Survey on Learning-Augmented Algorithms

Thang Chu
Department of Computing Science
University of Alberta
Edmonton, Canada
thang@ualberta.ca

1 Paper Review

1.1 The primal-dual method for learning augmented algorithms

Bamas et al. (2020) extend the online primal-dual (PD) algorithms (Buchbinder et al., 2009) to include machine-learned (oracle) predictions. The new algorithms, called PDLA, use prediction as an additional input to accelerate the updating of primary variables towards optimal solutions. The authors design novel algorithms to tackle three problems: online set cover, ski rental, and Transmission Control Protocol (TCP) acknowledgement. Also, the paper makes no assumptions about the model's construction.

The paper presents unique contributions relative to other studies. First, it proposes a novel way to combine the strengths of PD and learning-augmented (LA) approaches. The PD framework is systematic and flexible, capable of representing and solving a large class of online optimization problems efficiently. Meanwhile, all LA algorithms have *consistency* and *robustness* properties. These properties enable the algorithms to surpass the established lower bounds of classical online algorithms when the predictions are correct, while maintain the performance within a reasonable bound in the case of wrong predictions. Second, the paper introduces an alternative perspective on the definitions robustness and consistency. Although previous works defined consistency and robustness in terms of prediction error $\eta \in [0, \infty)$, this paper introduces using a single parameter $\lambda \in [0, 1]$ to define these two properties, which also alter the derived competitive ratio. Intuitively, a smaller value of λ indicates more confidence in the prediction's quality. We will refer this type of competitive ratio as the λ-*competitive ratio*.

There are several strengths related to the paper. First, PDLA significantly improves the competitive ratio over the previous ratio. Second, the algorithms and their proof have undergone minor changes compared to the previous version, making them easier to understand. Third, the discrete design of PDLA simplifies the implementation of the algorithm. Furthermore, the author provides an experiment for the TCP acknowledgement problem to support the effectiveness of the algorithm. Lastly, since the predictions are problem-dependent, there is potential to generalize the idea to solve other online problems.

The paper also exhibits several weaknesses. First, the main theorems exclude the additive term without providing an explanation. Ignoring this additive term can weaken the analysis, especially when the predictions are partially correct. Second, since PDLA relies on discrete PD algorithms, it does not capture the interrelation between primal and dual variables. Hence, designing a continuous PDLA will better reflect the essence of primal-dual schema. Third, the author only prove the optimal λ trade-off for the ski rental problem but does not prove it for the online set cover and TCP acknowledgment problem. Lastly, the paper does not mention a key assumption about the learnability of the predictor. Without this assumption, it is difficult to obtain a sample-efficient predictor within a reasonable time complexity. There is a minor issue such that the terms $S(A, I)$ and C_{nc} are only described in words without proper definition.

1.2 Competitive caching with machine learned advice

Lykouris and Vassilvitskii (2020) add machine-learned predictions to the basic Marker algorithm (Fiat et al., 1991) to create an effective eviction strategy. They refer to this new version "Predictive Marker." The paper focuses on the unweighted caching setting, in which each page incurs a cost of one. Also, the paper makes no assumption about the model's construction.

The paper presents some distinct contributions compared to others. First, Predictive Marker algorithm proposes a novel way to tackle the naive eviction strategy of the previous algorithm. The Marker algorithm runs in phases. Each phase ends with the cache being full and all pages being marked. Then, a new phase begins by unmarking all of pages from the cache. When a page arrives and it is in the cache, the algorithm will mark it. Otherwise, the algorithm randomly evicts one unmarked page from the cache. By incorporating the prediction, Predictive Marker tackles the naivety of the previous eviction strategy. Specifically, the Predictive Marker algorithm assigns each page with a time prediction of the next request and evicts an unmarked page with the furthest-in-future prediction. Second, the paper defines the robustness and consistency properties based on the total loss η between predictions and true labels. Particularly, the consistency property holds if the predictor has a small loss function. Otherwise, the worst-case analysis is still $O(\log k)$, where k is the cache size. We refer this type of competitive ratio *η-competitive ratio*.

