Supplementary material to the ICML 2014 paper
Putting MRFs on a Tensor Train

Alexander Novikov[1] NOVIKOV@BAYESGROUP.RU
Anton Rodomanov[1] ANTON.RODOMANOV@GMAIL.COM
Anton Osokin[1] OSOKIN@BAYESGROUP.RU
Dmitry Vetrov[1,2] VETROVD@YANDEX.RU
[1] Moscow State University, Moscow, Russia
[2] Higher School of Economics, Moscow, Russia

1. TT-representation for some types of potentials

In this section we derive explicit formulae for the TT-representations of two types of potentials: unary and pairwise Ising. The correctness of each formula immediately follows from the definition of the matrix product. Both the representations are of low TT-rank. One can use the derived formulae instead of TT-SVD during step 1 of the algorithm from section 5.1.

1.1. Unary potentials

The unary potentials $\mathbf{\Theta}_\ell(\boldsymbol{x}) = \sum_{i=1}^n f_i(x_i)$, where f_i $(i = 1, \ldots, n)$ are arbitrary univariate functions, admit a TT-representation $\mathbf{\Theta}_\ell(\boldsymbol{x}) = \prod_{i=1}^n G_i^\ell[x_i]$ with the following cores:

$$G_i^\ell[x_i] = \begin{bmatrix} 1 & 0 \\ f_i(x_i) & 1 \end{bmatrix}, \quad i = 2, \ldots, n-1,$$
$$G_1^\ell[x_1] = \begin{bmatrix} f_1(x_1) & 1 \end{bmatrix},$$
$$G_n^\ell[x_n] = \begin{bmatrix} 1 \\ f_n(x_n) \end{bmatrix}.$$

The maximal TT-rank equals 2.

1.2. Ising potentials

The Ising potential

$$\mathbf{\Theta}_\ell(x_i, x_j) = x_i x_j, \quad 1 \le i < j \le n,$$

can be represented in the TT-format as $\mathbf{\Theta}_\ell(x_i, x_j) = G_i^\ell[x_i] G_j^\ell[x_j]$, where the TT-cores are defined as follows:

$$G_i^\ell[x_i] = x_i, \quad G_j^\ell[x_j] = x_j,$$

i.e. each core is simply a number (a 1-by-1 matrix) for each value of x_i (or x_j). The TT-representation of the Ising potential is of maximal TT-rank equal to 1.

2. Proof of theorem 1

The main paper presents the algorithm that converts the energy tensor $\mathbf{E}$ into the TT-format (sec. 5.1). Theorem 1 states an upper bound on the maximal TT-rank of the resulting TT-representation.

Theorem 1. *If the order of each potential $\mathbf{\Theta}_\ell$, $\ell = 1, \ldots, m$ does not exceed p, then the algorithm in sec. 5.1 constructs a TT-representation for the energy $\mathbf{E}$ in such a way that its maximal TT-rank is polynomially bounded:*

$$\mathrm{r}(\mathbf{E}) \le d^{\frac{p}{2}} \cdot m,$$

where each variable x_i takes at most d possible values.

Proof. At first, let us estimate the rank of the TT-representation of the potential $\mathbf{\Theta}_\ell$ constructed after step 1 of the algorithm. According to Oseledets (2011, Th. 2.1) the TT-SVD algorithm finds a TT-decomposition with the ranks that are not higher than the TT-ranks of the corresponding unfolding matrices[1] Since the order of each potential does not exceed p, the i-th unfolding matrix A_i of $\mathbf{\Theta}_\ell$ is of dimensionality $\alpha \times \beta$ with $\alpha \le d^i$ and $\beta \le d^{p-i}$. Recall that the rank of an $\alpha \times \beta$ matrix cannot be greater than $\min\{\alpha, \beta\}$. Thus, inequality $\mathrm{r}(A_i) \le d^{\frac{p}{2}}$ holds and therefore after step 1 the following bound is true:

$$\mathrm{r}(\mathbf{\Theta}_\ell) \le d^{\frac{p}{2}}. \tag{1}$$

As mentioned in the algorithm description, during step 2 the TT-rank of the potential tensor is not increasing. Thus, after step 2 inequality (1) still holds for the tensor of each potential. During step 3 of the algorithm we compute the sum of the potentials, so the maximal TT-rank is growing additively. The fact that the total number of potentials is m completes the proof. □

[1] An i-th unfolding matrix of an n-dimensional tensor $\mathbf{A}$ is a matrix where the first index is defined by the first i dimensions of the tensor $\mathbf{A}$ and the second index – by the last $n - i$ dimensions of $\mathbf{A}$.

