Fast Stochastic Quadrature for Approximate Maximum-Likelihood Estimation —Supplementary Material—

Nico Piatkowski
Department of Computer Science
AI Group, TU Dortmund
44221 Dortmund, Germany

Katharina Morik
Department of Computer Science
AI Group, TU Dortmund
44221 Dortmund, Germany

PROOFS

Proof of Lemma 1. Each of the above coordinate-wise statistics have closed-form integrals for any $c \in \mathbb{N}$. Now, in the course of integrating the product $\int_{\mathcal{X}} \prod_{i=1}^{k} \phi(\boldsymbol{x})_{\boldsymbol{j}_i} \,\mathrm{d}\,\nu(\boldsymbol{x})$ for some fixed $\boldsymbol{j} \in [d]^k$, we can encounter the following situations. First, products of the same dimension, say $\phi(\boldsymbol{x})_i$, yield an expression of the same functional form, but with a power of $2c$ instead of c. Hence, those products have a closed-form integral. The integral of the product of different coordinate-wise statistics $\phi(\boldsymbol{x})_i$ and $\phi(\boldsymbol{x})_j$, which access different components of $\boldsymbol{x}$, say $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$, is simply the product of the corresponding integrals, i.e., $\int\int \phi(\boldsymbol{x})_i \phi(\boldsymbol{x})_j \,\mathrm{d}\,\nu(\boldsymbol{x}_a) \,\mathrm{d}\,\nu(\boldsymbol{x}_b) = \left(\int \phi(\boldsymbol{x})_i \,\mathrm{d}\,\nu(\boldsymbol{x}_a)\right)\left(\int \phi(\boldsymbol{x})_j \,\mathrm{d}\,\nu(\boldsymbol{x}_b)\right)$, which also has a closed-form. The last case is the integration of the product of statistics of different type which involve the same variable $\boldsymbol{x}_a = x$. Any such integral can be written in the form

$$\int x^{c_1} \ln(x)^{c_2} \,\mathrm{d}\,x \;, \tag{1}$$

where $c_1 \in \mathbb{Z}$ and $c_2 \in \mathbb{N}$. Then, for $c_1 \neq -1$,

$$\int x^{c_1} \ln(x)^{c_2} \,\mathrm{d}\,x = \frac{x^{c_1+1} \ln(x)^{c_2}}{c_1+1} - \frac{c_2}{c_1+1} \int x^{c_1} \ln(x)^{c_2-1} \,\mathrm{d}\,x \;.$$

c_2 is a known integer constant and we may repeat the unrolling of the integral until only $\int x^{c_1} \,\mathrm{d}\,x$ is left. Thus, we arrive at a summation over a constant number (c_2) of terms. In the case $c_1 = -1$, the indefinite integral is $\ln(x)^{c_2+1}/(c_2+1)$. ■

Proof of Lemma 2. Considering the semantic of index tuples, the number of index tuples that correspond to the same clique tuple is equal to the number of joint state assignments to all cliques in the tuple. The number of such assignments is the product of the state spaces $\prod_{l=1}^{i} |\mathcal{X}_{\boldsymbol{C}(\boldsymbol{j})_l}|$, which establishes the statement for $|[\![\boldsymbol{j}]\!]|$. Second, the number of permutations of l objects is $l! = l(l-1)(l-2)\ldots 2$. Here, we are interested in the number of ways how $h(\boldsymbol{C}) \leq i$ objects—$h(\boldsymbol{C})$ distinct cliques—can be distributed to i different places. For example, when we consider the cliques A and B and the clique tuple (A, A, B), its equivalence class $[\![(A, A, B)]\!]$ contains the tuples (A, A, B), (A, B, A), (A, B, B), (B, B, A), (B, A, B), and (B, A, A). Counting such multicombinations is a well known combinatorial enumeration problem, equivalent to the total number of surjective functions from $[i]$ to $[h(\boldsymbol{C})]$. The resulting number is $h(\boldsymbol{C})!\{i \;\; h(\boldsymbol{C})\}^{\top}$, where the factorial accounts for the number of permutations of distinct cliques. The second factor is the Stirling number of second kind $\{n \;\; k\}^{\top}$, which counts the number of ways to partition a set of n elements into k subsets. This establishes the second equality. Lastly, each clique tuple in the equivalence class $[\![\boldsymbol{C}]\!]$ corresponds to the same number of indices, namely $|[\![\boldsymbol{j}]\!]|$. Hence, the total number of indices covered by the equivalence class $|[\![\boldsymbol{j}]\!]^*|$ is the product of the sizes $|[\![\boldsymbol{C}]\!]|$ and $|[\![\boldsymbol{j}]\!]|$, which completes the proof. ■

