Comparing Recommendation Losses under Negative Sampling ⋆

Giulia Di Teodoro[1,*], Federico Siciliano[2,*], Nicola Tonellotto[1,*] and Fabrizio Silvestri[2,*]

[1]Information Engineering Department, University of Pisa, Pisa, Italy
[2]Department of Computer, Control and Management Engineering, Sapienza University of Rome, Italy

Abstract

Loss functions, such as categorical cross-entropy (CCE), binary cross-entropy (BCE), and Bayesian personalized ranking (BPR), play a central role in training modern recommender systems. Although evaluations are often based on ranking metrics, such as Normalized Discounted Cumulative Gain (NDCG) and Mean Reciprocal Rank (MRR), a direct understanding of how these losses relate to target metrics remains incomplete. Furthermore, full-item training is computationally prohibitive, which has led to the widespread use of negative sampling. In this extended abstract, we (i) derive theoretical equivalences and bounds relating these loss functions under negative sampling; (ii) prove that BPR and CCE become identical under a single negative sample; and (iii) show that BCE provides the tightest bound on NDCG and MRR when negative sampling is used. We complement our theoretical findings with empirical results on five datasets and four neural architectures, which consistently validate the theory.

Keywords

Recommender Systems, Loss Functions, Negative Sampling, Ranking Metrics

1. Introduction

Recommender systems (RSs) have become indispensable in e-commerce, media streaming, and social platforms. Training these models usually involves optimizing a loss function based on user-item interactions. Common choices include:

- Categorical Cross-Entropy (CCE): Treats recommendations as a multi-class classification problem across all items.
- Binary cross-entropy (BCE): Considers each positive interaction against sampled negatives in a binary classification framework.
- Bayesian Personalized Ranking (BPR)[2]: Directly optimizes pairwise ranking by contrasting positive and negative items.

Despite their prevalence[3, 4, 5, 6, 7, 8, 9, 10], the formal connection between these losses and downstream ranking metrics (e.g., NDCG[11], MRR) is often assumed rather than proven. Moreover, real-world systems rely on negative sampling [2, 12, 13, 14, 15, 16, 17], i.e. selecting a subset of unobserved items per positive, due to scalability concerns. Our prior work [1] addressed this issue by providing a unified theoretical framework to analyze how loss functions behave under negative sampling and how they bound the actual ranking objectives of interest.

IIR2025: 15th Italian Information Retrieval Workshop, 3th - 5th September 2025, Cagliari, Italy

⋆This work is an extended abstract based on the publication "A Theoretical Analysis of Recommendation Loss Functions under Negative Sampling" accepted at the International Joint Conference on Neural Networks 2025 [1].

*Corresponding authors.
giulia.di.teodoro@ing.unipi.it (G. Di Teodoro); siciliano@diag.uniroma1.it (F. Siciliano); nicola.tonellotto@unipi.it (N. Tonellotto); fsilvestri@diag.uniroma1.it (F. Silvestri)
https://coda.io/@federico-siciliano/federico-siciliano (F. Siciliano)
0000-0002-0418-0067 (G. Di Teodoro); 0000-0003-1339-6983 (F. Siciliano); 0000-0002-7427-1001 (N. Tonellotto); 0000-0001-7669-9055 (F. Silvestri)

2. Theoretical Analysis

We consider a user u and a set of items $\mathcal{I}$, where $\mathcal{P}_u \subset \mathcal{I}$ are positive (interacted) items and $\mathcal{N}_u$ are negative (non-interacted) items. Let $s(u, i)$ denote the model's score for item i.

2.1. Loss Definitions under Sampling

CCE - Categorical cross-entropy:

$$\mathcal{L}_{\text{CCE}} = -\log \frac{e^{s(u,i^+)}}{e^{s(u,i^+)} + \sum_{j=1}^{k} e^{s(u,i_j^-)}}$$

BCE - Binary cross-entropy:

$$\mathcal{L}_{\text{BCE}} = -\log \sigma\left(s(u, i^+)\right) - \sum_{j=1}^{k} \log\left(1 - \sigma(s(u, i_j^-))\right)$$

BPR - Bayesian personalized ranking:

$$\mathcal{L}_{\text{BPR}} = -\sum_{j=1}^{k} \log \sigma\left(s(u, i^+) - s(u, i_j^-)\right)$$

where $i^+ \in \mathcal{P}_u$ and $i_j^- \in \mathcal{N}_u$.

