Another View on Optimization as Probabilistic Inference

Felix Gonsior, Nico Piatkowski, and Katharina Morik

TU Dortmund, AI Group, Dortmund, Germany
http://www-ai.cs.tu-dortmund.de

Abstract. We convert an optimization model for Boolean Matrix Factorization (BMF) into a Bayesian probabilistic model by plugging it into a probabilistic context. We infer the parameter distribution by using a state of the art sampling method based on Langevin diffusions. A visual analysis of the sampled uncertainty values shows a connection to the model uncertainty.

Keywords: Matrix factorization, Bayesian models, Markov-Chain Monte Carlo, Langevin diffusion, uncertainty

1 Introduction

Most machine learning models are trained by minimizing a loss function using optimization, this can often be implemented in an efficient way. Training a model in this way however does not yield information about the uncertainty of the learned parameters and about possible alternative parameters. Using Bayesian inference, models are trained by inferring a probability distribution over their parameters. Using Markov-Chain Monte Carlo (MCMC) sampling, the parameter distribution of the Bayesian probabilistic model is approximated by a finite sample, providing estimates of parameter uncertainty as well as alternative parameters. We believe, that in certain cases it is possible to simply "plug in" an optimization model into a probabilistic context. We evaluate this idea using the Boolean Matrix Factorization (BMF) problem[4]. While it has a concise mathematical description,the interactions between model variables are nontrivial. Gaining insight into these interactions by observing the parameter uncertainty in a probabilistic model of BMF is another way to strengthen trust in BMF solutions. Its properties make BMF a challenging problem for sampling algorithms, but therefore also an interesting test case for our exploration.

2 Boolean Matrix Factorization

Boolean Matrix Factorization belongs to the more prominent machine learning models. However, multiple issues with the BMF approach require a lot of work

from the user of existing methods to verify that a given solution can be trusted. Therefore the focus of recent work by Hess et al. on the PAL-Tiling framework[2] lies on increasing the trustworthiness of BMF by increasing the interpretability of its results.

BMF is an unsupervised learning method, which can discover meaningful whole-parts relationships within boolean databases. A boolean database is represented as a matrix $D \in \{0,1\}^{m \times n}$ where the columns are labeled with items (features) $\mathcal{I} = \{1, \ldots, n\}$ and the rows are labeled with transactions (data cases) $\mathcal{T} = \{1, \ldots, m\}$. If an item occurs in a transaction the corresponding matrix entry is set to one. As an example take the database of a movie rental service, where transactions represent customers and items represent movies. For customers that have rented a certain movie at least once, the corresponding matrix entry is set to one. Subsets of users which have rented similar subsets of movies form patterns[1]. BMF assumes that the matrix D can be constructed by the multiplying two smaller factor matrices. Assume factor matrices $X \in \mathbb{B}^{n \times r}$ and $Y \in \mathbb{B}^{m \times r}$ with the *factorization rank* $r > 0$ and $r << \min(m, n)$. The goal is to minimize the reconstruction error $|D - XY^T|$ as follows[2]

$$\min_{X,Y} |D - XY^T| + |X| + |Y| \qquad X \in \mathbb{B}^{n \times r}, Y \in \mathbb{B}^{m \times r} \tag{1}$$

Having a small factorization rank reduces the amount of space available in the factor matrices. If it it is possible to construct the full matrix D from the smaller matrices, these must contain the same information as D itself. In this way BMF achieves a compression of D. Due to the nontrivial interactions within the model, in many cases it is not clear if a given BMF solution can be trusted. It is unknown if the dataset factors as assumed by BMF and with which factorization rank. Non separable patterns might yield many false positives. Column permutations of the factor matrices do not change the objective value, blowing up the search space. Finally, NP-completeness as well as APX-hardness have been proven[4].

3 Implementation

With the formulation in eq. (1) we still have a hard to solve combinatorial optimization problem. Hess et al. obtain an approximate solution by relaxing the the parameters to the interval $[0, 1]$, thereby obtaining the related Nonnegative Matrix Factorization problem (NMF)[5]. They solve this problem with gradient descent, followed by reconstruction of binary factor matrices by thresholding. In our work we follow this idea, but instead of optimizing we use a probabilistic model over factor matrices X, Y, from which we obtain a sample.

