The result easily holds for $T = 1$ since a real univariate non-zero polynomial can only have a finite set of roots. Assume now that the result holds for any polynomial in $T-1$ variables. We can write $P(X_1, X_2, \ldots, X_T)$ as a univariate polynomial in X_1 with coefficients polynomials in $\mathbb{R}[X_2, \ldots, X_T]$ as follows: $P(X_1, X_2, \ldots, X_T) = \sum_{i=0}^{d_1} Q_i(X_2, \ldots, X_T)X_1^i$, where d_1 is the maximum degree of X_1. For any arbitrary $x_2, \ldots, x_T \in S_2 \times \ldots \times S_T$, we know from the hypothesis that $P(c, x_2, \ldots, x_T) = 0, \forall c \in S_1$. Since S_1 is infinite we have that the univariate polynomial in X_1 is identical 0, i.e. $P(X, x_2, \ldots, x_T) \equiv 0$, which implies that $Q_i(x_2, \ldots, x_T) = 0$. However, $x_2, \ldots, x_T \in S_2 \times \ldots \times S_T$ were chosen arbitrarily, thus $Q_i(x_2, \ldots, x_T) = 0, \forall x_2, \ldots, x_T \in S_2 \times \ldots \times S_T$. Applying the induction hypothesis for $T-1$, one gets that all Q_i vanish on the full $\mathbb{R}^{T-1}$. Thus, $P(X, x_2, \ldots, x_T) \equiv 0, \forall (x_2, \ldots, x_T) \in \mathbb{R}^{T-1}$, which implies that $P(x_1, x_2, \ldots, x_T) = 0, \forall (x_1, x_2, \ldots, x_T) \in \mathbb{R}^T$. □

Lemma 9. *There exist $c_i \in [b_i - \epsilon, b_i + \epsilon], \forall i \in \{1, \ldots, T\}$ s.t. given any pointwise function h satisfying $h(b_i) = c_i, \forall 1 \leq i \leq T$, we have $\det(h(\mathbf{M})) \neq 0$.*

Proof. Assume the contrary, that $\forall c_i \in [b_i - \epsilon, b_i + \epsilon]$, $\det(h(\mathbf{M})) = 0$.

We note that, using the Leibniz formula of the determinant, one easily sees that $\det(\mathbf{M})$ can be written as

$P(b_1, \ldots, b_T)$, where $P \in \mathbb{R}[X_1, \ldots, X_T]$ is a multivariate polynomial in T variables. It is then easy to see that any pointwise h will change the determinant of M as: $\det(h(\mathbf{M})) = P(h(b_1), \ldots, h(b_T))$. Then, assuming this lemma is not true is equivalent with $P(c_1, \ldots, c_T) = 0, \forall c_i \in [b_i - \epsilon, b_i + \epsilon], \forall 1 \leq i \leq T$. Applying lemma 8 to sets $S_i = [b_i - \epsilon, b_i + \epsilon]$, one gets that $P(c_1, \ldots, c_T) = 0, \forall c_i \in \mathbb{R}, \forall i \in \{1, \ldots, T\}$. Taking $c_i = f(b_i)$ one obtains $\det(f(\mathbf{M})) = P(f(b_1), \ldots, f(b_T)) = 0$ which is a contradiction with our assumption on $\mathbf{M}$ and f. □

We now return to the proof of the main theorem. For each $i \in \{1, \ldots, T\}$, let us denote by $c_i \in [b_i - \epsilon, b_i + \epsilon]$ the values from lemma 9 that guarantee a non-zero determinant. We construct a pointwise bijective, piecewise differentiable, continuous and strictly increasing function $g : \mathbb{R} \to \mathbb{R}$ such that $g(b_i) = c_i$. It is obvious that $\det(g(\mathbf{M}))$ depends only on the values $g(b_i)$, so we are free to assign any other values to any other real input of g as long as the above constraints on g are satisfied. One example of such g is a piecewise linear function defined to match the following values: $g(b_i) = c_i, g(b_i + 2\epsilon) = b_i + 2\epsilon, \forall 1 \leq i \leq T, g(x) = x, \forall x < b_1 - 2\epsilon$ and $g(x) = x, \forall x > b_T + 2\epsilon$. It can be easily seen that such a function is bijective, piecewise differentiable, continuous and strictly increasing. □