2.2. Ranking Metrics

The Normalized Discounted Cumulative Gain (NDCG) is a widely used recommendation metric that accounts for the graded relevance of items depending on their position in the ranked list:

$$NDCG(r_+) = \frac{1}{log_2(1 + r_+)}$$

if there's only one relevant item and r_+ is its rank position.

Another key ranking measure is the Mean Reciprocal Rank (MRR), which computes the inverse of the rank position r_+ of the first relevant item in the recommendations:

$$MRR(r_+) = \frac{1}{r_+}$$

2.3. Equivalence of BPR and CCE

Under a *single* negative sample ($k = 1$), we prove that ℓ_{BPR} is equivalent to ℓ_{CCE}.

Proposition 1. *$\ell_{BPR} = \ell_{CCE}$ if one negative item $K = 1$ is sampled for each user.*

This highlights that, when sampling only one negative per positive, optimizing CCE or BPR leads to the same parameter updates.

2.4. Equivalence of Global Minima

We now present a result that establishes the equivalence of the global minima of the three loss functions when a single negative is sampled and item scores are bounded.

Proposition 2. *If $s_+, s_i \in [-S, S]$ for some $S > 0$, then:*

$$\arg\min_{s_+} \ell_{BCE} = \arg\min_{s_+} \ell_{BPR} = \arg\min_{s_+} \ell_{CCE} = S \quad \arg\min_{s_i} \ell_{BCE} = \arg\min_{s_i} \ell_{BPR} = \arg\min_{s_i} \ell_{CCE} = -S$$

This proposition implies that, under bounded scores and single negative sampling, BPR, BCE, and CCE converge to the same optimal solution. Practically, it means that the choice of loss function does not affect the ideal parameter configuration.

However, in deep neural networks, these extreme score values are rarely reached due to regularization, early stopping, and model inductive biases, which prevent overfitting and favour generalization [18, 19]. Hence, while useful, this result has limited applicability to real-world RS training scenarios.

2.5. Bounding Ranking Metrics

We now turn to the comparison of ranking losses from the perspective of their ability to upper bound ranking metrics, particularly $-\log(\text{NDCG})$, under uniform negative sampling.

Theorem 1. *When uniformly sampling K negative items, in the worst-case scenario, and $s_+ \geq 0$:*

$$\begin{aligned}\mathbb{P}(-\log \mathit{NDCG}(r_+) \leq \ell_{BCE}) \geq \\ \mathbb{P}(-\log \mathit{NDCG}(r_+) \leq \ell_{BPR}) \geq \\ \mathbb{P}(-\log \mathit{NDCG}(r_+) \leq \ell_{CCE})\end{aligned}$$

This result shows that BCE offers the tightest bound on NDCG among the three losses, followed by BPR and then CCE. While the exact behaviour depends on the rank r_+ of the positive item and the number of sampled negatives K, BCE consistently exhibits more favourable properties, especially when item embeddings remain well-distributed, avoiding embedding collapse [20].

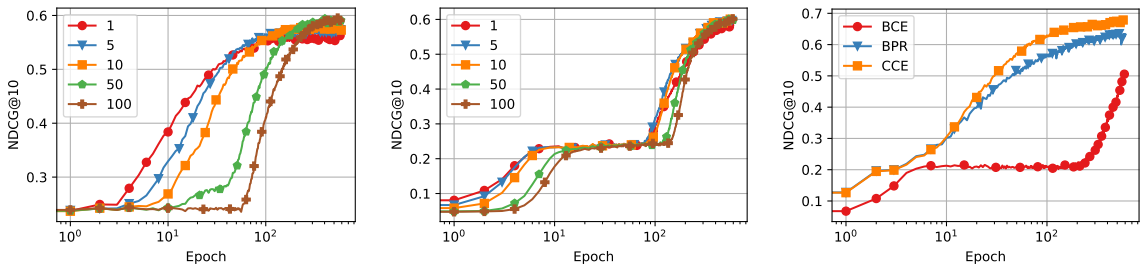
That said, practical dynamics during training, such as changing item ranks, differences in optimization behaviour between losses, and embedding concentration due to popularity bias, can affect these bounds. Thus, while BCE is theoretically preferable, its advantage may vary in real-world scenarios.

