A. Proof of Proposition 3.1

Let $\mathbf{X}$ be partially exchangeable with respect to the statistic T with values $\mathcal{T}$, let $|\mathcal{T}| = \mathbf{poly}(|\mathbf{X}|)$, and let, for any partial assignment $\mathbf{e}$, $S_{t,\mathbf{e}} := \{\mathbf{x} \mid T(\mathbf{x}) = t \text{ and } \mathbf{x} \sim \mathbf{e}\}$, where $\mathbf{x} \sim \mathbf{e}$ denotes that $\mathbf{x}$ and $\mathbf{e}$ agree on the variables in their intersection (Koller & Friedman, 2009). If we can in time $\mathbf{poly}(|\mathbf{X}|)$,

(1) for every $\mathbf{e}$ and every $t \in \mathcal{T}$, decide if there exists an $\mathbf{x} \in S_{t,\mathbf{e}}$ and, if so, construct such an $\mathbf{x}$,

then the complexity of MAP inference, that is, computing $\operatorname{argmax}_{\mathbf{y}} P(\mathbf{y}, \mathbf{e})$ for any partial assignment $\mathbf{e}$, is $\mathbf{poly}(|\mathbf{X}|)$. If, in addition, we can in time $\mathbf{poly}(|\mathbf{X}|)$,

(2) for every $\mathbf{e}$ and every $t \in \mathcal{T}$, compute $|S_{t,\mathbf{e}}|$,

then the complexity of marginal inference, that is, computing $P(\mathbf{e})$ for any partial assignment $\mathbf{e}$, is $\mathbf{poly}(|\mathbf{X}|)$.

Proof. We first prove statement (1). Let $\mathbf{e}$ be a given partial assignment and assume we want to compute $\operatorname{argmax}_{\mathbf{y}} P(\mathbf{y}, \mathbf{e})$. We construct an $\mathbf{x}_t \in S_{t,\mathbf{e}}$ for each $t \in \mathcal{T}$ and set $\hat{\mathbf{x}}_t := \operatorname{argmax}_{\mathbf{x}_t} P(\mathbf{x}_t)$. By assumption, this is possible in time $\mathbf{poly}(|\mathbf{X}|)$. Since we have that $\hat{\mathbf{x}}_t = \hat{\mathbf{y}}\mathbf{e}$ with $\hat{\mathbf{y}} := \operatorname{argmax}_{\mathbf{y}} P(\mathbf{y}, \mathbf{e})$ we can extract the solution in linear time.

To prove statement (2), let $\mathbf{e}$ be a partial assignment. We construct a $\mathbf{x}_t \in S_{t,\mathbf{e}}$ for each $t \in \mathcal{T}$ for which such an $\mathbf{x}_t$ exists, compute $|S_{t,\mathbf{e}}|$, and return $\sum_{t \in \mathcal{T}} P(\mathbf{x}_t)|S_{t,\mathbf{e}}|$. By assumption, this is possible in time $\mathbf{poly}(|\mathbf{X}|)$. □

We can utilize Proposition 3.1 to prove that probabilistic inference for a sequence of n exchangeable binary variables is tractable.

Example A.1 (Finite Exchangeability)**.** Let $\mathbf{X}$ be an exchangeable sequence of binary random variables. Let $\mathtt{n}(\mathbf{e})$ be the number of 1s in a partial assignment $\mathbf{e}$ to the variables $\mathbf{X}$. Clearly, we have that $\mathbf{X}$ is exchangeable with respect to the statistic $T(\mathbf{x}) = \mathtt{n}(\mathbf{x})$ with values $\mathcal{T} = \{0, ..., n\}$.

First, we prove that for every partial assignment $\mathbf{e}$ to k of the n variables and every $t \in \mathcal{T}$, we can decide if there exists an $\mathbf{x} \in S_{t,\mathbf{e}}$ and, if so, construct such an $\mathbf{x}$ in time $\mathbf{poly}(|\mathbf{X}|)$. If $\mathtt{n}(\mathbf{e}) > t$ or $n - k + \mathtt{n}(\mathbf{e}) < t$, then there does *not* exist such an $\mathbf{x}$. Otherwise it is possible to generate a $\mathbf{x}$ with $\mathtt{n}(\mathbf{x}) = t$ in linear time by assigning exactly $t - \mathtt{n}(\mathbf{e})$ ones to the unassigned variables and we have that $\mathbf{x} \in S_{t,\mathbf{e}}$. Hence, MAP inference is tractable.

