Fundamental Limits on the Regret of Online Network-Caching

Rajarshi Bhattacharjee
Dept. of Electrical Engineering
Indian Institute of Technology Madras
brajarshi91@gmail.com

Subhankar Banerjee
Dept. of Electrical Engineering
Indian Institute of Technology Madras
ee16s048@ee.iitm.ac.in

Abhishek Sinha
Dept. of Electrical Engineering
Indian Institute of Technology Madras
abhishek.sinha@ee.iitm.ac.in

ABSTRACT

Optimal caching of files in a content distribution network (CDN) is a problem of fundamental and growing commercial interest. Although many different caching algorithms are in use today, the fundamental performance limits of the network caching algorithms from an online learning point-of-view remain poorly understood to date. In this paper, we resolve this question in the following two settings: (1) a single user connected to a single cache, and (2) a set of users and a set of caches interconnected through a bipartite network. Recently, an online gradient-based *coded* caching policy was shown to enjoy sub-linear regret. However, due to the lack of known regret lower bounds, the question of the optimality of the proposed policy was left open. In this paper, we settle this question by deriving tight non-asymptotic regret lower bounds in the above settings. In addition to that, we propose a new Follow-the-Perturbed-Leader-based *uncoded* caching policy with near-optimal regret. Technically, the lower-bounds are obtained by relating the online caching problem to the classic probabilistic paradigm of *balls-into-bins*. Our proofs make extensive use of a new result on the expected load in the most populated half of the bins, which might also be of independent interest. We evaluate the performance of the caching policies by experimenting with the popular MovieLens dataset and conclude the paper with design recommendations and a list of open problems.

CCS CONCEPTS

• Networks → Network performance analysis; **• Mathematics of computing → Probabilistic algorithms**.

KEYWORDS

network-caching; online algorithms; regret bounds; fundamental limits

ACM Reference Format:
Rajarshi Bhattacharjee, Subhankar Banerjee, and Abhishek Sinha. 2020. Fundamental Limits on the Regret of Online Network-Caching. In *ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '20 Abstracts), June 8–12, 2020, Boston, MA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3393691.3394189

1 INTRODUCTION

The classical caching problem, which seeks to make popular content quickly accessible by prefetching them in a low-latency storage, has been extensively studied in the literature. Among the online caching policies, the Least Recently Used policy (LRU), the Least Frequently Used policy (LFU), the FIFO policy, and online coded caching policies have been studied extensively. However, the performance guarantees available for most of these policies are highly contingent upon some a priori assumptions on the generative model of the file request sequence. With frequent addition of new content to the library, mobility of the users, Femtocaching with small caches, and change in the popularity distribution with time, the assumption of stationary file popularity barely holds in practice. This prompts us to consider the problem of caching from an online learning point-of-view with no a priori statistical assumptions on the file request sequence. Our work is inspired by the recent paper [2], which describes an online gradient-based coded caching policy (OGA), and proves a sub-linear regret upper-bound for the same. The authors also showed that popular uncoded caching policies, such as LRU, LFU, and FIFO, suffer from linear regrets. In fact, no uncoded caching policy with a sub-linear regret is known previously in the literature.

We formulate the network-caching problem as an online convex optimization problem with a piecewise-linear reward function and polytope constraints. Though a few results are known for the regret lower bounds for online convex optimization problems [1], to the best of our knowledge, except for [2], the regret for a linear cost function with simplex and several box constraints, which feature in the single-cache problem, has not been studied before. Moreover, the problem of deriving a tight lower bound on the regret for a piecewise-linear cost function with polytope constraints, which arise in the context network-caching, is entirely open. The above considerations prompt us to ask the following two questions:
Question 1. What is the fundamental performance limit of all online caching policies *regardless* of their operational constraints or computational complexity?
Question 2. Can a simple, distributed network-caching strategy be designed which meets the above fundamental limit?
In this paper, we derive universal regret lower bounds answering Question 1. Moreover, we answer Question 2 in the affirmative by proposing a new Follow-the-Perturbed-Leader-based uncoded caching policy that has near-optimal regret.