Numerous advantages are present in this paper. First, the author rigorously shows an improved competitive ratio of Predictive Marker algorithm using both L1 and L2 loss functions. Second, the next arrival time is more natural choice of prediction compared to the previous paper. Third, experiments comparing Predictive Marker with the popular Least Recently Used (LRU) algorithm are also provided.

However, several weaknesses are present. First, the algorithm is designed for unweighted caching, where each page has the same eviction cost. In a weighted setting, inaccurate predictions for high-cost pages could lead to an unbounded competitive ratio. Incorporating page weights into the eviction decision may mitigate this. Second, the analysis focuses only on L1 and L2 loss functions. Exploring alternatives such as a more robust square root loss function could be beneficial. Third, the current definition of loss function is sensitive to duplication. For instance, if we double the sequence of page requests, the total error will also double. Even though the maximum competitive ratio is $O(\log k)$, normalizing the loss over the length of sequence may address the issue. Additionally, the proof could be clarified by better motivating the concept of "chain eviction" and including a detailed example.

2 Proposal

Lykouris and Vassilvitskii (2020) and Bamas et al. (2020) established the foundational work on how to design online learning-augmented algorithms. Since then, researchers have conducted numerous studies to explore various ideas and settings. Therefore, our goal is to do a comprehensive **survey** about the field of LA algorithms.

Previous survey (Mitzenmacher and Vassilvitskii, 2020) proposed separating the LA algorithms based on what problems they solved. This strategy relies on the concept that online algorithms are usually designed to solve a specific problem. It has the benefit of showing differences among algorithms within the same problem. However, there are several drawbacks with this approach. First, it is unable to show the relationship among various algorithms across different problems. For instance, Bamas et al. (2020) suggested representing multiple problems under the same covering problem and used PDLA to solve all of them. Second, some papers (Antoniadis et al., 2022; Anand et al., 2022b) focused on generalizing the LA setting, instead of solving for a particular problem. Lastly, since robustness and consistency influence the algorithm design, algorithms (Lykouris and Vassilvitskii, 2020; Antoniadis et al., 2022) solving different problems are not necessarily distant. In order to overcome the limitations of the previous survey, we suggest conducting a systematic review of the papers from another viewpoint along the following dimensions:

- **Prediction: limited versus full information**: This setting aims to limit the advice's information through reducing the number of queries or constraining the size of the advice. In certain applications, it may not be optimal to query predictions every round due to potentially repetitive information or the high cost associated with such queries. While the previous two papers assume the algorithm can query prediction when a new item arrives, there are several papers that studied this setting for the caching (Antoniadis et al., 2022) and ski rental (Drygala et al., 2023) problem.

- **Predictor: single versus multiple**: The setting assumes the existence of multiple predictors advising the algorithm. In practice, if we want to train an image classifier, we can obtain multiple models from various algorithms. Previous two papers assume that we have access to one single predictor, while there are several papers generalizing the setting to solve set cover, caching (Anand et al., 2022a) and ski rental (Gollapudi and Panigrahi, 2019) problems.

- **Predictor: black-box versus grey-box**: PDLA and Predictive Marker assume there exists a *black-box* prediction model, meaning the algorithms have no knowledge about the model's construction. In practice, it is more reasonable to assume we have some information such as what kind of algorithm is used to train the model. Therefore, this setting allows the algorithms to access additional information about the model and enable them to make more informed decisions. There are some works for this setting to address the ski rental (Anand et al., 2022b) and prophet inequality (Diakonikolas et al., 2021) problem.