D. Proof of Lemma 4

Proof. If $\langle \mathbf{w}_i, \mathbf{h}_{j_i} \rangle$ are distinct from all the other entries in the matrix A, one can design the following pointwise function:

$$f(x) = \begin{cases} 1 & \text{if } \exists i \text{ s.t. } x = \langle \mathbf{w}_i, \mathbf{h}_{j_i} \rangle \\ 0 & \text{else} \end{cases}$$

Then, let $\mathbf{B}$ be the $M \times M$ submatrix of A consisting of all its M rows and the M columns indexed by j_i's. It is then clear that $f(\mathbf{B}) = \mathbf{I}_M$, which is obviously full rank. □

E. Proof of Theorem 6

Proof. We will make use of the following folklore lemmas:

Lemma 10. *Let $\mathcal{M} = \cup_i M_i$ be a finite union of Riemannian manifolds of dimension m, embedded in $\mathbb{R}^k$, with Riemannian metric g_i inherited from $\mathbb{R}^k$. Then, any finite union S of submanifolds of the $\mathcal{M}_i$'s of dimensions strictly smaller than m is a set of null measure[6]. In other words, any point from $\mathcal{M}$ is almost surely not in S.*

Proof. (sketch) any submanifold of $\mathcal{M}$ of strictly smaller dimension than m has volume or measure zero. The result then follows from the fact that a finite union of sets of measure zero has also measure zero. □

Lemma 11. *The set O_k^N of rank-k matrices of size $N \times N$ with $0 < k < N$ is a Riemannian manifold of dimension $2kN - k^2$ embedded in $\mathbb{R}^{N \times N}$.*

Proof. See e.g. (Shalit et al., 2012). The Riemannian metric for embedded manifolds is simply the Euclidean metric restricted to the manifold. □

We now return to the main proof of the theorem. From lemma 11 we have that $\dim(O_{N-1}^N) = N^2 - 1$. We want to prove that the subset of O_{N-1}^N of rank $N-1$ matrices for which x^2 is not increasing their rank has dimension strictly smaller than $\dim(O_{N-1}^N)$. In this case, using lemma 10, the measure of all ill-behaved matrices would be 0, so any matrix from O_{N-1}^N is almost surely well-behaved, i.e. the rank of $\mathbf{A}^{\odot 2}$ is almost surely full rank N for $\mathbf{A} \in O_{N-1}^N$.

We begin by removing from O_{N-1}^N the set of all matrices that have two proportional columns, a set that we name Ξ^N. This is a finite[7] union of manifolds of dimension $N(N-1)+1$, namely all sets of matrices for which column i is proportional to column j, for all $1 \leq i < j \leq N$ [8]. Using lemma 10, we derive that the measure or volume of Ξ^N is 0.

Now, for any arbitrary $\mathbf{A} \in O_{N-1}^N \setminus \Xi^N$ with columns $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)} \in \mathbb{R}^N$, one can easily derive that $\exists \gamma_i \in \mathbb{R}$ not all equal to 0 s.t. $\sum_{i=1}^N \gamma_i \mathbf{x}^{(i)} = 0$. We know that at least one $\gamma_i \neq 0$ from the fact that $\mathbf{A} \in O_{N-1}^N$; let us denote by Γ^i the set of such matrices $\mathbf{A} \in O_{N-1}^N$. Since O_{N-1}^N is the (finite) union of the Γ^i's, we want to show that the set of ill-behaved matrices in each Γ^i is contained in a manifold of dimension strictly smaller than that of O_{N-1}^N, which will conclude, using the fact that a finite union of null measure sets has null measure.