3. Proof of theorem 2 and corollary 1

We propose algorithm 1 to compute the partition function of MRF. Reminding the notation the approximation of the product of the sequence of matrices $\{B_j\}_{j=i}^{n}$ in the TT-format is denoted by $\boldsymbol{f}_i$. We have $\boldsymbol{f}_n = B_n$ and $\boldsymbol{f}_1 = \tilde{Z}$. Using the matrix-by-vector product and the TT-rounding procedure algorithm 1 sequentially computes values $\boldsymbol{f}_i$:

$$\boldsymbol{f}_i = \operatorname{round}(B_i \boldsymbol{f}_{i+1}, \varepsilon),$$

where the TT-rounding precision controls the relative accuracy:

$$\left\| B_i \boldsymbol{f}_{i+1} - \boldsymbol{f}_i \right\|_2 \le \varepsilon \left\| B_i \boldsymbol{f}_{i+1} \right\|_2 . \tag{2}$$

Theorem 2. *For an MRF and a rounding parameter $\varepsilon \ge 0$ the absolute error of the partition function estimation $\tilde{Z}$ computed by algorithm 1 is bounded as follows:*

$$\begin{aligned}
&\left| Z - \tilde{Z} \right| \le \|B_1\|_2 \dots \|B_{n-2}\|_2 \cdot \|B_{n-1}\boldsymbol{f}_n - \boldsymbol{f}_{n-1}\|_2 + \\
&+ \|B_1\|_2 \dots \|B_{n-3}\|_2 \cdot \|B_{n-2}\boldsymbol{f}_{n-1} - \boldsymbol{f}_{n-2}\|_2 + \dots + \\
&+ \|B_1 \boldsymbol{f}_2 - \boldsymbol{f}_1\|_2 .
\end{aligned} \tag{3}$$

We start the proof with the following lemma.

Lemma 1. *For all $i = 1, \dots, n-1$ the following inequality holds:*

$$\begin{aligned}
&\|B_i \dots B_n - \boldsymbol{f}_i\|_2 \le \\
&\le \|B_i\|_2 \dots \|B_{n-2}\|_2 \cdot \|B_{n-1}\boldsymbol{f}_n - \boldsymbol{f}_{n-1}\|_2 + \\
&+ \|B_i\|_2 \dots \|B_{n-3}\|_2 \cdot \|B_{n-2}\boldsymbol{f}_{n-1} - \boldsymbol{f}_{n-2}\|_2 + \\
&+ \dots + \|B_i \boldsymbol{f}_{i+1} - \boldsymbol{f}_i\|_2 .
\end{aligned} \tag{4}$$

Proof. We prove the lemma by induction. Indeed, for $i = n-1$ we have

$$\|B_{n-1}B_n - \boldsymbol{f}_{n-1}\|_2 = \|B_{n-1}\boldsymbol{f}_n - \boldsymbol{f}_{n-1}\|_2 ,$$

where equality $\boldsymbol{f}_n = B_n$ holds by definition of $\boldsymbol{f}_n$.

Now suppose that (4) is true for all $i = j+1, \dots, n-1$. For $i = j$ we obtain

