Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Aleksandra Burashnikova

PhD Program: Engineering Systems

Title of Thesis: Large-scale sequential learning for recommender and engineering systems

Supervisor: Assistant Professor Yury Maximov, Skoltech

Professor Massih-Reza Amini, University Grenoble Alpes

The thesis document includes the following changes in answer to the external review process.

Dear Jury Members,

I would like to thank you for making relevant comments and suggestions. Addressing those has led me to a revision of the thesis, which I feel has significantly improved it. In this document, my responses to your comments.

Sincerely yours,
Aleksandra Burashnikova

The change log:

- Substituted “mathcal F” with “\mathcal F”, p. 25.
- Corrected Eq. 2.6 by changing “-“ to “+”, p. 28.
- Corrected "hyperplan” on “hyperplane”, a typo, p. 31, 36.
- Provided the references to the sources of the Figures, p. 48-55.
- Provided a reference for the Table 6.2 in the text, p. 95.
- Added the explanation of the surrogate part in formula 4.3, Chapter 4.
- Reduced “A pictorial depiction…”, redundant formulation from Figure 4-1.
- Provided proof reading of Theorem 4, extended it with the additional parts mentioned in Supplementary materials of the corresponding paper published on the basis of Chapter 4 in ECML-PKDD’2019.
- Corrected the notation, now all vectors are in bold, p.31- 36, p.38-39, p.45-46, p.65-71, p.80-84, p.94-98.
- Figures 3-2 and 3-3 are deleted as redundant. Figure 3-4 (Figure 3-3 in new numeration) is changed on another one with good quality. Figure 3-6 (Figure 3-5 in new numeration) is changed on the C. Olah’s version.
- Added the definition of the Admittance matrix, p. 98.
- Notation of Chapter 6 is corrected, p. 95-98.
- Author’s contribution is added, p.17, Chapter 1.
- Added the contribution with the emphasis on the thesis results in Concluding Remarks, Chapter 7.
- Joined the Subsection 3.2.2 with the Section 3.2.
- Deleted Figure 3.7 as non-informative.
- Subsection 3.2.3 (3.2.2 in new numeration) was modified with focus on Prod2Vec.

- Added an explanation of AUC measure in the Subsection 3.3.5.
- Added the information about modeling recommender system as bipartite graph into the Subsection 3.2.3, Chapter 3.
- The description of the contribution of CNN and RNN for recommender systems is extended, p.49-52.
- The description of the quality measures is extended with the descriptions of their properties and prerequisites, p. 56-59.
- The nomenclature for Chapters 4-6 is added, p.9-10.
- The explanation about the benefits of SAROS and BPR with respect to all the remains state-of-the-art approaches is added into the Section 4.5, p.77.

Comments:

1. "What was the loss function used?"

   - In SAROS algorithm we minimize pairwise ranking loss over blocks of consecutive items constituted by a sequence of non-clicked items followed by a clicked one for each user. The surrogate part of the minimized loss function is represented by logistic loss. I added this information into the description of Formula 4.3, Chapter 4.

2. "How weights ω were structured?"

   - We represent each user *u* and each item *i* respectively by vectors $\boldsymbol{U}_u \in \mathbb{R}^k$ and $\boldsymbol{V}_i \in \mathbb{R}^k$ in the same latent space of dimension k. The set of weights to be found $\omega = (\boldsymbol{U}, \boldsymbol{V})$ are then matrices formed by the vector representations of users $\boldsymbol{U} = (\boldsymbol{U}_u)_{u\in[\mathrm{N}]} \in \mathbb{R}^{\mathrm{N}\times\mathrm{k}}$ and items $\boldsymbol{V} = (\boldsymbol{V}_i)_{i\in[\mathrm{M}]} \in \mathbb{R}^{\mathrm{M}\times\mathrm{k}}$. More details are given on the page 64, Chapter 4.

3. "What type of convergence is considered in Theorem 4?"

   - In theorem 4 we show that the proposed algorithm converges uniformly to the global minimum of the convex ranking loss:

$$\mathcal{L}(\omega) = \mathbb{E}_u\left[\frac{1}{|I_u^+||I_u^-|}\sum_{i\in I_u^+}\sum_{i'\in I_u^-} l_{u,i,i'}(\omega)\right].$$

4. "How stationary components are defined?"

   - As a measure of non-stationarity we used memory parameter inferred with GPH estimator. When the memory parameter is large, the time series tends to have a sample autocorrelation function with large spikes at several lags which is well known to be the signature of non-stationarity. We estimate memory parameter for each sequence of user's interactions over items ordered with respect to time. To do it, we extracted the gradients of loss function over the items after SAROS was trained. As because in the latent space each item is represented by the vector $\boldsymbol{V}_i \in R^k$ and *k* after cross-validation was installed to 4, then the gradient of loss function over each item has the dimension 4. The memory parameter then was estimated with GPH over 4d time series of gradients for each user and as the output was represented by the vector with 4 components. We then classify the time series as stationary if each component of memory parameter $d \leq 1/2$, and as non-stationary otherwise.

5. "What is the admittance matrix?"

   - Admittance is a measure of how easily a circuit or device will allow a current to flow. Admittance matrix is an *N x N* matrix describing a linear power system with *N* buses. It

represents the nodal admittance of the buses in a power system. The general mathematical expression of each element of the admittance matrix Y is represented as following:

$$Y_{ij} = \begin{cases} y_i + \sum_{\substack{k=1,..,N \\ k \neq i}} y_{ik}, & i = j \\ -y_{ij}, & i \neq j \end{cases}$$

Where y_i is the admittance of linear loads connected to bus *i* as well as the admittance-to ground at bus *i*, and y_{ik} is the admittance between the bus *i* and another bus *k* connected to bus *i*. This information was added into the Chapter 6, p. 98.