Without loss of generality, let us assume that $\mathbf{A} \in \Gamma^N$, *i.e.* that $\gamma_N \neq 0$. Let us note that

$$\Gamma^N = \{\mathbf{A} \in O_{N-1}^N : \gamma_N = 1\}, \tag{24}$$

by substituting each γ_i with γ_i / γ_N for $1 \leq i \leq N-1$.

[6] w.r.t. the volume form of the manifold, *i.e.* locally w.r.t. to the m-dimensional Lebesgue measure.

[7] More precisely, of $\frac{N(N-1)}{2}$ manifolds, one per each pair of columns.

[8] The $N(N-1)+1$ dimension comes from the fact that there are N-1 independent columns, plus a scalar, namely the multiplication factor between column i and column j.

If $\mathbf{A}^{\odot 2}$ is not full rank, there exist $\alpha_1, ..., \alpha_N \in \mathbb{R}$ such that

$$\sum_{i=1}^{N-1} \alpha_i (\mathbf{x}^{(i)})^{\odot 2} = \alpha_N \left(\sum_{i=1}^{N-1} \gamma_i \mathbf{x}^{(i)} \right)^{\odot 2} . \quad (25)$$

For fixed $\alpha_1, ..., \alpha_N \in \mathbb{R}$, denote by M_α the subset of the solutions $\{\mathbf{x}^{(i)}\}_{1 \le i \le N-1} \subset \mathbb{R}^N$ of the above equation.

Define

$$\varphi : (x_k^{(1)}, ..., x_k^{(N-1)}) \in \mathbb{R}^{N-1} \mapsto \sum_{i=1}^{N-1} \alpha_i (x_k^{(i)})^2 - \alpha_N \left(\sum_{i=1}^{N-1} \gamma_i x_k^{(i)} \right)^2 . \quad (26)$$

This can be re-written $\varphi(\mathbf{x}) = \mathbf{x}^T \mathbf{G} \mathbf{x}$ with

$$G_{ij} = \delta_{ij}(\alpha_i - \alpha_N \gamma_i^2) - (1 - \delta_{ij}) \alpha_N \gamma_i \gamma_j$$

It can be easily shown that since $\mathbf{A}$ is not in Ξ^N, $\mathbf{G}$ is not the null matrix. Indeed, if $\mathbf{G} = \mathbf{0}$, then either $\alpha_N = 0$ – and then $\alpha_i = \alpha_N \gamma_i^2 = 0$ for all i, which is excluded – or $\alpha_N \neq 0$, and then $\alpha_N \gamma_i \gamma_j = 0$ for all $i \neq j$, meaning only one γ_{i_0} is non-zero, *i.e.* $\mathbf{x}^{(N)} = -\gamma_{i_0} \mathbf{x}^{(i_0)}$ and hence $\mathbf{A} \in \Xi^N$.

Note that since $\mathbf{G}$ is not the null matrix, $dim(\ker \mathbf{G}) < N - 1$. Furthermore, let $U := \mathbb{R}^{N-1} \setminus \ker \mathbf{G}$. Invoking the Pre-Image theorem, the set $U \cap \varphi^{-1}(\{0\})$ is a submanifold of $\mathbb{R}^{N-1}$ of dimension $(N-1) - 1 = N - 2$. Therefore, $\varphi^{-1}(\{0\})$ is a finite union of manifolds of dimensions smaller than (or equal to) $N - 2$.

Since eq. (25) can be written as an intersection of N equations as the one defined by φ (*i.e.* one per coordinate), the set M_α of solutions of eq. (25) is included in a finite union of manifolds of dimensions smaller than (or equal to) $N(N-2)$.

Finally, the total set X of matrices we are after – *i.e.* of rank $N - 1$ and which cannot be made full ranked by pointwise square – can be defined as the union over α of all M_α, *i.e.* $X = \cup_\alpha M_\alpha$. As X has the structure of a fiber bundle, with base space the set of α's (of dimension N), X is a subset of submanifolds of dimensions smaller than $N + N(N-2) = N^2 - N < N^2 - 1$ for $N > 1$, which concludes the proof.

