Caching with Calibrated Predictions

Helia Karisani
University of Massachusetts
Amherst, MA, USA

Mohammadreza Daneshvaramoli
University of Massachusetts
Amherst, MA, USA

Adam Lechowicz
University of Massachusetts
Amherst, MA, USA

Mingda Qiao
University of Massachusetts
Amherst, MA, USA

Mohammad Hajiesmaili
University of Massachusetts
Amherst, MA, USA

ABSTRACT

Learning-augmented online algorithms use predictions to improve performance beyond worst-case guarantees while preserving robustness to prediction errors. In caching, existing approaches typically rely on signals such as next-arrival times or ranking scores, whose quality is measured through aggregate or worst-case error notions. These notions lack explicit probabilistic semantics, making it difficult to relate prediction quality directly to algorithmic performance.

We study online caching with *calibrated probabilistic predictions*. Every cached page receives a succinct *phase prediction* in $[0, 1]$ representing the probability that it will *not* reappear in the next phase. Calibration ensures that these probabilities match empirical frequencies, providing a meaningful notion of predictor reliability. As a preliminary result, we give a threshold-based algorithm that labels a cached page as evictable whenever its phase prediction exceeds the closed-form threshold $T^\star = \frac{1}{1+H_k}$ induced by the asymmetric paging loss. Our main result bounds the expected competitive ratio in terms of the optimal cost and the quality of the predictions, with an additive calibration term that vanishes as the predictor approaches perfect calibration.

1. INTRODUCTION

The natural interplay between online algorithms and machine learning in decision-making under uncertainty has led to a growing body of work on *learning-augmented* algorithm design [5]. Broadly speaking, the goal is to design algorithms with strong *consistency* (good performance when predictions are accurate) while maintaining strong *robustness* (performance that does not degrade too much when predictions are inaccurate). However, most prior work treats predictions as a black box, relying on coarse or worst-case error measures.

A few works have explored richer prediction models that carry explicit reliability information [9, 2]. Most relevant to our work, Shen, Vitercik, and Wikum [8] showed that calibration guarantees translate directly into performance bounds for ski rental and non-clairvoyant scheduling—calibration requires that predicted probabilities match empirical frequencies (e.g., among all inputs predicted 0.7, roughly 70% are truly positive). In this preliminary work, we investigate how calibrated predictors can be used in the context of *caching*, a fundamental online problem with wide-ranging applications.

Learning-Augmented Algorithms: Theory and Applications (LATA) 2026
Ann Arbor, MI, USA

Problem. We consider a caching problem over a universe $\mathcal{U}$ of n pages with a cache of size $k < n$. Given a request sequence $\sigma = (r_1, \ldots, r_T)$, each cache miss incurs unit cost. We let $\mathrm{ALG}(\sigma)$ and $\mathrm{OPT}(\sigma)$ denote the costs of the online algorithm and the optimal offline algorithm, respectively, and define the competitive ratio as $\mathrm{CR} := \sup_{\forall \sigma} \frac{\mathbb{E}[\mathrm{ALG}(\sigma)]}{\mathrm{OPT}(\sigma)}$.

Phase decomposition. In the analysis of caching algorithms (such as the classic Marker algorithm [4]), it is standard to decompose the request sequence into *phases*. We partition any request sequence σ into phases $\Phi_1, \ldots, \Phi_m$, where Φ_1 is the shortest prefix of σ containing k distinct pages, and for $j > 1$, Φ_j is the shortest prefix of the remaining sequence containing exactly k distinct pages. This decomposition depends only on the request sequence, not on the algorithm. To preserve this algorithm-independence, we adopt the *marking rule*: whenever a new page is fetched into the cache during a phase, it is marked and cannot be evicted until the phase ends. This ensures that the algorithm performs at most k evictions per phase and that its eviction decisions do not alter the phase decomposition.

Calibrated predictions. To define a calibrated prediction model for caching, we adopt a notion of *phase predictions* introduced by Antoniadis et al. [1]. Under their original model, each page has a true binary $\{0, 1\}$ label such that a *1-page* will *not* be requested in the next phase (therefore, it should be evicted from the cache) and a *0-page will* reappear (therefore, it should be retained in the cache).