Additional theorems, full proofs, and the extension to MRR can be found in the original work [1].

3. Empirical Evaluation

We validate our theoretical insights on five benchmarks (MovieLens-1M[21], Amazon-Beauty[22], Amazon-Books[22], Yelp [23], and Foursquare NYC [24]) and four architectures (matrix factorization[25], Self-attentive Sequential Recommendation (SASRec) [26], GRU4Rec[27], and LightGCN[28]). For each setting, we vary $k \in \{1, 5, 10, 20\}$ negatives per positive and measure NDCG@10 and MRR [29].

3.1. Effect of Negative Sampling



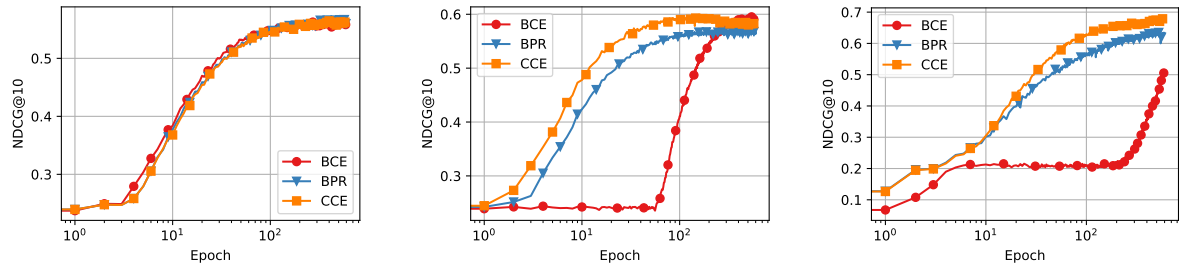
(a) BCE – GRU4Rec (b) BCE – SASRec (c) BCE – GRU4Rec (Foursquare)

Figure 1: NDCG@10 over training epochs for BCE on ML-1M (GRU4Rec, SASRec) and Foursquare.

We analyze how varying the number of negative items affects training on ML-1M using BCE. As shown in Fig. 1, fewer negatives yield faster improvements in early epochs, while a larger number (e.g., 100) leads to slower starts but better final performance. This reflects a trade-off: fewer negatives ease

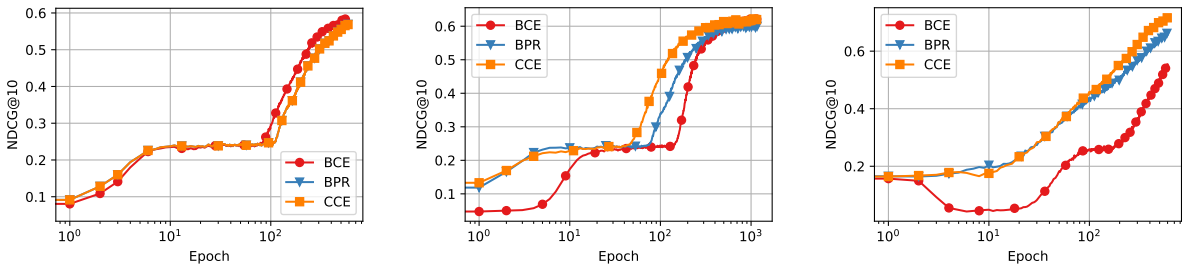
early learning, but more negatives improve generalization by providing harder contrasts. For BPR and CCE, we observe similar trends with slightly more stable early-phase training (see complete results in the original paper).