□

F. Proof of Theorem 7

Proof. Let $h : [-T, T]$ be any increasing function defined on $[-T, T]$. Assume bounded derivatives, i.e. $\exists R > 0$ s.t. $|h'(x)| < R, \forall x \in [-T, T]$. Then, for a fixed positive integer K, we consider the knots $l_i = -T + \frac{2Ti}{K}, \forall 0 \le i \le K$. Next, using standard linear interpolation, we define a piecewise linear function $f_K : [-T, T] \to \mathbb{R}$ s.t. $f_K(l_i) = h(l_i), \forall 0 \le i \le K$. Since h is increasing, one obtains that f_K is also increasing. It is then easy to see that f_K is a PLIF function. Moreover, the slopes are given by the formula: $s_i = \frac{h(l_{i+1}) - h(l_i)}{l_{i+1} - l_i}$.

We define the additional function $g_K(x) := f_K(x) - h(x)$. We wish to prove that $\lim_{K \to \infty} \max_{x \in [-T,T]} |g_K(x)| = 0$. For this, we first use Cauchy's theorem deriving that $\exists c_i \in (l_{i+1}, l_i)$ s.t. $s_i = \frac{h(l_{i+1}) - h(l_i)}{l_{i+1} - l_i} = h'(c_i)$. Thus, since h' is bounded by R, we get that $|s_i| < R, \forall i$. This further implies that $|g'_K(x)| < 2R, \forall x \in [-T, T]$. Moreover, from the definition of f_K we have that $g_K(l_i) = 0, \forall i$. Finally, for any $x \in [-T, T]$, let $[l_{i+1}, l_i]$ be the interval in which x lies. We have that:

$$\begin{aligned} |g_K(x)| &= |g_K(x) - g_K(l_i)| = \\ &= \frac{|g_K(x) - g_K(l_i)|}{|x - l_i|} |x - l_i| \le \\ &\le 2R|x - l_i| \le 2R \frac{2T}{K} \end{aligned} \quad (27)$$

where the first inequality happens from the same argument derived from Cauchy's theorem as above. It is now trivial to prove that $\lim_{K \to \infty} \max_{x \in [-T,T]} |g_K(x)| = 0$, which concludes our proof.

□

G. Effect of the Dirichlet concentration

See fig. 5.

H. Additional Synthetic Experiments

See figs. 6 to 8.