We assume a Gibbs distribution $p(\cdot) = \frac{1}{Z} e^{\mathrm{E}(\cdot)}$, where $\mathrm{E}(\cdot)$ is called the energy function and Z is the normalization constant[2]. To plug an optimization model into this formula, we reinterpret the objective of the optimization problem as an energy function over parameters X, Y given the data D. In our case we chose the

[1] This is interpreted as different users having a similar taste in movies.
[2] As an integral over the whole parameter space, it is often intractable.

PANPAL[2] objective for the reconstruction error $F(X,Y,D) = \frac{1}{2}\|D - YX^T\|_2^2$ without the regularizer. Using F as the energy and adding a prior distribution $p(X,Y) = p(x_{11},\ldots,x_{nr},y_{11},\ldots,y_{mr}) \sim \mathcal{N}(\mathbf{0},\boldsymbol{I})$ in place of the regularizer we have

$$p(X,Y|D) = \frac{1}{Z(D)} e^{-F(X,Y,D)}\, p(X,Y)$$

Therefore, low reconstruction errors are associated with high probabilities. Standard normal priors for each parameter model sparsity assumptions for the factor matrices. Recent research has focused on gradient based MCMC sampling approaches inspired by physical models of motion. The Langevin diffusion $dX = \frac{1}{2}\nabla \log L_D(X) + dB$ with $\frac{dB}{dt} \sim \mathcal{N}(0,\sigma)$ models a particle moving in an external potential $L_D(X)$ and influenced by random effects dB[3]. In our case, the potential $L_D(X)$ is given by the log-likelihood function of our probabilistic model. This continuous stochastic process is guaranteed to converge in its stationary distribution[6] to $P_D(X) \sim L_D(X)$. Starting with the work of Welling and Teh[7], multiple samplers based on this principle were constructed. Some use minibatches $\tilde{D} \subset D$ to simulate the noise term dB for resource constrained operation. Recent work by Mandt et al.[3] gives an optimal step size ϵ^* for log-quadratic likelihoods. They develop the Iterate averaging Stochastic Gradient (IASG) sampler with the update equation

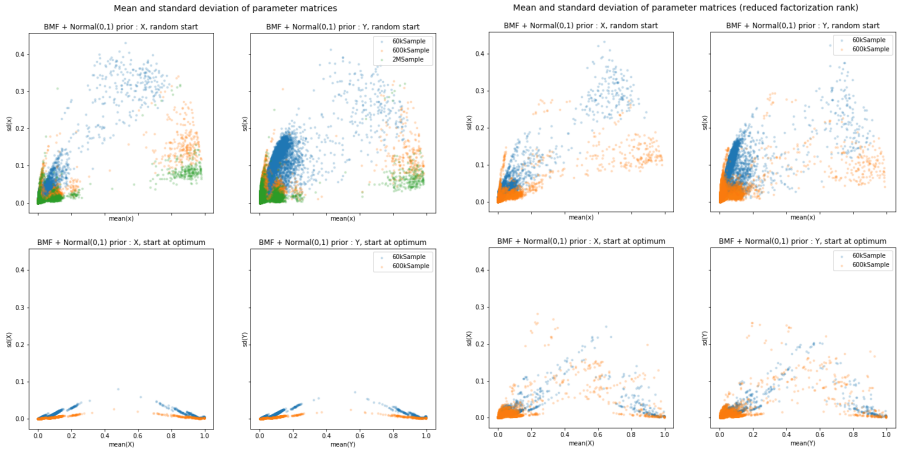
$$X_t = X_{t-1} + \epsilon^* \nabla_X \log L_D(X_{t-1}, \tilde{D}_t) \tag{2}$$

Here $\epsilon^* = \frac{N}{S}\frac{1}{\mathcal{F}(X)_{ii}}$, the optimal step size depends on the number of data cases N and the batch size S as well as on the diagonal of the empirical Fisher Information[1] $\mathcal{F}(X) = \widehat{\mathrm{cov}}\,[\nabla_X \log L_D(X)]$. Some changes to eq. (2) are necessary in our case, they are not presented here due to space constraints. The update equation is iterated starting from a given X_0. Decorrelated samples are produced by Polyak averaging over $\frac{N}{S}$ intermediate results X_t.