Proof of Theorem 2. By Hoeffding's inequality (Hoeffding, 1963), for N independent samples from a random variable $\boldsymbol{Y}$, bounded in $[a; b]$, we have

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y}^i - \mathbb{E}[\boldsymbol{Y}]\right| \geq t\right] \leq 2\exp\left(-\frac{2Nt^2}{(b-a)^2}\right) \;.$$

Setting $t = \varepsilon|\mathcal{X}|$ in combination with $|\hat{Z}^k_{\boldsymbol{\zeta}}(\boldsymbol{\theta}) - Z(\boldsymbol{\theta})| \leq \varepsilon|\mathcal{X}|$, Theorem 1 and the triangle inequality, we get

$$\begin{aligned}
&\mathbb{P}[|\hat{Z}^{N,k}_{\boldsymbol{\zeta}}(\boldsymbol{\theta}) - \hat{Z}^k_{\boldsymbol{\zeta}}(\boldsymbol{\theta})| \geq \varepsilon|\mathcal{X}|] \\
&\leq 2\exp\left(-N\frac{\varepsilon^2|\mathcal{X}|^2}{\tau^2 2\|\boldsymbol{\theta}\|^{2k'}_{\infty}}\right) = \delta \\
&\Rightarrow \mathbb{P}[|\hat{Z}^{N,k}_{\boldsymbol{\zeta}}(\boldsymbol{\theta}) - \hat{Z}^k_{\boldsymbol{\zeta}}(\boldsymbol{\theta})| + |\hat{Z}^k_{\boldsymbol{\zeta}}(\boldsymbol{\theta}) - Z(\boldsymbol{\theta})| \geq 2\varepsilon|\mathcal{X}|] \\
&\leq \delta \Rightarrow \mathbb{P}[|\hat{Z}^{N,k}_{\boldsymbol{\zeta}}(\boldsymbol{\theta}) - Z(\boldsymbol{\theta})| \geq 2\varepsilon|\mathcal{X}|] \leq \delta \;.
\end{aligned} \tag{2}$$

Now, we apply the error bound for Chebyshev approximations (Xiang et al., 2010) and Hölder's inequality (Hölder, 1889).

$$\varepsilon 2|\mathcal{X}| \leq \frac{4 \exp \|\boldsymbol{\theta}\|_1}{\pi\,(k-1)\,k!} 2|\mathcal{X}| \leq \epsilon \frac{|\mathcal{X}|}{\exp \|\boldsymbol{\theta}\|_1} < \epsilon Z(\boldsymbol{\theta}) \quad (3)$$

The rightmost inequality is known as naive mean field lower bound (Wainwright and Jordan, 2008). More precisely,

$$\begin{aligned} \log Z(\boldsymbol{\theta}) &= \sup_{\boldsymbol{\mu} \in \mathcal{M}} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle + \mathcal{H}(\boldsymbol{\mu}) \\ &\geq \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle + \mathcal{H}(\boldsymbol{\mu}) \quad \forall \boldsymbol{\mu} \in \mathcal{M}(G) \,, \end{aligned}$$

where $\mathcal{M}(G)$ is the marginal polytope and $\mathcal{H}(\boldsymbol{\mu})$ is the entropy of the density implied by $\boldsymbol{\mu}$. Since the inequality is valid for all $\boldsymbol{\mu} \in \mathcal{M}(G)$, we may choose the fully factorized density with uniform marginals. Hence,

$$\begin{aligned} \mathcal{H}(\boldsymbol{\mu}) &= -\sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) \\ &= \sum_{\boldsymbol{x} \in \mathcal{X}} \frac{1}{\prod_{v \in V} |\mathcal{X}_v|} \log \prod_{v \in V} |\mathcal{X}_v| = \log |\mathcal{X}| \,. \end{aligned}$$

