Why (and When) do Asymptotic Methods Work so well?

Nicolas Gast

Queueing Systems Vol. 100
Special Issue: 100 views on queues

Why (and When) do Asymptotic Methods Work so well?

Nicolas Gast

February 21, 2022

1 Introduction

Asymptotic methods are widely used in queueing systems to obtain very accurate approximations of various performance metrics. Typical examples are load-balancing systems, where mean-field approximations provide very accurate estimates of response time [11]; cache replacement policies, where TTL or fixed point approximations provide extremely accurate estimators of the miss rate [4,5]; or scheduling problems where asymptotic relaxations such as Whittle index [14] provide simple control policies that are very close to the optimal policies.

These approximations are often justified by studying the asymptotic behavior of the model when its size grows [1]. For instance, in a load balancing system, if n is the number of servers, and X_i^n is the fraction of servers in state i in steady-state, then it can be shown that, for many load-balancing policies, X_i^n is close to a deterministic value x_i^n as n goes to infinity. The mean field approximation is to then write:

$$\mathbb{E}[f(X^n)] \approx f(x^n), \tag{1}$$

where f is our performance metric: Here $f(x) = \sum_i i x_i$ is the average queue length.

The same method is used for caches where TTL-approximations are shown to be asymptotically exact as the cache size grows [6], or for scheduling policies where Whittle index is shown to provide asymptotically optimal resource allocations [13]. Yet, we argue that, if Equation (1) guarantees that the approximation is asymptotically exact, it does not explain why for small system sizes, the approximation might be extremely accurate for some models, like caching [4], and much less for others, like some load-balancing policies [12].

In this document, we discuss two models for which the approximation works much better than expected, and provide ideas and challenges to study why.

Nicolas Gast
Univ. Grenoble Alpes, Inria, Grenoble, France. E-mail: nicolas.gast@inria.fr

Table 1 Exact hit probability v.s. TTL-approximation for $n = 6$ objects and a cache of size $C = 3$.

λ_k	6	5	4	3	2	1
Exact π_k^n	0.738	0.674	0.589	0.478	0.341	0.181
TTL-approximation $\pi_k^{n,\mathrm{TTL}}$	0.729	0.663	0.581	0.479	0.353	0.195

2 Problems Statements

Caches and TTL-approximation. Consider a system with n items and a cache of size C^n. Object k is requested according to a Poisson process of intensity λ_k. Assume that the cache is managed according to LRU (Least Recently Used), *i.e.*, the C^n most recently requested objects are stored in the cache. We want to approximate π_k^n, the probability that object k is in this cache of C^n objects. Let τ^n be a (deterministic) duration such that the expected number of objects requested in the last τ^n seconds equals C^n, that is: $\sum_{k=1}^n (1 - e^{-\lambda_k \tau^n}) = C^n$. The TTL-approximation [5] is to do as if all objects were removed from the cache after τ^n seconds. This translates in $\pi_k^n \approx \pi_k^{n,\mathrm{TTL}} := 1 - e^{-\lambda_k \tau^n}$.

This approximation is shown to be asymptotically exact as n goes to infinity [6, 10]. The approach to show this result is as follows. Let X_k be the time since the last request of object k and let $X_{(C^n)}$ be C^nth smallest value of X_k. It is shown in [6] that $X_{(C^n)}$ concentrates on τ_n as n grows (at rate $O(1/\sqrt{n})$). In practice, however, the TTL-approximation is almost exact even for very small values of n, as illustrated in Table 1. The asymptotic analysis does not explain the accuracy of the approximation as the error bound derived in [6, 10] is far from the true error.

Open question Can we compute a sharp, non-asymptotic, error bound on the error of the TTL-approximation, $\sum_{k=1}^n |\pi_k^{\mathrm{TTL}} - \pi_k|$?

Whittle-index Restless bandits are optimization problems composed of n independent "arms". At each decision epoch, the decision maker chooses which arms to activate and receives a reward in exchange. Each arm then makes a Markovian transition that is a function of the actions sent by the decision maker. One typical application of restless bandits is for scheduling problems [9]. The arms are the tasks to execute and activating an arm corresponds to processing a task.