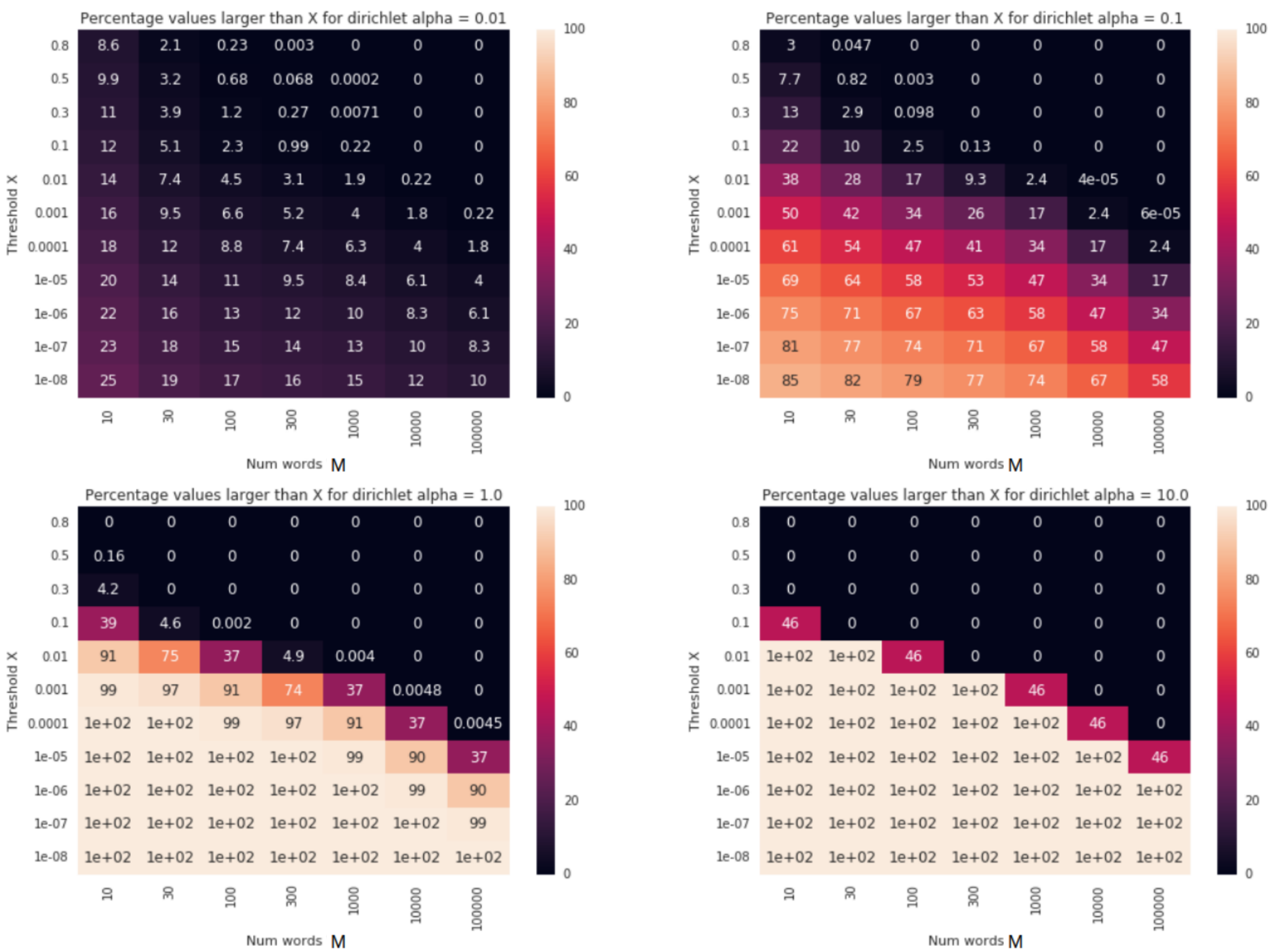

Figure 5. Distribution of M-class discrete distributions sampled from a Dirichlet prior. Larger concentration parameters result in close to uniform distributions, while low values result in sparse or long-tail distributions.