Combining this with $\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle > -\|\boldsymbol{\theta}\|_1$ implies $\log Z(\boldsymbol{\theta}) > -\|\boldsymbol{\theta}\|_1 + \log |\mathcal{X}|$ and thus $Z(\boldsymbol{\theta}) > |\mathcal{X}| / \exp \|\boldsymbol{\theta}\|_1$, which explains the last inequality in (3). The statement of the theorem is then derived by plugging (3) into the probability that is complementary to (2). ■

Proof of Theorem 3. We call an index tuple realizable, if no states of the induced state tuple contradict each other[1]. The χ^i_ϕ value of non-realizable index tuples is 0, hence, it suffices to do the summation only over realizable tuples. The value $\chi^i_\phi(\boldsymbol{j})$ of any realizable index tuple depends only on the associated clique tuple $\boldsymbol{C}(\boldsymbol{j})$. This implies that all tuples from the same equivalence class $[\![\boldsymbol{j}]\!]_\phi$ will contribute equally to the sum. We can thus rewrite the summation over all index tuples $[d]^i$ in terms of a summation over all index tuples $\mathcal{C}^i$ and multiply each summand by the size of the corresponding equivalence class:

$$\begin{aligned} \|\chi^i_\phi\|_1 &= \sum_{\boldsymbol{j} \in [d]^i} |\chi^i_\phi(\boldsymbol{j})| \\ &= \sum_{\boldsymbol{C} \in \mathcal{C}^i} |[\![\boldsymbol{j}(\boldsymbol{C})]\!]_\phi| \chi^i_\phi(\boldsymbol{j}(\boldsymbol{C})) \,, \end{aligned}$$

where $\boldsymbol{j}(\boldsymbol{C})$ is an arbitrary index tuple associated with clique tuple $\boldsymbol{C}$.

Furthermore, note that χ^i_ϕ is permutation invariant, e.g., for any fixed indices $a, b, c, \ldots$ and any permutation σ, we have $\chi^i_\phi(a, b, c, \ldots) = \chi^i_\phi(\sigma(a, b, c, \ldots))$. In other words, χ^i will yield the same function value for all $\boldsymbol{j}$ which correspond to clique tuples $\boldsymbol{C}$ that are in the same equivalence class $[\![\boldsymbol{C}]\!]$. We hence sum only over the elements of $\mathcal{P}(\mathcal{C}, i)$ and multiply each term in the sum by the size of the corresponding equivalence class:

$$\begin{aligned} \|\chi^i_\phi\|_1 &= \sum_{\boldsymbol{C} \in \mathcal{C}^i} |[\![\boldsymbol{j}(\boldsymbol{C})]\!]_\phi| \chi^i_\phi(\boldsymbol{j}(\boldsymbol{C})) \\ &= \sum_{[\![\boldsymbol{C}]\!] \in \mathcal{P}(\mathcal{C}, i)} |[\![\boldsymbol{C}]\!]| |[\![\boldsymbol{j}(\boldsymbol{C})]\!]_\phi| \chi^i_\phi([\![\boldsymbol{j}(\boldsymbol{C})]\!]_\phi) \,. \end{aligned}$$