$$\begin{aligned}
&\|B_j \dots B_n - \boldsymbol{f}_j\|_2 = \\
&= \|(B_j \dots B_n - B_j \boldsymbol{f}_{j+1}) + (B_j \boldsymbol{f}_{j+1} - \boldsymbol{f}_j)\|_2 \le \\
&\le \|B_j\|_2 \|B_{j+1} \dots B_n - \boldsymbol{f}_{j+1}\|_2 + \|B_j \boldsymbol{f}_{j+1} - \boldsymbol{f}_j\|_2 \le \\
&\le \|B_j\|_2 (\|B_{j+1}\|_2 \dots \|B_{n-2}\|_2 \cdot \|B_{n-1}\boldsymbol{f}_n - \boldsymbol{f}_{n-1}\|_2 + \\
&+ \|B_{j+1}\|_2 \dots \|B_{n-3}\|_2 \cdot \|B_{n-2}\boldsymbol{f}_{n-1} - \boldsymbol{f}_{n-2}\|_2 + \\
&+ \dots + \|B_{j+1}\boldsymbol{f}_{j+2} - \boldsymbol{f}_{j+1}\|_2) + \|B_j \boldsymbol{f}_{j+1} - \boldsymbol{f}_j\|_2 ,
\end{aligned}$$

which is (4) for $i = j$. □

Proof of the theorem 2. Recall that $Z = B_1 \dots B_n$ and $\tilde{Z} = \boldsymbol{f}_1$. Therefore

$$|Z - \tilde{Z}| = |B_1 \dots B_n - \boldsymbol{f}_1| = \|B_1 \dots B_n - \boldsymbol{f}_1\|_2 .$$

The latter equation follows from the fact that both $B_1 \dots B_n$ and $\boldsymbol{f}_1$ are actually real numbers and in this case the absolute value and the vector L_2-norm coincide. Applying lemma 1 to the equation above completes the proof. □

Corollary 1. *For an MRF and a rounding parameter $\varepsilon \ge 0$ the absolute error of the partition function estimation $\tilde{Z}$ computed by algorithm 1 is bounded as follows:*

$$\left| Z - \tilde{Z} \right| \le \|B_1\|_2 \dots \|B_n\|_2 \left((1+\varepsilon)^{n-1} - 1\right). \tag{5}$$

We start with proving lemmas 2 and 3.

Lemma 2. *Inequality (6) holds for all $i = 1, \dots, n$:*

$$\|\boldsymbol{f}_i\|_2 \le \|B_i\|_2 \dots \|B_n\|_2 (1+\varepsilon)^{n-i}. \tag{6}$$

Proof. We prove the lemma by induction. For $i = n$ the statement immediately follows from the definition of $\boldsymbol{f}_n$.

Let (6) be true for $i = j+1 \le n$. Inequality (2) implies that

$$\begin{aligned}
\|\boldsymbol{f}_j\|_2 &= \|\boldsymbol{f}_j - B_j \boldsymbol{f}_{j+1} + B_j \boldsymbol{f}_{j+1}\|_2 \le \\
&\le \varepsilon \|B_j\|_2 \|\boldsymbol{f}_{j+1}\|_2 + \|B_j\|_2 \|\boldsymbol{f}_{j+1}\|_2 = \\
&= \|B_j\|_2 \|\boldsymbol{f}_{j+1}\|_2 (1+\varepsilon) \le \\
&\le \|B_j\|_2 \dots \|B_n\|_2 (1+\varepsilon)^{n-j},
\end{aligned}$$

where the last inequality follows from the induction assumption. □

Lemma 3. *Inequality (7) holds for all $i = 1, \dots, n$:*

$$\|B_i \boldsymbol{f}_{i+1} - \boldsymbol{f}_i\|_2 \le \|B_i\|_2 \dots \|B_n\|_2 \, \varepsilon (1+\varepsilon)^{n-i-1}. \tag{7}$$

Proof. The statement immediately follows from lemma 2 and the inequality

$$\|B_i \boldsymbol{f}_{i+1} - \boldsymbol{f}_i\|_2 \le \varepsilon \|B_i \boldsymbol{f}_{i+1}\|_2 \le \varepsilon \|B_i\|_2 \|\boldsymbol{f}_{i+1}\|_2 .$$

□

Proof of the corollary 1. By applying lemma 3 to inequality (3) we obtain

$$\begin{aligned}
|Z - \tilde{Z}| \le &\|B_1\|_2 \dots \|B_n\|_2 \, \varepsilon + \\
&+ \|B_1\|_2 \dots \|B_n\|_2 \, \varepsilon (1+\varepsilon) + \dots + \\
&+ \|B_1\|_2 \dots \|B_n\|_2 \, \varepsilon (1+\varepsilon)^{n-2} = \\
&= \|B_1\|_2 \dots \|B_n\|_2 \, \varepsilon \sum_{j=0}^{n-2} (1+\varepsilon)^j = \\
&= \|B_1\|_2 \dots \|B_n\|_2 \left((1+\varepsilon)^{n-1} - 1\right).
\end{aligned}$$