2 SYSTEM MODEL

2.1 Model for a Single Cache

We consider a library of N distinct files. Time is slotted and a cache with a limited storage capacity can store at most C files at any time slot. A user may request at most one file at a slot. The file

SIGMETRICS '20 Abstracts, June 8–12, 2020, Boston, MA, USA

ACM ISBN 978-1-4503-7985-4/20/06.
https://doi.org/10.1145/3393691.3394189

request at time slot t is represented by an N-dimensional one-hot encoded binary vector $\boldsymbol{x}_t \in \{0,1\}^N$, Files can be either uncoded, or encoded with rateless erasure codes [3]. Following an online caching policy, files are cached at every time slot *before* the request for that slot arrives. We do not make *any* statistical assumptions on the file request sequence $\{\boldsymbol{x}_t\}_{t\geq 1}$. The cache configuration at time t is represented by the vector $\boldsymbol{y}_t \in [0,1]^N$, where $y_{t,f}$ denotes the fraction of the file f cached at time t. Naturally, in the uncoded case $y_{t,f} \in \{0,1\}, \forall f, t$. The set of all admissible caching configuration is given by $\mathcal{Y} = \{\boldsymbol{y} \in [0,1]^N : \sum_{f=1}^N y_f \leq C\}$. The total reward accrued by a caching policy up to time T is given by: $Q(\{\boldsymbol{x}_t\}_1^T, \{\boldsymbol{y}_t\}_1^T) = \sum_{t=1}^T q(\boldsymbol{x}_t, \boldsymbol{y}_t)$. Here, $q(\boldsymbol{x}_t, \boldsymbol{y}_t)$ is the slot wise hit-rate, given by: $q(\boldsymbol{x}_t, \boldsymbol{y}_t) \equiv \boldsymbol{x}_t \cdot \boldsymbol{y}_t$. We define the regret $R_T^\pi(\{\boldsymbol{x}_t\}_1^T)$ for a request sequence $\{\boldsymbol{x}_t\}_1^T$ to be the difference in the reward obtained by the best static caching configuration $\boldsymbol{y}^*$ and that of the online policy π:

$$R_T^\pi(\{\boldsymbol{x}_t\}_1^T) := \sup_{\boldsymbol{y}^* \in \mathcal{Y}} \Big(Q(\{\boldsymbol{x}_t\}_1^T, \{\boldsymbol{y}^*\}_1^T) - Q(\{\boldsymbol{x}_t\}_1^T, \{\boldsymbol{y}_t\}_1^T)\Big). \quad (1)$$

The regret R_T^π of any caching policy π up to time T is defined to be the maximum regret over all admissible request sequences, *i.e.,* $R_T^\pi := \sup_{\{\boldsymbol{x}_t\}_1^T} R_T^\pi(\{\boldsymbol{x}_t\}_1^T)$.

2.2 Model for a Bipartite Caching Network

In a bipartite Content Distribution Network (CDN), a set of users $\mathcal{I} = \{1, 2, \ldots, I\}$ is connected to a set of caches $\mathcal{J} = \{1, 2, \ldots, J\}$ in the form of a bipartite graph $(\mathcal{I}, \mathcal{J}, E)$. The set of caches connected to a user $i \in \mathcal{I}$ is denoted by $\partial^+(i) \equiv \{j \in \mathcal{J} : (i,j) \in E\}$. Similarly, the set of users connected to a cache $j \in \mathcal{J}$ is denoted by $\partial^-(j) \equiv \{i \in \mathcal{I} : (i,j) \in E\}$. The *in-degree* of the cache j is defined as $d_j \equiv |\partial^-(j)|, j \in \mathcal{J}$. For simplicity, assume that $d_j = d, \forall j \in \mathcal{J}$, and that each cache has equal capacity C. A user i may request at most one file (represented by the one-hot encoded vector $\boldsymbol{x}_t^i, \forall i$) at slot t. Each file-request may be served by any (one or more) neighboring caches. The configuration of the j^{th} cache at slot t is denoted by the vector $\boldsymbol{y}_t^j, \forall j$. We consider caching two types of content- (1) *Elastic contents* (*e.g.,* multi-resolution HD videos) and (2) *Inelastic contents* (*e.g.,* databases, documents). Receiving multiple layers (resolutions) of an elastic content improves its overall utility. Consequently, the reward function for elastic contents is defined as $q_{\text{elastic}}(\boldsymbol{x}_t, \boldsymbol{y}_t) \equiv \sum_{i\in\mathcal{I}} \boldsymbol{x}_t^i \cdot (\sum_{j\in\partial^+(i)} \boldsymbol{y}_t^j)$. On the other hand, receiving multiple copies of an *Inelastic Content* does not increase its utility (single copy suffices for full decoding). Hence, the reward function for inelastic contents is defined as $q_{\text{inelastic}}(\boldsymbol{x}_t, \boldsymbol{y}_t) \equiv \sum_{i\in\mathcal{I}} \boldsymbol{x}_t^i \cdot \min\{\boldsymbol{1}, (\sum_{j\in\partial^+(i)} \boldsymbol{y}_t^j)\}$.