4 Results

We devised two scenarios to test our implementation. In one case the sampling process begins at a random point in parameter space. In the other case sampling starts at a known local optimum. For each scenario we generated test datasets with data matrices of size 512×512 and factorization ranks $r \in \{20, 30, 40\}$ using the method given in [2] leading to a parameter space of dim. 40960 for large instances. In the first test runs we drew samples of size 60k, which for the random start scenario did not yield proper estimates of the posterior parameter distribution but still showed tendencies. With samples of size 600k and 2M we observed improved sample quality. In figures 1a and 1b we have visualized samples for different configurations as scatterplots. Each point in the diagram relates to mean and standard deviation of an entry within a factor matrix with means on

[3] Brownian motion



(a) correct factorization rank (30) (b) reduced factorization rank (20)

Fig. 1: Scatter plot of the mean values vs. the standard deviations for parameter matrices X, Y for a BMF/NMF model with $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ prior. Minibatch size 512.

the x-axis and s.d. on the y-axis. Different colors mark samples of different sizes. As the factors are sampled on the relaxed (NMF) problem, they take values in $[0, 1]$. Mean values near to 0.5 signal internal conflicts in the model, mostly due to noisy data. In NMF this effect is known as fuzzy assignment between items and transactions. In this way the parameter values in NMF themselves express a type of uncertainty within the model. The standard deviation values on the y-axis are obtained through the sampling process and express uncertainty in the sense of our probabilistic model.

In the upper rows of figures 1a and 1b we show samples of different sizes when the element values for the initial parameter X_0, Y_0 were chosen uniformly random out of $[0, 1]$. For each sample size (color) we observe a dense cluster of mean values around zero (sparsity) and also a lower density cluster of mean values with high standard deviations. With increasing sample size this cluster moves towards mean one and towards lower standard deviations. As there is an abundance of zeros within the data and also a sparsity prior, it becomes clear that the correct assignment of the crucial nonzero values is only validated after many observations of the parameter space, i.e. after a long sampling period.

A totally different situation presents itself in the lower rows of figures 1a and 1b. In these plots the sampling process has been started at a known local optimum for the factor matrices. Also, in figure 1a the factorization rank matches the rank used when generating the data (30) while it is mismatched in figure 1b. In figure 1a we observe a very tight clustering of mean values around zero and one. In addition for each cluster we observe an almost linear relationship between the mean values and the corresponding standard deviations. With increasing sample size the standard deviations diminish, expressing very low probabilistic

uncertainty, while the spread in the mean values is still visible. For a matched factorization rank, both types of uncertainty express related facts about the model fit. The situation is different with a mismatched factorization rank, as shown in figure 1b. While we observe clustering at means zero and one we also observe values in between with large standard deviations. With increasing sample size these large standard deviations prevail meaning that some conflicts between data and model are not resolvable in terms of fuzzy assignments. In this way probabilistic uncertainty adds valuable information which cannot be obtained by only looking at fuzzy assignments.

5 Conclusion

We have "plugged in" the BMF/NMF problem into a probabilistic model and analyzed different samples. With mismatched factorization rank, the information about the rank mismatch is carried in the probabilistic uncertainty but not in the fuzzy assignment. Developing this insight is target of further research. Realizing sampling for BMF/NMF posed nontrivial challenges even with state of the art methods, requiring further research into sampling methods for high dimensional spaces.

Acknowledgement This research has been funded by the Federal Ministry of Education and Research of Germany (BMBF) as part of the competence center for machine learning ML2R (01|S18038A).

References

1. Girolami, M., Calderhead, B.: Riemann manifold Langevin and Hamiltonian Monte Carlo methods: Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**(2), 123–214 (Mar 2011)
2. Hess, S., Morik, K., Piatkowski, N.: The PRIMPING routine—Tiling through proximal alternating linearized minimization". Data Mining and Knowledge Discovery **31**(4), 1090–1131 (Jul 2017)
3. Mandt, S., Hoffman, M.D., Blei, D.M.: Stochastic Gradient Descent As Approximate Bayesian Inference. J. Mach. Learn. Res. **18**(1), 4873–4907 (Jan 2017)
4. Miettinen, Pauli and Mielikäinen, Taneli and Gionis, Aristides and Das, Gautam and Mannila, Heikki: The Discrete Basis Problem. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD. pp. 335–346. Springer (2006)
5. Paatero, P., Tapper, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics **5**(2), 111–126 (Jun 1994)
6. Roberts, G.O., Tweedie, R.L.: Exponential Convergence of Langevin Distributions and Their Discrete Approximations. Bernoulli **2**(4), 341 (Dec 1996)
7. Welling, M., Teh, Y.W.: Bayesian Learning via Stochastic Gradient Langevin Dynamics. In: Getoor, L., Scheffer, T. (eds.) Proceedings of the 28th International Conference on Machine Learning (ICML). pp. 681–688. ACM, New York, NY, USA (2011)