PAGING WITH SUCCINCT PREDICTIONS

Antonios Antoniadis · Joan Boyar · Marek Eliáš · Lene M. Favrholdt
Ruben Hoeksma · Kim S. Larsen · Adam Polak · Bertrand Simon

University of Twente · University of Southern Denmark · Bocconi · CNRS · CC IN2P3 · max planck institut informatik

The online paging problem

Which page to evict from cache to make room for a new page?

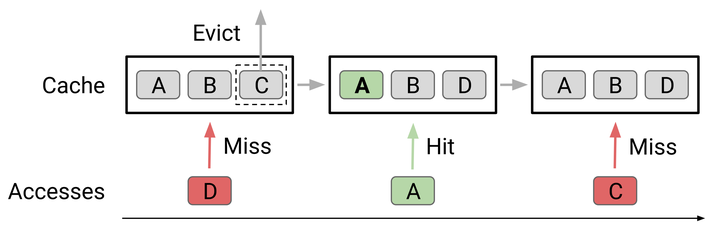

Image credit: Liu et al. *An Imitation Learning Approach for Cache Replacement* [ICML'20]

Goal: minimize number of **cache misses**.

requested page no longer in cache and has to be reloaded

k = cache size

Best classic algorithm: randomized $O(\log k)$-competitive [Fiat et al., '91]

$$\mathbb{E}[\mathrm{ALG}] \leqslant O(\log k) \cdot \mathrm{OPT} + \mathrm{const}$$

↑ cost of a best in hindsight choice of evictions

Matching lower bound: any algorithm is $\Omega(\log k)$-competitive.

Paging with predictions

You can bypass the $\log k$ barrier…
…if you have access to sufficiently accurate *predictions* about:

predicted information	bits per request	
time of reoccurrence of this page	$\log T$	[Lykouris, Vassilvitskii, ICML'18]
next action of OPT	$\log k$	[Antoniadis et al., ICML'20]
all requests until reoccurrence	$\log n$	[Jiang et al., ICALP'20]
relative order of reoccurrences	$\log k$	[Bansal et al., SODA'22]
if OPT evicts this page before reuse	**1**	**this work**
if this page appears in next phase	**1**	**this work**

Lower bound: **$o(1)$ bits** per request **do not suffice** to go below $\log k$, even with perfectly accurate prediction of any kind. [Mikkelsen, ICALP'16]

DISCARD-predictions setup

Would OPT evict this page before it is requested again?

Each page request r_i comes with *prediction* p_i:

$$p_i = \begin{cases} \mathbf{0}, & \text{if OPT } \textbf{keeps } r_i \text{ in cache until it is requested again,} \\ \mathbf{1}, & \text{if OPT } \textbf{evicts } r_i \text{ before it is requested again.} \end{cases}$$

There are available models trained to output such predictions, e.g.,:

- *Hawkeye* SVM [Jain, Lin, ISCA'16]
- *Glider* deep neural network with LSTM [Shi et al., MICRO'19]

Deterministic algorithm w/ DISCARD predictions

On each cache miss:

- evict a page that is predicted as safe to evict ($p_i = 1$), if it exists;
- otherwise, flush the cache, i.e., evict all pages (with $p_i = 0$).

$$\mathrm{ALG} \leqslant 1 \cdot \mathrm{OPT} + (k-1) \cdot \eta_0 + 1 \cdot \eta_1 + \mathrm{const}$$

Lower bound: coefficients 1, $(k-1)$, 1 cannot be improved (using a deterministic algorithm)

Randomized algorithm w/ DISCARD predictions

Immediately evict every page with prediction 1.
Perform a randomized marking strategy on pages with predictions 0.

intricate; see paper for details

$$\mathbb{E}[\mathrm{ALG}] \leqslant 1 \cdot \mathrm{OPT} + 2H_k \cdot \eta_0 + 1 \cdot \eta_1 + \mathrm{const}$$

↑ $H_k = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k} = O(\log k)$

Lower bound: coefficients 1, $2H_k$, 1 are optimal up to constant factors.

How do we measure prediction error?

η_0 = number of **zeros** that **should be ones**

η_1 = number of **ones** that **should be zeros**

PHASE-predictions setup

Is this page going to be requested in next phase?

maximal subsequence of k distinct pages

Each page request r_i comes with *prediction* p_i:

$$p_i = \begin{cases} \mathbf{0}, & \text{if } r_i \text{ is requested in next phase,} \\ \mathbf{1}, & \text{if } r_i \text{ is not requested in next phase.} \end{cases}$$

Algorithm with PHASE predictions

foreach request r_i **do**
  if r_i not in cache **then**
    if all pages in cache are marked **then**
      unmark all pages
    if there is an unmarked page with prediction 1 **then**
      evict a random unmarked page with prediction 1
    else
      evict a random unmarked page with prediction 0
  mark r_i

$$\mathbb{E}[\mathrm{ALG}] \leqslant 2 \cdot \mathrm{OPT} + H_k \cdot \eta_0 + 1 \cdot \eta_1 + \mathrm{const}$$
$$\mathbb{E}[\mathrm{ALG}] \leqslant O(\log(\eta_1/\mathrm{OPT})) \cdot \mathrm{OPT} + H_k \cdot \eta_0 + \mathrm{const}$$

Lower bounds:

- coefficient H_k is optimal up to an additive constant;
- asymptotic dependence on η_1/OPT cannot be improved.