Algorithm 1 FTPL policy for Single Cache

```
count ← 0, $\eta \leftarrow \frac{1}{(4\pi \log N)^{1/4}}\sqrt{\frac{T}{C}}$
for $t = 1$ to $T$ do
    count ← count + $\boldsymbol{x}_t$
    Sample $\boldsymbol{\gamma}_t \sim \mathcal{N}(0, \boldsymbol{I}_{N\times N})$
    perturbed count ← count + $\eta\boldsymbol{\gamma}_t$
    Sort perturbed count in decreasing order and load the top
    $C$ files in the cache.
end for
```

3 KEY CONTRIBUTIONS

We make the following key theoretical contributions in the paper:

(1) Lower bound for a single cache: The regret R_T^π of any online caching policy π in the single cache setting, for a library size N and cache capacity C with $N \geq 2C$, is lower bounded as

$$R_T^\pi \geq \sqrt{\frac{CT}{2\pi}} - \Theta(\frac{1}{\sqrt{T}}), \ \forall\, T \geq 1. \quad (2)$$

The above result improves upon the previously known *asymptotic* lower bound in [2] by removing the dependence on N from the regret bound.

(2) Lower bounds for caching networks: For caching *elastic* contents in a bipartite network in the above set up with $N \geq 2C$, the regret of any online caching policy π is lower bounded as:

$$R_T^\pi \geq d|\mathcal{J}|\sqrt{\frac{CT}{2\pi}} - \Theta(\frac{1}{\sqrt{T}}), \ \forall T \geq 1. \quad (3)$$

For caching *inelastic* contents in a bipartite network in the above set up with $N \geq 2C|\mathcal{J}|$, the regret of any online caching policy π is lower bounded as:

$$R_T^\pi \geq d\sqrt{\frac{|\mathcal{J}|CT}{2\pi}} - \Theta(\frac{1}{\sqrt{T}}), \ \forall T \geq 1. \quad (4)$$

(3) Near-optimal caching policies: In Algorithm 1, we propose a simple uncoded caching policy based on the Follow-the-Perturbed-Leader (FTPL) paradigm in online learning. In the single cache setting, the FTPL uncoded caching policy achieves the following upper bound on the expected regret, where the expectation is taken over the intrinsic randomness of the algorithm:

$$\mathbb{E}_{\{\boldsymbol{\gamma}_t\}_{t\geq 1}}(R_T^{\mathsf{FTPL}}) \leq 1.51(\log N)^{1/4}\sqrt{CT}. \quad (5)$$

Eqns. (2) and (5) show that FTPL is regret-optimal up to a poly-log factor. We also prove that the lower bounds in Eqns. (3) and (4) are tight up to a constant factor and up to a factor of $O(\sqrt{|\mathcal{J}|})$ resp. The OGA policy [2] achieves the corresponding upper bounds. A version of the FTPL policy is also shown to be regret-optimal up to a constant factor in the bipartite cache setting for elastic contents.

(4) New proof techniques: We establish the lower bounds in Eqns. (2), (3), and (4) by relating the online caching problem to the classic probabilistic setup of *balls-into-bins*. In particular, we derive a new non-asymptotic lower bound to the expected total load in the *most populated n* bins when m balls are randomly thrown into $2n$ bins. This is the first paper where a connection between online learning and the framework of balls-into-bins has been explicitly brought out and exploited in proving regret lower bounds.

4 ACKNOWLEDGEMENT

This work is supported by the Qualcomm grant IND-417880.

REFERENCES

[1] Elad Hazan and Sanjeev Arora. 2006. *Efficient algorithms for online convex optimization and their applications.* Princeton University Princeton.

[2] Georgios S Paschos, Apostolos Destounis, Luigi Vigneri, and George Iosifidis. 2019. Learning to Cache With No Regrets. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications.* IEEE, 235–243.

[3] Amin Shokrollahi. 2006. Raptor codes. *IEEE/ACM Transactions on Networking (TON)* 14, SI (2006), 2551–2567.