We have $\chi^i_\phi([\![\boldsymbol{j}(\boldsymbol{C})]\!]_\phi) = |\mathcal{X}| / |\mathcal{X}_{[\![\boldsymbol{C}]\!]}| = |\mathcal{X}| / |[\![\boldsymbol{j}(\boldsymbol{C})]\!]_\phi|$. Plugging this into the equation above and invoking Lemmas 2 and 3 to compute the sizes of equivalence classes, yields

$$\|\chi^i_\phi\|_1 = |\mathcal{X}| \sum_{[\![\boldsymbol{C}]\!] \in \mathcal{P}(\mathcal{C}, i)} h(\boldsymbol{C})! \left\{ {i \atop h(\boldsymbol{C})} \right\} \,.$$

Here, $\boldsymbol{C}$ is an arbitrary member of the equivalence class $[\![\boldsymbol{C}]\!]$, and $h(\boldsymbol{C})$ is the number of distinct cliques which appear in the tuple $\boldsymbol{C}$. We finally partition the summation over $\mathcal{P}(\mathcal{C}, i)$, into i separate summations, each over those members of $\mathcal{P}(\mathcal{C}, i)$ which have size $1 \leq l \leq i$. Hence, $h(\boldsymbol{C}) = l$ in each of these separate summations:

$$\|\chi^i_\phi\|_1 = |\mathcal{X}| \sum_{l=0}^{i} \sum_{[\![\boldsymbol{C}]\!] \in \mathcal{P}(\mathcal{C}, i) : h(\boldsymbol{C}) = l} \left\{ {i \atop l} \right\} l! \,.$$

(7) follows from the fact that number of terms in each inner sum can be computed via binomial coefficients. Note that $\{0\ 0\}^\top = 1$, $0! = 1$, and that the empty set is contained in $\mathcal{P}(\mathcal{C}, 0)$. The runtime reported in the theorem follows from the identity $\sum_{l=0}^{i} \left\{ {i \atop l} \right\} \binom{|\mathcal{C}|}{l} l! = |\mathcal{C}|^i$. ■

Proof of Lemma 3. We have $\chi^i_\phi(\boldsymbol{j}) = 0$ whenever an index tuple is non-realizable. I.e., at least two indices in the tuple imply different assignments to the same variable. For the clique tuple $\boldsymbol{C}(\boldsymbol{j})$, observe that $\mathcal{X}_{\boldsymbol{C}(\boldsymbol{j})}$ is the full joint state space of all unique variables in $\boldsymbol{C}$. If $\boldsymbol{C}$ contains cliques that share some vertices, those vertices must have the same state if and only if $\boldsymbol{j}$ is realizable. Thus, there cannot be more than $|\mathcal{X}_{\boldsymbol{C}(\boldsymbol{j})}|$ distinct realizable index tuples. The second equality is then a direct consequence of Lemma 2. ■

Proof of Theorem 4. If $\boldsymbol{j}$ is non-realizable, then $\boldsymbol{y} \notin \mathcal{X}_{[\![\boldsymbol{C}]\!]}$ and both $\mathbb{P}_{\boldsymbol{\zeta}, \phi}(\boldsymbol{J} = \boldsymbol{j} \mid I = i) = 0$ and the proposed factorization will assign mass 0. Now, assume that $\boldsymbol{j}$ is realizable and hence $\boldsymbol{y} \in \mathcal{X}_{[\![\boldsymbol{C}]\!]}$. By definition of the

[1] We refer to the proof of Lemma 2 in (Piatkowski and Morik, 2016) for more details on non-realizable instances.

above quantities:

$$\begin{aligned}&\mathbb{P}(\boldsymbol{C} \mid [\![\boldsymbol{C}]\!], l, i)\mathbb{P}(\boldsymbol{y} \mid [\![\boldsymbol{C}]\!], i)p([\![\boldsymbol{C}]\!] \mid l)\mathbb{P}(l \mid i)\\ =&\frac{1}{|\mathcal{X}_{[\![\boldsymbol{C}]\!]}| \sum_{h=0}^{i} \left\{ {i \atop h} \right\} \binom{|\mathcal{C}|}{h} h!} \,. \end{aligned} \tag{4}$$