In our model, we relax the binary prediction to a probabilistic one. Each time page i is requested during phase j, the algorithm receives a prediction $p_{j,i} \in [0, 1]$ alongside the request and stores it with page i. The prediction represents the probability that page i is a 1-page, i.e., will *not* be requested in phase $j+1$. At the phase boundary, the algorithm reads the stored predictions for all pages currently in the cache (each of which was requested at least once in phase j, so every cached page has a fresh stored prediction) and can use them to compute the eviction threshold. Following [8], we let $Y_{j,i} \in \{0, 1\}$ indicate whether page i is truly a 1-page at phase boundary j. We define a predictor's calibration as follows: for a bin $B \subseteq [0, 1]$, we define $\mathrm{CalErr}(B) = \big|\mathbb{E}[Y \mid p \in B] - \mathbb{E}[p \mid p \in B]\big|$. The predictor is *perfectly calibrated* if $\mathrm{CalErr}(B) = 0$ for all B, meaning that the empirical frequencies of 1-pages match the predicted probabilities at every level set. It is *ε-calibrated* if $\sup_{B \subseteq [0,1]} \mathrm{CalErr}(B) \le \varepsilon$ (see [6] for a closely related ℓ_1 formulation). Compared to the binary $\{0, 1\}$ prediction model of [1], calibrated predictions carry explicit uncertainty semantics, allowing the algorithm to reason about expected

mistake rates for a given set of predictions.

Related work. Learning-augmented caching was introduced by Lykouris and Vassilvitskii [5] and has since been studied under several prediction models, including next-request times [7], costly predictions [3], and succinct predictions [1]. The closest prior work to ours is [1], where the predictor provides a binary label in $\{0, 1\}$ for each page, and the competitive guarantee is expressed in terms of the number of incorrect binary predictions. Our work generalizes that binary setting to a probabilistic one and incorporates calibration in the spirit of [8], which applied calibrated predictions to ski rental and other online problems.

2. ALGORITHM AND ANALYSIS

Algorithm 1 describes our threshold-based caching policy that receives a prediction $p_{j,i} \in [0, 1]$ online alongside each request for page i in phase j, storing it with the page. At the phase boundary, the algorithm reads the stored predictions of all k cached pages and labels each page as a *1-page* if $p_{j,i} \geq T^\star$ and as a *0-page* otherwise, where $T^\star = \frac{1}{1+H_k}$. Within the phase, a modified marking rule handles cache misses: pages are initially unmarked and become marked upon their first request; when an eviction is required, the algorithm evicts a uniformly random unmarked 1-page (if one exists), otherwise evicting a random unmarked 0-page.

Phase-local mistake objective. Fix a phase boundary $j{+}1$ with prediction vector $\vec{p}_j = (p_{j,1}, \ldots, p_{j,k})$. Using threshold $T \in [0, 1]$, page i is classified as a 1-page iff $p_{j,i} \geq T$. Two error types arise:

$$\mathbb{E}\Big[\eta_0^{(j)}\Big]\,(T, \vec{p}_j) = \sum_{i=1}^{k} \mathbf{1}[p_{j,i} < T]\; p_{j,i}, \tag{1}$$

$$\mathbb{E}\Big[\eta_1^{(j)}\Big]\,(T, \vec{p}_j) = \sum_{i=1}^{k} \mathbf{1}[p_{j,i} \geq T]\; (1 - p_{j,i}). \tag{2}$$

Here $\eta_0^{(j)}$ counts *incorrect 0-predictions*: pages labeled as 0-pages (predicted to reappear) that are truly 1-pages. These hold a cache slot unnecessarily; if evicted later to make room, the eviction falls on a true 0-page instead, causing up to H_k extra faults in phase $j + 1$ [1]. Alternatively, $\eta_1^{(j)}$ counts *incorrect 1-predictions*: pages labeled as 1-pages (predicted not to reappear) that are truly 0-pages. Under the modified marking rule these are evicted first on a cache miss, causing one extra fault when they are re-requested. This asymmetrical cost motivates an objective in terms of T:

$$f_j(T) := H_k\, \mathbb{E}\Big[\eta_0^{(j)}\Big]\,(T, \vec{p}_j) + \mathbb{E}\Big[\eta_1^{(j)}\Big]\,(T, \vec{p}_j). \tag{3}$$

Closed-form threshold. The objective (3) admits a closed-form minimizer. Page i is cheaper to label as a 1-page iff $1 - p_{j,i} \leq H_k p_{j,i}$, i.e., $p_{j,i} \geq \frac{1}{1+H_k}$.Since f_j decomposes over pages, an optimal decision boundary is the phase- and prediction-independent threshold $T^\star := \frac{1}{1+H_k}$.

From phase faults to a global bound. Let ALG_{j+1} denote faults in phase Φ_{j+1} and let c_j denote the number of new distinct pages entering the cache in phase j. The per-phase cost satisfies $\mathrm{ALG}_{j+1} \leq \mathrm{base}_j + H_k\, \eta_0^{(j)} + \eta_1^{(j)}$, where base_j is the unavoidable cost of phase Φ_{j+1} (at least one fault per new distinct page). Taking expectations, using $T^\star$, and summing over all phases with the standard paging bound $\sum_j \mathrm{base}_j \leq 2\,\mathrm{OPT}(\sigma)$ for marker-based algorithms:

$$\mathbb{E}[\mathrm{ALG}(\sigma)] \;\leq\; 2\,\mathrm{OPT}(\sigma) + \sum_j f_j(T^\star). \tag{4}$$