Next, we prove that for every partial assignment $\mathbf{e}$ to k variables and every $t \in \mathcal{T}$, we can compute $|S_{t,\mathbf{e}}|$ in time $\mathbf{poly}(|\mathbf{X}|)$. But this is possible since $|S_{t,\mathbf{e}}| = \binom{n-k}{t-\mathtt{n}(\mathbf{e})}$. Hence, marginal inference is tractable.

Please note that Example A.1 implies tractability results for numerous important special cases of finite exchangeability such as parity and threshold functions.

There are forms of finite partial exchangeability (Diaconis & Freedman, 1980a) that go beyond the notion of *full* finite exchengeability and, therefore, cardinality-based potentials (Gupta et al., 2007; Tarlow et al., 2010) of Example A.1. We provide three examples.

Example A.2 (Block Exchangeability)**.** Let w be a fixed constant. For a sequence of binary random variables $\mathbf{X}$ let $\mathcal{X} = \{\mathbf{X}_1, ..., \mathbf{X}_w\}$ be a partition of the variables $\mathbf{X}$ into w subsequences, and let $\mathtt{n}_{\mathbf{Y}}(\mathbf{x})$ be the number of 1s in an assignment $\mathbf{x}$ projected onto the variables $\mathbf{Y} \subseteq \mathbf{X}$. Now, let $T(\mathbf{x}) = (\mathtt{n}_{\mathbf{X}_1}(\mathbf{x}), ..., \mathtt{n}_{\mathbf{X}_w}(\mathbf{x}))$.

It is straight-forward to verify that $|\mathcal{T}| = \mathbf{poly}(|\mathbf{X}|)$. Moreover, with arguments similar to those made in Example A.1 one can show that conditions (1) and (2) of Proposition 3.1 are met. Hence, MAP and marginal inference are tractable for the statistic T.

Example A.3. Let $\mathbf{X}$ be a sequence of n binary random variables and let $\tau_{0\to1}(\mathbf{x})$ be the number of times 01 occurs as a substring[2] in $\mathbf{x}$. Now, consider the statistic

$$T(\mathbf{x}) = \tau_{0\to1}(\mathbf{x}).$$

For example, for $\mathbf{x} = 11011111$ we have $T(\mathbf{x}) = 1$ and for $\mathbf{x} = 01010101$ we have $T(\mathbf{x}) = 4$. We also have that $|\mathcal{T}| = \lfloor n/2 \rfloor + 1 = \mathbf{poly}(|\mathbf{X}|)$.

Now, let $\mathbf{e}$ be a partial assignment to k of the n variables and let $0 \leq t \leq \lfloor n/2 \rfloor$ be a value of the statistic. Let $\mathbf{b} = \{0, 1, *\}^n$ be a string where the characters 0 and 1 encode the assignments to variables according to $\mathbf{e}$ and the character * encodes unassigned variables. We now partition $\mathbf{b}$ into four sets G_{ij}, $i, j \in \{0, 1\}$, of substrings defined as $G_{ij} := \{\mathbf{s} \sqsubseteq \mathbf{b} \mid s_1 = i, s_{|\mathbf{s}|} = j, s_\ell = * \text{ for } 1 \leq i < \ell < j \leq |\mathbf{s}|\}$, where $\sqsubseteq$ denotes the substring relation. We can now complete the partial assignment $\mathbf{e}$ to a joint assignment $\mathbf{x}$ with $T(\mathbf{x}) = t$ if and only if (1) $\tau_{0\to1}(\mathbf{b}) + |G_{01}| \leq t$ and (2) $\tau_{0\to1}(\mathbf{b}) + \sum_{\mathbf{s}\in G_{00}} \left\lceil \frac{|\mathbf{s}|-2}{2} \right\rceil + \sum_{\mathbf{s}\in G_{01}} \left\lfloor \frac{|\mathbf{s}|}{2} \right\rfloor + \sum_{\mathbf{s}\in G_{10}} \left\lfloor \frac{|\mathbf{s}|-2}{2} \right\rfloor + \sum_{\mathbf{s}\in G_{11}} \left\lceil \frac{|\mathbf{s}|-2}{2} \right\rceil \geq t$. When these two conditions are met, the full assignment $\mathbf{x}$ can be constructed by completing the substring in the groups G_{ij} so as to make $T(\mathbf{x}) = t$ and this is possible in linear time. Hence, MAP inference is tractable.