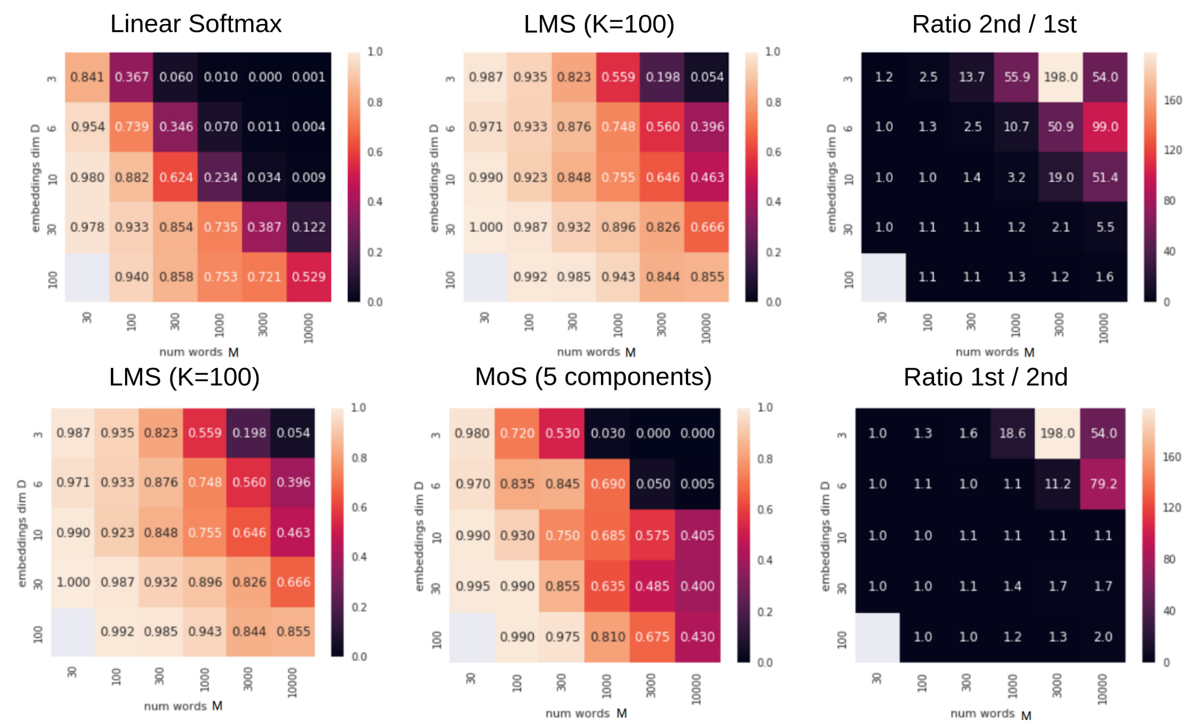

Figure 6. Percentage of contexts j for which the modes of true and parametric distributions match, i.e $\arg\max_i P^*(x_i|c_j) = \arg\max_i Q_\Theta(x_i|c_j)$. Higher the better. Dirichlet concentration $\alpha = 0.01$.