We have $\chi_\phi^i(\boldsymbol{j}) = |\mathcal{X}|/|\mathcal{X}_{\bigcup_{i=1}^{k} C_{j_i}}|$ for any realizable $\boldsymbol{j}$. The denominator will be the same for all clique tuples $\boldsymbol{C}$ from the same equivalence class $[\![\boldsymbol{C}]\!]$, since those share the same variables. Hence, $\chi_\phi^i(\boldsymbol{j}) = |\mathcal{X}|/|\mathcal{X}_{[\![\boldsymbol{C}]\!]}|$. Multiplication of (4) by $1 = |\mathcal{X}|/|\mathcal{X}|$ hence yields

$$\begin{aligned}&\mathbb{P}(\boldsymbol{C} \mid [\![\boldsymbol{C}]\!], l, i)\mathbb{P}(\boldsymbol{y} \mid [\![\boldsymbol{C}]\!], i)\mathbb{P}([\![\boldsymbol{C}]\!] \mid l)\mathbb{P}(l \mid i)\\ =&\frac{\chi_\phi^i(\boldsymbol{j})}{|\mathcal{X}| \sum_{h=0}^{i} \left\{ {i \atop h} \right\} \binom{|\mathcal{C}|}{h} h!} \,. \end{aligned} \tag{5}$$

By Definition, we have $\mathbb{P}_{\boldsymbol{\zeta},\phi}(\boldsymbol{J} = \boldsymbol{j} \mid I = i) = \chi_\phi^i(\boldsymbol{j})/\|\chi_\phi^i\|_1$ for any realizable $\boldsymbol{j}$. Thus, the theorem follows by plugging the result of Theorem 3 into (5). ■

Proof of Theorem 5. We analyze the worst-case complexity of each step separately and assume that uniform random numbers can be drawn in $\mathcal{O}(1)$. To draw a sample according to $p(l \mid i)$, we employ inversion sampling. Computing each probability requires the evaluation of a Stirling number of second kind ($\mathcal{O}(l^2)$ steps), a binomial coefficient ($\mathcal{O}(l)$ steps), and a factorial (also $\mathcal{O}(l)$). These probabilities have to be computed for each $1 \leq l \leq i$. Since i is at most k, the total runtime for the probability computation is $\mathcal{O}(k^3)$. Drawing an actual inversion sample from $\{1, 2, \ldots, i\}$ hence requires $\mathcal{O}(k^4)$ steps. Whenever multiple samples must be drawn, we can reuse the probabilities. Thus, any subsequent sample requires only $\mathcal{O}(k)$ steps. Moreover, the terms that appear in the uniform sampling steps in lines 2 and 3 have already be computed for the first step. Consequently, their worst-case complexity is $\mathcal{O}(1)$.

We explained earlier that each clique equivalence class corresponds to a subset of l cliques. The algorithm presented in (Buckles and Lybanon, 1977) is used to directly generate the a-th clique combination (in lexicographic order). The runtime of this algorithm is equal to the largest element in the combination, and the worst-case complexity of line 4 is hence $\mathcal{O}(|\mathcal{C}|)$.

In contrast, we are not aware of an algorithm that generates the b-th composition of $\{1, 2, \ldots, i\}$ with l subsets directly. Instead, we generate such compositions by first computing an unordered partition of i elements into exactly l blocks, followed by a particular l-permutation.

As an example, let $[\![\boldsymbol{C}]\!] = \{A, B, C\}$ and $i = 5$ (the tuple length). Hence, $l = 3$ cliques must be assigned to $i = 5$ places. After determining the tuple, say $(1, 2, 2, 3, 1)$, a permutation of (A, B, C) must be determined to find the actual clique tuple. in case of the identity permutation, the resulting tuple would be (A, B, B, C, A). Another permutation would lead to (C, A, A, B, C)—there are $3! = 6$ of such permutation in total.