It is possible to construct novel tractable statistics by nesting statistics that are known to be tractable.

[2] As opposed to subsequences, substrings are consecutive parts of a string.

Example A.4 (Nested Tractable Statistics)**.** Let $\mathbf{X}$ be an $n \times n$ array of binary random variables. For instance, $\mathbf{X}$ could represent a binarized image with n rows and n columns. Let k be a fixed integer constant and let ℓ be the integer such that $n = k\ell$. We assume without loss of generality that such an integer exists. We partition the original array into ℓ^2 squares of dimension $k \times k$. For $1 \leq i \leq \ell^2$, let $\mathbf{S}_i$ be the variables of square i. Now, let $T_1 : \{0,1\}^{k^2} \rightarrow \{0,1\}$ be the statistic defined as

$$T_1(\mathbf{s} = (s_1, ..., s_{k^2})) = [[\sum_{i=1}^{k^2} s_i > \tau]],$$

for some τ with $0 \leq \tau < k^2$. That is, $T_1(\mathbf{s}) = 1$, if the number of 1s in a given square exceeds a threshold of τ and $T_1(\mathbf{s}) = 0$ otherwise. Please note that for $\tau = 0$ this corresponds to max-pooling. Now, let $T : \{0,1\}^{n^2} \rightarrow \{0, ..., \ell^2\}$ be the statistic defined as follows:

$$T(\mathbf{x}) = \sum_{i=1}^{\ell^2} T_1(\mathbf{s}_i).$$

Based on the tractability of the two statistics, it is straightforward to verify that both MAP and marginal inference is tractable for the statistic T.

Please note that the presented theoretical framework facilitates the discovery and development of novel tractable non-local potentials.

B. Proof of Theorem 3.2

Let $X_1, ..., X_n$ be a sequence of random variables with joint distribution P, let T be a statistic with distinct values $t_0, ..., t_k$, and let $X_1, ..., X_n$ be partially exchangeable with respect to T. The ML estimates for N examples, $\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}$, are $\texttt{MLE}[(w_0, ..., w_k)] = \left(\frac{c_0}{N}, ..., \frac{c_k}{N}\right)$, where $c_i = \sum_{j=1}^{N} [[T\left(\mathbf{x}^{(j)}\right) = t_i]]$.

Proof. Let $\theta = (w_0, ..., w_k)$. By Theorem 2.3, the log-likelihood for N examples $\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}$ is

$$\texttt{L}(\theta) = \sum_{j=1}^{N} \log \left(\sum_{i=0}^{k} w_i U_i \left(\mathbf{x}^{(j)}\right) \right).$$

Let $c_i = \sum_{j=1}^{N} [[T\left(\mathbf{x}^{(j)}\right) = t_i]]$ and let $\hat{\mathbf{x}}_i$ be a joint assignment with $T(\hat{\mathbf{x}}_i) = t_i$. Then, $\texttt{L}(\theta) = \sum_{i=0}^{k} c_i \log(w_i U_i(\hat{\mathbf{x}}_i)) = \sum_{i=0}^{k} c_i [\log(w_i) + \log(U_i(\hat{\mathbf{x}}_i))] = \sum_{i=0}^{k} c_i \log(w_i) + \sum_{i=0}^{k} c_i \log(U_i(\hat{\mathbf{x}}_i))$. The second term is free of parameters and, hence, finding the ML estimates amounts to maximizing the first sum. This is equivalent to finding the maximum likelihood estimate of a multinomial which can be solved with Lagrange multipliers. Hence, $\texttt{MLE}(w_i) = \frac{c_i}{N}$, for $0 \leq i \leq k$. □

C. Proof of Proposition 3.3

The following statements are necessary conditions for exchangeability of a finite sequence of random variables $X_1, ..., X_n$. For all $i, j, i', j' \in \{1, ..., n\}$ with $i \neq j$ and $i' \neq j'$

(1) $\mathbf{E}(X_i) = \mathbf{E}(X_j)$;

(2) $\mathbf{Var}(X_i) = \mathbf{Var}(X_j)$; and

(3) $\mathbf{Cov}(X_i, X_j) = \mathbf{Cov}(X_{i'}, X_{j'}) \geq -\frac{\mathbf{Var}(X_i)}{(n-1)}$.