- **Performance measure: λ- versus η-competitive ratio**: From above paper review, there are different ways of defining robustness and consistency leading to $\lambda-$ and $\eta-$competitive ratio. The difference in these definitions may lead to different algorithm designs, as shown in these two papers (Bamas et al., 2020; Lykouris and Vassilvitskii, 2020). An understanding of the motivation behind these definitions and algorithm design can be achieved through a systematic review of algorithms based on these definitions.

- **Algorithm: systematic versus ad-hoc**: Previous LA algorithms were designed as extensions of randomized algorithms. These algorithms aimed to solve specific problems, such as caching (Lykouris and Vassilvitskii, 2020) and secretary (Dütting et al., 2020) problems. In this document, we refer to these algorithms as *ad-hoc* designs. However, Bamas et al. (2020) proposes to design online LA algorithms more systematically using a primal-dual framework. This framework is powerful and flexible, capable of representing a wide range of optimization problems. More recent works have been done in this direction, including online covering (Grigorescu et al., 2022) and packing (Grigorescu et al., 2024) problems. Therefore, reviewing these papers has the potential to unify previous approaches under the same class.

Our survey will concentrate on the five dimensions mentioned above. For each dimension, we divide various algorithms into the problem-specific sub-category. We will conduct a comprehensive survey by summarizing the key contributions, strengths, and drawbacks of each method and comparing them with one another. For the ad-hoc algorithm design, we plan only to recognize algorithms for specific problems in the simplest setting to avoid duplications with other dimensions of the analysis. Afterwards, we will identify the open problems for each dimension. By reading this survey, readers can gain a deeper understanding and intuition of the developments in the learning-augmented setting.

Acknowledgements

This proposal was prepared through discussions with Prof. Xiaoqi Tan and Prof. Csaba Szepesvári. I also would like to thank Shivam Garg, M Duc Nguyen, Haruto Tanaka, Jiamin He, Henry Du, and N Minh Pham for detailed feedback of this proposal.

References

Keerti Anand, Rong Ge, Amit Kumar, and Debmalya Panigrahi. 2022a. Online algorithms with multiple predictions.

Keerti Anand, Rong Ge, and Debmalya Panigrahi. 2022b. Customizing ml predictions for online algorithms.

Antonios Antoniadis, Joan Boyar, Marek Eliáš, Lene M. Favrholdt, Ruben Hoeksma, Kim S. Larsen, Adam Polak, and Bertrand Simon. 2022. Paging with succinct predictions.

Niv Buchbinder, Joseph Seffi Naor, et al. 2009. The design of competitive online algorithms via a primal–dual approach. *Foundations and Trends® in Theoretical Computer Science*, 3(2–3):93–263.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, Ali Vakilian, and Nikos Zarifis. 2021. Learning online algorithms with distributional advice. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2687–2696. PMLR, 18–24 Jul.

Marina Drygala, Sai Ganesh Nagarajan, and Ola Svensson. 2023. Online algorithms with costly predictions. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 8078–8101. PMLR, 25–27 Apr.

Paul Dütting, Silvio Lattanzi, Renato Paes Leme, and Sergei Vassilvitskii. 2020. Secretaries with advice.

Amos Fiat, Richard M Karp, Michael Luby, Lyle A McGeoch, Daniel D Sleator, and Neal E Young. 1991. Competitive paging algorithms. *Journal of Algorithms*, 12(4):685–699.

Sreenivas Gollapudi and Debmalya Panigrahi. 2019. Online algorithms for rent-or-buy with expert advice. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2319–2327. PMLR, 09–15 Jun.

Elena Grigorescu, Young-San Lin, Sandeep Silwal, Maoyuan Song, and Samson Zhou. 2022. Learning-augmented algorithms for online linear and semidefinite programming.

Elena Grigorescu, Young-San Lin, and Maoyuan Song. 2024. A simple learning-augmented algorithm for online packing with concave objectives.

Thodoris Lykouris and Sergei Vassilvitskii. 2020. Competitive caching with machine learned advice.

Michael Mitzenmacher and Sergei Vassilvitskii. 2020. Algorithms with predictions.