Each subsequent pair of phases (e.g., j and $j{+}1$) forces OPT to pay at least one fault per new distinct page in phase $j{+}1$, yielding the uniform bound (for marker-based algorithms) $\mathrm{OPT}(\sigma) \geq \frac{1}{2}\sum_j c_j$ [1]. Combining and dividing by $\mathrm{OPT}(\sigma)$:

$$\mathrm{CR} \;\leq\; 2 + \frac{\sum_{j=1}^{m} 2 f_j(T^\star)}{\sum_{j=1}^{m} c_j}. \tag{5}$$

Incorporating ε-calibration. Under ε-calibration, we bound the resulting phase-wise mistake objective by $f_j(T) \leq f_j^{ideal}(T) + \varepsilon(H_k + 1)k$, where f_j^{ideal} is the objective under perfect calibration. The additive per-phase penalty $\varepsilon(H_k + 1)k$ accumulates to $m\,\varepsilon(H_k + 1)k$ globally.

THEOREM 1. *Let $T^\star = \frac{1}{1+H_k}$. For a ε-calibrated predictor, the competitive ratio of Algorithm 1 satisfies*

$$\mathrm{CR} \;\leq\; 2 + \frac{\sum_{j=1}^{m} 2 f_j^{\mathrm{ideal}}(T^\star)}{\sum_{j=1}^{m} c_j} + O(\varepsilon H_k k). \tag{6}$$

In particular, as $\varepsilon \to 0$, the calibration penalty vanishes and the ratio is governed entirely by the phase-wise weighted mistake costs evaluated at the fixed threshold $T^\star$.

When predictions are both calibrated and sharp (concentrated away from $\frac{1}{2}$), $T^\star$ classifies most pages correctly and $f_j(T^\star) \ll H_k k$; thus, when both the calibration penalty and the phase-wise mistake terms are small, the bound is close to 2.

Algorithm 1 Threshold-Based Caching Algorithm

```
Initialize cache C with k arbitrary pages; label all as 0-pages; unmark all.
Set T* ← 1/(1+H_k).
for each phase Φ_j, j = 1, 2, ... do
    Phase boundary: unmark all pages in C.
    Label each i ∈ C: 1-page if p_{j,i} ≥ T*, else 0-page.
    for each request r_t in Φ_j do
        Receive prediction p_{j,r_t} ∈ [0, 1] with page r_t.
        if r_t ∈ C then cache hit; mark r_t.
        else cache miss:
            if ∃ unmarked 1-page in C then
                evict uniformly random unmarked 1-page.
            else
                evict uniformly random unmarked 0-page.
            Insert r_t into C; mark r_t; label r_t as 0-page.
```

3. OPEN QUESTIONS

We identify the following open questions for future work:

Question 1: Can we develop a general (not necessarily threshold-based) algorithm that attains a better (lower) competitive ratio?

Question 2: Can we derive a lower bound applying to all algorithms that characterizes the best-achievable competitive ratio for any algorithm given ε-calibrated predictions?

Question 3: Are there any natural online problems that similarly admit a clean model of calibrated predictions and a corresponding competitive analysis?

4. REFERENCES

[1] A. Antoniadis, J. Boyar, M. Eliáš, L. M. Favrholdt, R. Hoeksma, K. S. Larsen, A. Polak, and B. Simon. Paging with succinct predictions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 952–968. PMLR, 2023.

[2] V. Cohen-Addad, T. d'Orsi, A. Gupta, E. Lee, and D. Panigrahi. Max-cut with ϵ-accurate predictions, 2024.

[3] M. Drygala, S. G. Nagarajan, and O. Svensson. Online algorithms with costly predictions. In F. Ruiz, J. Dy, and J.-W. 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 2023.

[4] A. Fiat, R. M. Karp, M. Luby, L. A. McGeoch, D. D. Sleator, and N. E. Young. Competitive paging algorithms. *Journal of Algorithms*, 12(4):685–699, 1991.

[5] T. Lykouris and S. Vassilvitskii. Competitive caching with machine learned advice. *Journal of the ACM*, 68(4):1–25, 2021.

[6] M. Qiao and L. Zheng. On the distance from calibration in sequential prediction. In A. Cohen and E. Hazan, editors, *Proceedings of the Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4307–4357. PMLR, 2024.

[7] D. Rohatgi. Near-optimal bounds for online caching with machine learned advice. In *Proceedings of the 31st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1834–1845, 2020.

[8] J. H. Shen, E. Vitercik, and A. Wikum. Algorithms with calibrated machine learning predictions. In *Forty-second International Conference on Machine Learning*, 2025.

[9] B. Sun, J. Huang, N. Christianson, M. Hajiesmaili, A. Wierman, and R. Boutaba. Online algorithms with uncertainty-quantified predictions. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.