These conditions are well-known and are straight-forward to prove. Nevertheless, for the sake of completeness, we prove statement (3).

Proof. It is straight-forward to prove statements (1) and (2). In order to prove statement (3) we use statements (2) to write

$$\begin{aligned} 0 &\leq \mathbf{Var}(X_1 + \cdots + X_n) \\ &= \mathbf{Var}(X_1) + \cdots + \mathbf{Var}(X_n) + 2 \sum_{i<j} \mathbf{Cov}(X_i, X_j) \\ &= n\mathbf{Var}(X_i) + n(n-1)\mathbf{Cov}(X_i, X_j). \end{aligned}$$

Hence, $\mathbf{Cov}(X_i, X_j) \geq -\frac{\mathbf{Var}(X_i)}{(n-1)}$. □

D. Proof of Theorem 4.4

The mixtures of EVMs family is globally optimal under zero-one loss for

1. Conjunctions and disjunctions of attributes;
2. Symmetric Boolean functions such as
   - Threshold (m-of-n) functions
   - Parity functions
   - Counting functions
   - Exact value functions

Proof. Let $\mathbf{X}$ be the sequence of variables under consideration. We write $y(\mathbf{x})$ for the (hidden) class value of example $\mathbf{x}$. For conjunctions of attributes, let $\hat{\mathbf{X}} \subseteq \mathbf{X}$ be the sequence of variables that are part of the conjunction. Conditioned on the binary class variable being either 0 or 1, we partition the variables into the two blocks $\hat{\mathbf{X}}$ and $\mathbf{X} - \hat{\mathbf{X}}$. We set the parameters of the MEVM as follows.

$q_{(\hat{\mathbf{X}})}(\ell \mid 1) = 1.0$ if $\ell = |\hat{\mathbf{X}}|$ and $q_{(\hat{\mathbf{X}})}(\ell \mid 1) = 0.0$ otherwise;

$q_{(\hat{\mathbf{X}})}(\ell \mid 0) = 0.0$ if $\ell = |\hat{\mathbf{X}}|$ and $q_{(\hat{\mathbf{X}})}(\ell \mid 0) = \frac{\binom{|\hat{\mathbf{X}}|}{\ell}}{2^{|\hat{\mathbf{X}}|}}$ otherwise;

$q_{(\mathbf{X}-\hat{\mathbf{X}})}(\ell \mid 1) = \frac{\binom{|\mathbf{X}|-|\hat{\mathbf{X}}|}{\ell}}{2^{|\mathbf{X}|-|\hat{\mathbf{X}}|}}$; $q_{(\mathbf{X}-\hat{\mathbf{X}})}(\ell \mid 0) = \frac{\binom{|\mathbf{X}|-|\hat{\mathbf{X}}|}{\ell}}{2^{|\mathbf{X}|-|\hat{\mathbf{X}}|}}$;

$p(1) = \frac{2^{|\mathbf{X}|-|\hat{\mathbf{X}}|}}{2^{|\mathbf{X}|}}$; and $p(0) = \frac{(2^{|\hat{\mathbf{X}}|}-1)(2^{|\mathbf{X}|-|\hat{\mathbf{X}}|})}{2^{|\mathbf{X}|}}$.

Then, we have that $\mathbf{P}(1 \mid \mathbf{x}) > 0$ if $y(\mathbf{x}) = 1$ and $\mathbf{P}(1 \mid \mathbf{x}) = 0$ otherwise. Moreover, $\mathbf{P}(0 \mid \mathbf{x}) = 0$ if $y(\mathbf{x}) = 1$ and $\mathbf{P}(0 \mid \mathbf{x}) > 0$ otherwise. Hence, the MEVM classifier always returns the correct class value. A similar argument can be made to prove the optimality for disjunctions of attributes.

To prove the second statement, we consider an MEVM model with a binary class variable and the following block structure. For each of the class variable's values y, $y \in \{0, 1\}$, we have that $\mathcal{X}_y = \{X_1, ..., X_n\}$. That is, conditioned on each class value, the attributes are assumed to be exchangeable (see Figure 3; right). It is straightforward to verify that this particular MEVM can learn *arbitrary* discrete distributions over any symmetric Boolean function. □