6. "What are feature vectors *x*, φ and ψ?"

   - x^t are the measurements of the parameters of power grid at each time step *t* (the notation was corrected, p. 95-98). ψ is a feature vector that is represented as a product between the bus voltages variations ΔU before and during the faults and the admittance matrix before the faults. φ is a mistake on the figure 6-1, it was fixed.

7. "What is the rationale behind convolutional structure of the net?"

   - In the Chapter 6 we don't define the architecture for the prediction of the faults in a power grid. We propose the idea of the improvement for the already existed model based on the modification of the loss function, where we include the additional term responsible for the predictions the neighbours of the faulted line. This modification suggests that if the fault happens in the particular line then it possibly something wrong with the its neighbours and these problems also should be fixed.

8. "The author's contribution needs to be clearly stated as all papers of Ms. Burashnikova were published with a large number of co-authors."

   - The personal contribution to the papers includes all the experimental parts, except memory estimation in Chapter 5 as well as the partial contribution in the theoretical parts under the supervision of Yury Maximov, Marianne Clausel and Massih-Reza Amini. This information now is mentioned on the p.17, Chapter 1.

9. "The contribution should be stated more clearly with the emphasis on the thesis results impact."

   - The information about the contribution on the thesis results impact is added in the Chapter 7, in concluding remarks. The main contribution of the thesis is the first part devoted to recommender systems, where we designed SAROS algorithm and provided theoretical provement on the convergence property of the proposed approach for the case when the loss function is convex and for the general case. Also, in this part was considered the way of the algorithm improvement by suggesting the strategy of filtering non-stationary users from the training dataset. The second part of the thesis is mostly devoted on the practical application of ranking model in engineering systems with the contribution on the improvement of the existing model for ranking the faulted lines in power grids by taking into account the neighbours of the faulted line in the loss function.

10. "In terms of the notation, it is strange that all the vectors are not in bold."

    - I've corrected the notation, now all vectors are in bold, Corrections are done on the p.31- 36, p.38-39, p.45-46, p.65-71, p.80-84, p.94-98.

11. "Section 3.2.2 on scheduling approaches seems very small compared to the literature on the subject and the positioning of the thesis."

- Thank you for taking it into account, you are completely right, now I joined this section to the description of the Section 3.2, Chapter 3.

12. "Figures 3-7 and 3-8 are not clear enough: they do not allow a good understanding of the contribution of the architectures."

    - The Figure 3-7 was deleted, I agree that it doesn't clearly describe the architecture of the model. Regarding Figure 3-8 (Figure 3-6 in new numeration), I keep this figure, as it describes the main idea of the model, how the mini-batches was built as because in recommender systems we're dealing with the sessions of completely different sizes with the distinction by orders of magnitude. The description of the idea of building mini-batches in the thesis was a bit extended with more comprehensive explanation.

13. "To keep the anchoring in recommender systems, section 3.2.3 could have been focused directly on prod2vec, by simply explaining the linguistic origin of the approach."

    - This part (3.2.2 in new numeration) was extended with more focus on prod2vec.

14. "Given the interest of this thesis for ranking systems, metrics based on AUC could be mentioned (AUC itself or ATOP for example). In terms of organization, the highlighting of MSE flaws would benefit from being presented in this last section rather than in the matrix factorization."

    - The Subsection 3.3.5 is created with the AUC explanation. We decided to keep MSE in the Subsection 3.2.1 instead of Subsection 3.3.5, as because it doesn't usually used in the raking problem as the ranking task isn't mostly considered as the regression task.

15. "The description of GCNN could have been preceded by an explanation of how to model the recommender system as a bi-partite graph."

    - The explanation of how to model recommender system as a bipartite graph is added into the Subsection 3.2.3 of Chapter 3.

16. "The description of the contributions of CNN & RNN for recommender systems is a bit succinct. In particular, the use of these architectures for modeling user sessions would deserve analyses regarding the representation of items and possible new tasks. The figures representing the architectures would have benefited from being more focused on the recommendation systems to better highlight the contribution of these approaches to this application. Matrix factorization is an interesting approach to bridge the gap between classical learning techniques and deep learning: the general organization of the section could have benefited from a transition around these aspects."

    - The description of the contribution of CNN and RNN for recommender systems is extended, p.49-52. Also, the benefits of the architectures for CNN and RNN baseline models from the figures 3-4 and 3-6 are added into the description of these models in the Subsection 3.2.3. The explanation of the transition between classical learning technique and deep learning is presented in the Subsection 3.2.2, where the similarities in terms of learning users and items representations for classical matrix factorization approach and neural networks based approaches are considered.

17. "The description is clear and factual but could have put more emphasis on the properties and prerequisites of the different metrics"

    - The descriptions of the quality measures are extended with the descriptions of their properties and prerequisites, p. 56-59, Chapter 3.

18. “In terms of form, a standardization of notations with the previous chapter would have benefited both chapters.”

    - It’s quite difficult to provide the same notation for Chapter 3 with Chapters 4 and 5, as because Chapter 3 has an overview nature and includes the background materials necessary for further understanding of the main contribution to the thesis. That is why we have divided the thesis into 2 parts: Part 1 “State-of-the-art” and Part 2 “Contribution”, where the nomenclature for the second part of the thesis is given on the pages 9-10.

19. “As a minor remark, we notice that the performance in this setting is very favorable to BPR and SAROS compared to more recent approaches: a more thorough analysis of this fact would be interesting.”

    - We suggest that this is an effect of the usage of both kind of feedback information (positive and negative) in SAROS and BPR, whereas last published popular neural networks-based approaches Caser, GRU4Rec and SASRec used only positive feedback in training the predictions. This information is added into the Section 4.5, p.77.