We employ the algorithm from (Ehrlich, 1973) (Section 5.2.2) to generate all unordered partitions of i elements into exactly l blocks. There are $\{i \; l\}^\top$ such partitions and the algorithm from (Ehrlich, 1973) requires $\mathcal{O}(1)$ steps to generate the successor of any partition. Another algorithm from (Ehrlich, 1973) can be used to generate all l-permutations in $\mathcal{O}(l!)$ steps. Alternatively, there are algorithms which do not require to store all permutations but generate the q-th permutation (for any q) directly—such procedures require $\mathcal{O}(l^2)$ steps. In total, it requires $\mathcal{O}(\{i \; l\}^\top + l!)$ steps (and memory) to precompute all partitions and permutations—any subsequent execution of line 5 (for the same combination of i and l) will take $\mathcal{O}(1)$ steps. Since clique tuples $\boldsymbol{C}$ may involve all $|V| = n$ vertices, the worst-case complexities of lines 6 and 7 are $\mathcal{O}(kn)$ and $\mathcal{O}(n)$, respectively. The joint state $\boldsymbol{y}$, computed in line 8, is the c-th element of the product space $\bigotimes_{v \in S} \mathcal{X}_v$. Converting c into the particular state $\boldsymbol{y}$ is done by a series of $|S|$ subtractions and divisions. As explained above, S may contain all vertices, and line 8 can thus be computed in $\mathcal{O}(n)$ steps. By employing an array of offsets to access the first parameter index of any clique, the conversion of the pair $(\boldsymbol{C}, \boldsymbol{y})$ to the index tuple $\boldsymbol{j}$ is done in $\mathcal{O}(k)$ steps. Combining these insights yields the statement of the theorem. ■

Proof of Theorem 6 Let $l = \min\{\hat{Z}_{\boldsymbol{\zeta}}^{N,k}(\boldsymbol{\theta}), Z(\boldsymbol{\theta})\}$ and $u = \max\{\hat{Z}_{\boldsymbol{\zeta}}^{N,k}(\boldsymbol{\theta}), Z(\boldsymbol{\theta})\}$. Applying the mean value theorem to the logarithm, there is $\xi \in [l, u]$ such that $\xi |\ln l - \ln u| = |l - u|$. By Theorem 2,

$$\mathbb{P}[|\hat{Z}_{\boldsymbol{\zeta}}^{N,k}(\boldsymbol{\theta}) - Z(\boldsymbol{\theta})| < \epsilon Z(\boldsymbol{\theta})] \geq 1 - \delta \,,$$

and hence

$$\mathbb{P}[\xi |\ln \hat{Z}_{\boldsymbol{\zeta}}^{N,k}(\boldsymbol{\theta}) - \ln Z(\boldsymbol{\theta})| < \epsilon Z(\boldsymbol{\theta})] \geq 1 - \delta \,.$$

Dividing by ξ and using the fact $\xi \geq l$ implies the desired result. ■

References

Bill P. Buckles and M. Lybanon. Algorithm 515: Generation of a vector from the lexicographical index [g6]. *ACM Transactions on Mathematical Software*, 3(2):180–182, June 1977. ISSN 0098-3500. doi: 10.1145/355732.355739.

Gideon Ehrlich. Loopless algorithms for generating permutations, combinations, and other combinatorial

configurations. *Journal of the ACM*, 20(3):500–513, 1973. doi: 10.1145/321765.321781.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Otto Ludwig Hölder. Ueber einen Mittelwerthssatz. *Nachrichten von der Königlichen Gesellschaft der Wissenschaften und der Georg-August-Universität Göttingen*, 2:38–47, 1889.

Nico Piatkowski and Katharina Morik. Stochastic discrete clenshaw-curtis quadrature. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 3000–3009. JMLR.org, 2016.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1 (1–2):1–305, 2008.

Shuhuang Xiang, Xiaojun Chen, and Haiyong Wang. Error bounds for approximation in Chebyshev points. *Numerische Mathematik*, 116(3):463–491, 2010.