□

4. Details of the experimental setup

In our experiments we use the Ising model as the main playground. The energy of the model is defined as follows:

$$\mathbf{E}(\boldsymbol{x}) = -\frac{1}{T}\left(\sum_{i=1}^{n} x_i h_i + \sum_{\{i,j\}\in\mathcal{E}} c_{ij} x_i x_j\right), \qquad (8)$$

where variables x_i, $i = 1, \dots, n$ take values from the set $\{-1, 1\}$ and $\mathcal{E}$ is the connectivity system. We refer to coefficients h_i as unary weights, to c_{ij} as pairwise weights, and to parameter T as temperature. The connectivity system $\mathcal{E}$ defines pairwise connections between the variables. We typically use square 4-connected grids of 10×10 nodes. If all pairwise weights are equal ($c_{ij} = c$) we call the Ising model homogeneous and heterogeneous otherwise.

In sec. 6.2 we construct the plots in fig. 3 using the heterogeneous Ising model of size 10 with with unary and pairwise weights generated uniformly from $[-1, 1]$ with the temperature T set to 1.

In sec. 7.1 (fig. 2) we use a series of homogeneous Ising models based on 4-connected grids of increasing sizes: from 1×1 to 12×12. All the unary weights h_i are generated from the uniform distribution on segment $[-1, 1]$, the pairwise weight c equals 1, the temperature T equals 10.

In sec. 7.2 for experiment 1 (fig. 4a) we use a set of homogeneous Ising models of size 10×10 where unary weights h_i are generated again from the uniform distribution on segment $[-1, 1]$, the pairwise weight c equals 1, the temperature T varies from 10^{-1} to 10^{3}. For each value of the temperature we generate 50 models and report the absolute error of the logarithm of the computed partition functions (we show the median, lower and upper quartiles[2]).

In sec. 7.2 for experiment 2 (fig. 4b) we use a series of models generated by the authors of the WISH method (Ermon et al., 2013). These are heterogeneous Ising models (mixed attractive and repulsive potentials) of size 10×10, where the unary weights are generated uniformly on $[-1, 1]$, the temperature is fixed to 1, and the pairwise weights are generated uniformly from $[-f, f]$ with the parameter f varying from 0.25 to 3. We report the absolute error of the logarithm of the computed partition functions.

In sec. 7.2 for experiment 3 (fig. 5) we use homogeneous Ising models of size 10×10 where the unary potentials are generated uniformly from $[-1, 1]$, the pairwise weight equaled 1, the temperature varies from 10^{-1} to 10^{1}. For each value of the temperature we average results over 10 models.

In sec. 7.3 (fig. 6) we use heterogeneous Ising models of size 10×10, where the unary weights are generated uniformly from $[-1, 1]$, the temperature is fixed to 1, and the pairwise weights are generated uniformly from $[-f, f]$ with parameter f varying from 0 to 3. For each value of parameter f we generate 50 models and report the average absolute error of the marginal for the "+1" class, i.e. $0.01 \sum_{i=1}^{100} |\tilde{p}(x_i = +1) - p(x_i = +1)|$, where $\tilde{p}(x_i = +1)$ and $p(x_i = +1)$ are the approximate, and the true marginal probabilities of variable x_i taking value "+1" correspondingly. For each value of the parameter f we report the median, the lower and upper quartiles w.r.t. all the generated models.

[2] The median, lower and upper quartiles are defined as 50%, 25%, and 75% quantiles respectively.

References

Ermon, S., Gomes, C., Sabharwal, A., and Selman, B. Taming the curse of dimensionality: Discrete integration by hashing and optimization. In *International Conference on Machine Learning (ICML)*, 2013.

Oseledets, I. V. Tensor-Train decomposition. *SIAM J. Scientific Computing*, 33(5):2295–2317, 2011.