3.2. Loss Comparison: 1 vs 100 Negatives



(a) ML-1M (1 negative) (b) ML-1M (100 negatives) (c) Foursquare (100 negatives)

Figure 2: Loss comparison (BCE, BPR, CCE) with GRU4Rec on ML-1M and Foursquare.



(a) ML-1M (1 negative) (b) ML-1M (100 negatives) (c) Foursquare (100 negatives)

Figure 3: Loss comparison (BCE, BPR, CCE) with SASRec on ML-1M and Foursquare.

Figs. 2 and 3 compare loss functions using 1 and 100 negative samples. With a single negative, BPR and CCE perform identically on SASRec, as predicted by theory. BCE shows superior final performance, confirming its tighter bound to ranking metrics. On GRU4Rec, differences between losses are smaller.

When using 100 negatives, CCE generally performs better than BPR early in training, while BCE starts slower but steadily improves, surpassing both losses in later epochs. On Foursquare (Figs. 2c and 3c), BCE again starts behind but shows strong late-phase gains. However, due to its slower convergence, CCE often remains the most stable choice in early-to-mid training.

4. Conclusion and Future Work

We presented a unified theoretical framework that (i) links popular recommendation losses under negative sampling, (ii) uncovers an equivalence between BPR and CCE for a single negative, and (iii) establishes BCE as the preferred surrogate for ranking metrics. Future directions include extending the analysis to dynamic sampling schemes and to include gradient descent dynamics.

Acknowledgments

This work was partially supported by projects FAIR (PE0000013) and SERICS (PE00000014), under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU, and

project NEREO (Neural Reasoning over Open Data), funded by the Italian Ministry of Education and Research (PRIN) Grant no. 2022AEFHAZ.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 and DeepL for grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] G. Di Teodoro, F. Siciliano, N. Tonellotto, F. Silvestri, A theoretical analysis of recommendation loss functions under negative sampling, 2025 International Joint Conference on Neural Networks (IJCNN) (2025).

[2] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09, AUAI Press, Arlington, Virginia, USA, 2009, p. 452–461.

[3] A. V. Petrov, C. Macdonald, gsasrec: Reducing overconfidence in sequential recommendation trained with negative sampling, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 116–128. URL: https://doi.org/10.1145/3604915.3608783. doi:10.1145/3604915.3608783.

[4] C. Xu, Z. Zhu, J. Wang, J. Wang, W. Zhang, Understanding the role of cross-entropy loss in fairly evaluating large language model-based recommendation, 2024. arXiv:2402.06216.

[5] J. Wu, X. Wang, X. Gao, J. Chen, H. Fu, T. Qiu, On the effectiveness of sampled softmax loss for item recommendation, ACM Trans. Inf. Syst. 42 (2024). URL: https://doi.org/10.1145/3637061. doi:10.1145/3637061.

[6] S. Bruch, X. Wang, M. Bendersky, M. Najork, An analysis of the softmax cross entropy loss for learning-to-rank with binary relevance, in: Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval, 2019, pp. 75–78.

[7] Y. Pu, X. Chen, X. Huang, J. Chen, D. Lian, E. Chen, Learning-efficient yet generalizable collaborative filtering for item recommendation, in: Forty-first International Conference on Machine Learning (ICML), 2024.

[8] W. Yang, J. Chen, X. Xin, S. Zhou, B. Hu, Y. Feng, C. Chen, C. Wang, Psl: Rethinking and improving softmax loss from pairwise perspective for recommendation, in: The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

[9] A. Bacciu, F. Siciliano, N. Tonellotto, F. Silvestri, Integrating item relevance in training loss for sequential recommender systems, in: Proceedings of the 17th ACM Conference on Recommender Systems, 2023, pp. 1114–1119.

[10] F. Siciliano, S. Lagziel, I. Gamzu, G. Tolomei, Robust training of sequential recommender systems with missing input data (2024).

[11] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques, ACM Trans. Inf. Syst. 20 (2002) 422–446. URL: https://doi.org/10.1145/582415.582418. doi:10.1145/582415.582418.