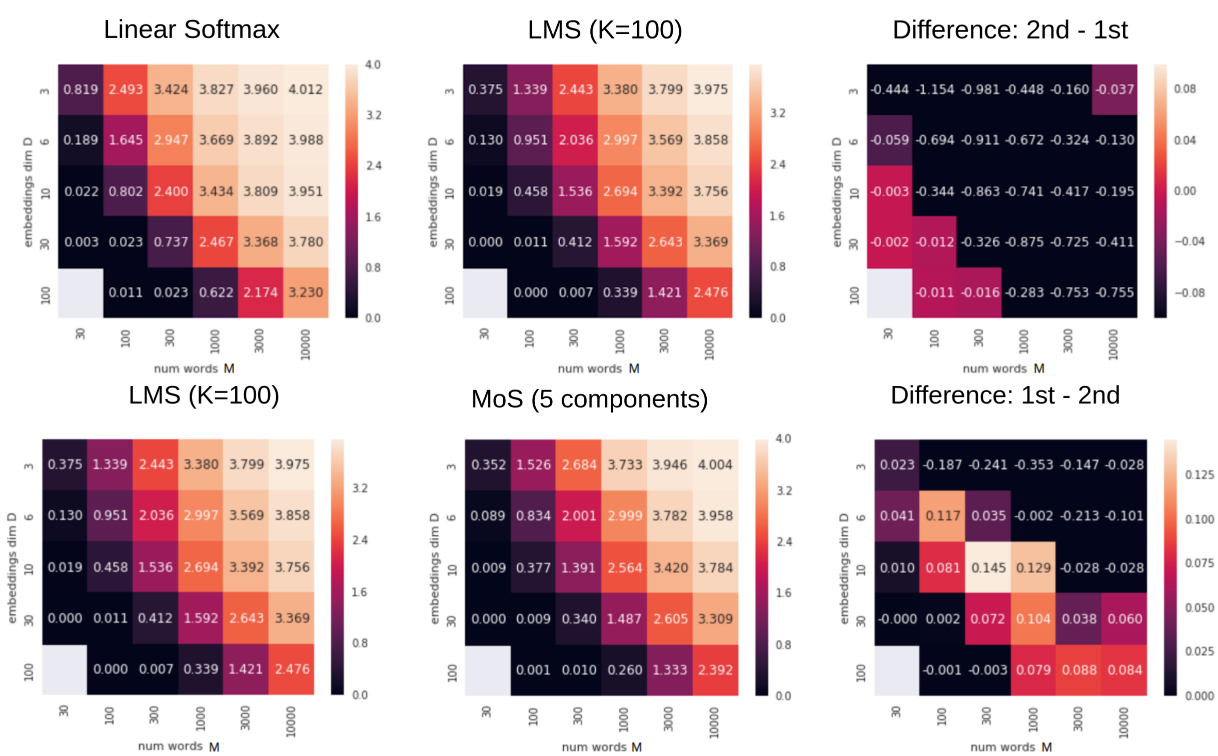

Figure 7. Average $KL(P^*||Q_\Theta)$ (across all contexts). Lower the better. Dirichlet concentration $\alpha = 0.01$.

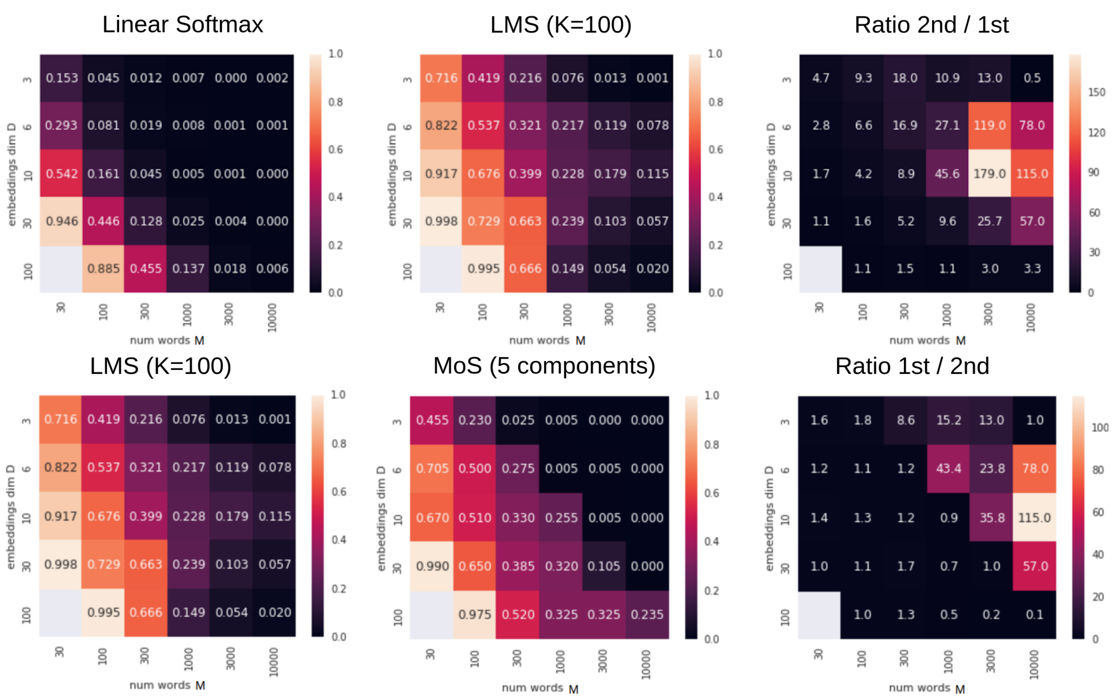

Figure 8. Percentage of contexts j for which the modes of true and parametric distributions match, i.e $\arg\max_i P^*(x_i|c_j) = \arg\max_i Q_\Theta(x_i|c_j)$. Higher the better. Dirichlet concentration $\alpha = 1$.