While these problems are computationally hard, Whittle introduced in [14] a notion of index – now known as Whittle index – and that generalizes the notion of Gittins index to restless bandits. Contrary to Gittins, the Whittle index policy is in general not optimal for a given system size n. Yet, under mild conditions, this policy is shown in [13] to be asymptotically optimal as n goes to infinity: If V_*^n is the maximal average reward that a decision maker can obtain for the n-arm problem and V_{WIP}^n the one of the Whittle index policy, then $|V_*^n - V_{WIP}^n| \to 0$. To show that, the authors [13] show that the dynamics of the Whittle index policy converges to a deterministic dynamics as n grows. While no rate of convergence is given in [13], many examples indicate that the Whittle index policy is almost optimal in practice, even for a small number of arms [9].

Open question Can we obtain a non-asymptotic error bound on $V_*^n - V_{WIP}^n$ that would explain why this policy works so well for very small n?

3 Discussion, challenges and opportunities

The natural approach to study the accuracy of Equation (1) is to study the distance between the stochastic system X^N and its deterministic counterpart x^n, but this often leads to bounds in $O(1/\sqrt{n})$ that are not sharp enough. Studying the distance between $\mathbb{E}[f(X^n)]$ and $f(x^n)$ provides sharper bounds but is in general more challenging.

A natural direction to pursue is to build on the recent developments of Stein's method, that can provide non-asymptotic bounds. This approach leads to a good understanding of the error of mean field methods for smooth Markovian systems composed of homogeneous entities, and can also be used to propose refinements [8]. Yet, applying this method to the above problem requires to lift a number of difficulties: non-homogeneity of objects and non-Markovian evolution of $X_{(C^n)}$ for LRU; non-smooth dynamics for the Whittle index policy.

Solving these challenges might prove to be quite difficult but has potentially important impacts in guiding when to use these approximations: What is special to a LRU cache or to Whittle index that makes the asymptotic methods so accurate? Why are these methods less accurate in other contexts? Could higher-order diffusion approximations [2] be useful here? Answers to such questions might also explain why the mean field approximation is so accurate for the multi-scale models of [3], or why the pair-approximation seems almost exact for models with a spatial component [7].

References

1. M. Benaïm and J.-Y. Le Boudec. A class of mean field interaction models for computer and communication systems. *Performance Evaluation*, 65(11):823–838, Nov. 2008.
2. A. Braverman, J. Dai, and X. Fang. High order steady-state diffusion approximations. *arXiv preprint arXiv:2012.02824*, 2020.
3. F. Cecchi, S. C. Borst, and J. van Leeuwaarden. Mean-field analysis of ultra-dense CSMA networks. *ACM SIGMETRICS Performance Evaluation Review*, 43(2):13–15, 2015.
4. A. Dan and D. Towsley. An approximate analysis of the LRU and FIFO buffer replacement schemes. In *Proceedings of the 1990 ACM SIGMETRICS conference*, pages 143–152, 1990.
5. R. Fagin. Asymptotic miss ratios over independent references. *Journal of Computer and System Sciences*, 14(2):222–250, 1977.
6. C. Fricker, P. Robert, and J. Roberts. A versatile and accurate approximation for LRU cache performance. In *2012 24th International Teletraffic Congress (ITC 24)*, pages 1–8. IEEE, 2012.
7. N. Gast. The power of two choices on graphs: the pair-approximation is accurate. *ACM SIGMETRICS Performance Evaluation Review*, 43(2):69–71, 2015.
8. N. Gast and B. Van Houdt. A refined mean field approximation. *Proc. ACM Meas. Anal. Comput. Syst.*, 1(2):33:1–33:28, Dec. 2017.
9. Y.-P. Hsu, E. Modiano, and L. Duan. Scheduling algorithms for minimizing age of information in wireless broadcast networks with random arrivals. *IEEE Transactions on Mobile Computing*, 19(12):2903–2915, 2019.
10. B. Jiang, P. Nain, and D. Towsley. On the convergence of the TTL approximation for an LRU cache under independent stationary request processes. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS)*, 3(4):1–31, 2018.
11. M. Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems*, 12(10):1094–1104, 2001.
12. D. Mukherjee, S. Borst, J. Van Leeuwaarden, and P. Whiting. Universality of power-of-d load balancing schemes. *ACM SIGMETRICS Performance Evaluation Review*, 44(2):36–38, 2016.
13. R. R. Weber and G. Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990.
14. P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25(A):287–298, 1988.