[12] J. Weston, S. Bengio, N. Usunier, Wsabie: scaling up to large vocabulary image annotation, in: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three, IJCAI'11, AAAI Press, 2011, p. 2764–2770.

[13] C. J. C. Burges, From RankNet to LambdaRank to LambdaMART: An Overview, Technical Report, Microsoft Research, 2010. URL: http://research.microsoft.com/en-us/um/people/cburges/tech_reports/MSR-TR-2010-82.pdf.

[14] F. Yuan, G. Guo, J. M. Jose, L. Chen, H. Yu, W. Zhang, Lambdafm: Learning optimal ranking with factorization machines using lambda surrogates, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, Association for Computing

Machinery, New York, NY, USA, 2016, p. 227–236. URL: https://doi.org/10.1145/2983323.2983758. doi:10.1145/2983323.2983758.

[15] S. Rendle, C. Freudenthaler, Improving pairwise learning for item recommendation from implicit feedback, in: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, WSDM '14, Association for Computing Machinery, New York, NY, USA, 2014, p. 273–282. URL: https://doi.org/10.1145/2556195.2556248. doi:10.1145/2556195.2556248.

[16] D. Lian, Q. Liu, E. Chen, Personalized ranking with importance sampling, in: Proceedings of The Web Conference 2020, 2020, pp. 1093–1103.

[17] Y. Zhao, R. Chen, R. Lai, Q. Han, H. Song, L. Chen, Augmented negative sampling for collaborative filtering, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 256–266. URL: https://doi.org/10.1145/3604915.3608811. doi:10.1145/3604915.3608811.

[18] P. Baldi, Z. Lu, Complex-valued autoencoders, Neural Networks 33 (2012) 136–147. URL: https://www.sciencedirect.com/science/article/pii/S089360801200127X. doi:https://doi.org/10.1016/j.neunet.2012.04.011.

[19] L. Palagi, Global optimization issues in deep network regression: an overview, J. Glob. Optim. 73 (2019) 239–277.

[20] X. Zhao, M. Wang, X. Zhao, J. Li, S. Zhou, D. Yin, Q. Li, J. Tang, R. Guo, Embedding in recommender systems: A survey, arXiv preprint arXiv:2310.18608 (2023).

[21] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, ACM Trans. Interact. Intell. Syst. 5 (2015). URL: https://doi.org/10.1145/2827872. doi:10.1145/2827872.

[22] J. McAuley, C. Targett, Q. Shi, A. van den Hengel, Image-based recommendations on styles and substitutes, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 43–52. URL: https://doi.org/10.1145/2766462.2767755. doi:10.1145/2766462.2767755.

[23] N. Asghar, Yelp dataset challenge: Review rating prediction, arXiv preprint arXiv:1605.05362 (2016).

[24] D. Yang, D. Zhang, V. W. Zheng, Z. Yu, Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns, IEEE Transactions on Systems, Man, and Cybernetics: Systems 45 (2015) 129–142. doi:10.1109/TSMC.2014.2327053.

[25] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer 42 (2009) 30–37.

[26] W. Kang, J. McAuley, Self-attentive sequential recommendation, in: 2018 IEEE International Conference on Data Mining (ICDM), IEEE Computer Society, Los Alamitos, CA, USA, 2018, pp. 197–206. URL: https://doi.ieeecomputersociety.org/10.1109/ICDM.2018.00035. doi:10.1109/ICDM.2018.00035.

[27] B. Hidasi, A. Karatzoglou, L. Baltrunas, D. Tikk, Session-based recommendations with recurrent neural networks, in: International Conference on Learning Representations (ICLR), 2016.

[28] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang, Lightgcn: Simplifying and powering graph convolution network for recommendation, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 639–648. URL: https://doi.org/10.1145/3397271.3401063. doi:10.1145/3397271.3401063.

[29] F. Betello, A. Purificato, F. Siciliano, G. Trappolini, A. Bacciu, N. Tonellotto, F. Silvestri, A reproducible analysis of sequential recommender systems, IEEE Access (2024).