Modeling and Performance Evaluation for the Least Recently Used Cache Eviction Policy

Modern Problems in Theoretical and Applied Probability
85th Anniversary of Alexander Borovkov
Novosibirsk State University

Carl Graham[1] (speaker)
Felipe Olmos[2] Alain Simonian[2]

[1]École Polytechnique, CNRS - [2]OrangeLabs

August 23, 2016

Cache Servers for Efficient Network Use

The efficient operation of caches is a major industrial concern.

Caches play a major role in Content Delivery Networks (CDN) which are a key component of today's Internet, as well as in the emergent Information Centric Networking (ICN) architecture.

Cache Servers for Efficient Network Use

A network which delivers content (such as videos)
proposes to the users
a catalog of documents which evolves in time.

The documents are all stored in a central server.

In order to save user time and network resources,
cache servers are placed close to the users
to store a fraction of these documents.

Caches are managed using real-time distributed
algorithms called cache eviction policies.

Cache Policy Performance Indicators

Upon any user request for some document in the catalog, one of the two following events may happen.

Hit: The document subject to the request is already stored in the cache.

Miss: The document is not in cache.

Their complementary probabilities are called

the hit probability, and
the miss probability.

These constitute important quality of service (QoS) indicators.

Cache Policy Performance Evaluation

We present here the essentials of the study,[1]

part of the PhD thesis of Felipe Olmos.

This interdisciplinary PhD in mathematics and computer science took place between industry and academia

at OrangeLabs and École polytechnique.

The aim was to use stochastic modeling and analysis, in order to exploit the proprietary Internet traces of Orange using statistics and data processing. The models were calibrated and validated on real and simulated data.

---

[1] Felipe Olmos, Carl Graham, and Alain Simonian (2015). "Cache miss estimation for non-stationary request processes". In: *ArXiv:1511.07392*.

LRU Performance Evaluation

This study evaluated the miss probability for the Least Recently Used (LRU) cache eviction policy.

The cache is modeled by a list of $C \geq 1$ documents which are stored in C consecutive slots.

The LRU policy is based only on hits and misses and acts at each user request.

LRU Cache Policy: Hits

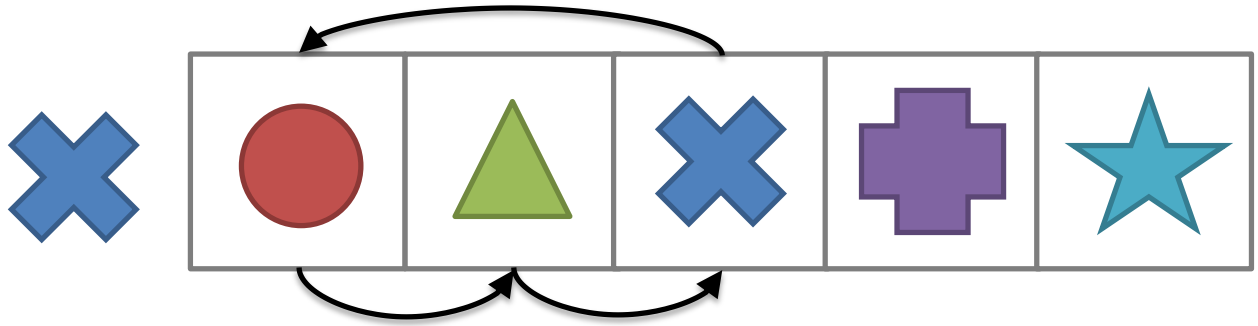
If the request is a:

- **Hit: The document is already in cache.**

Then the cache:

- Uploads this document to the user.
- Moves this document to the front of the list.
- Shifts down one slot all documents that were in front of this document in the cache.

Hit Request



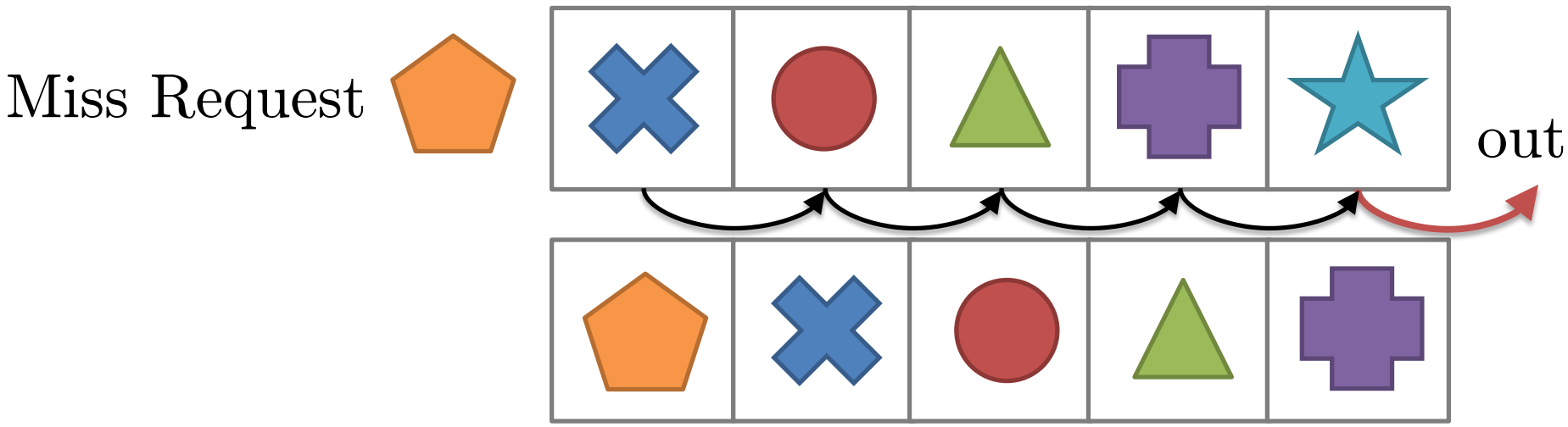
LRU Cache Policy: Misses

If the request is a:

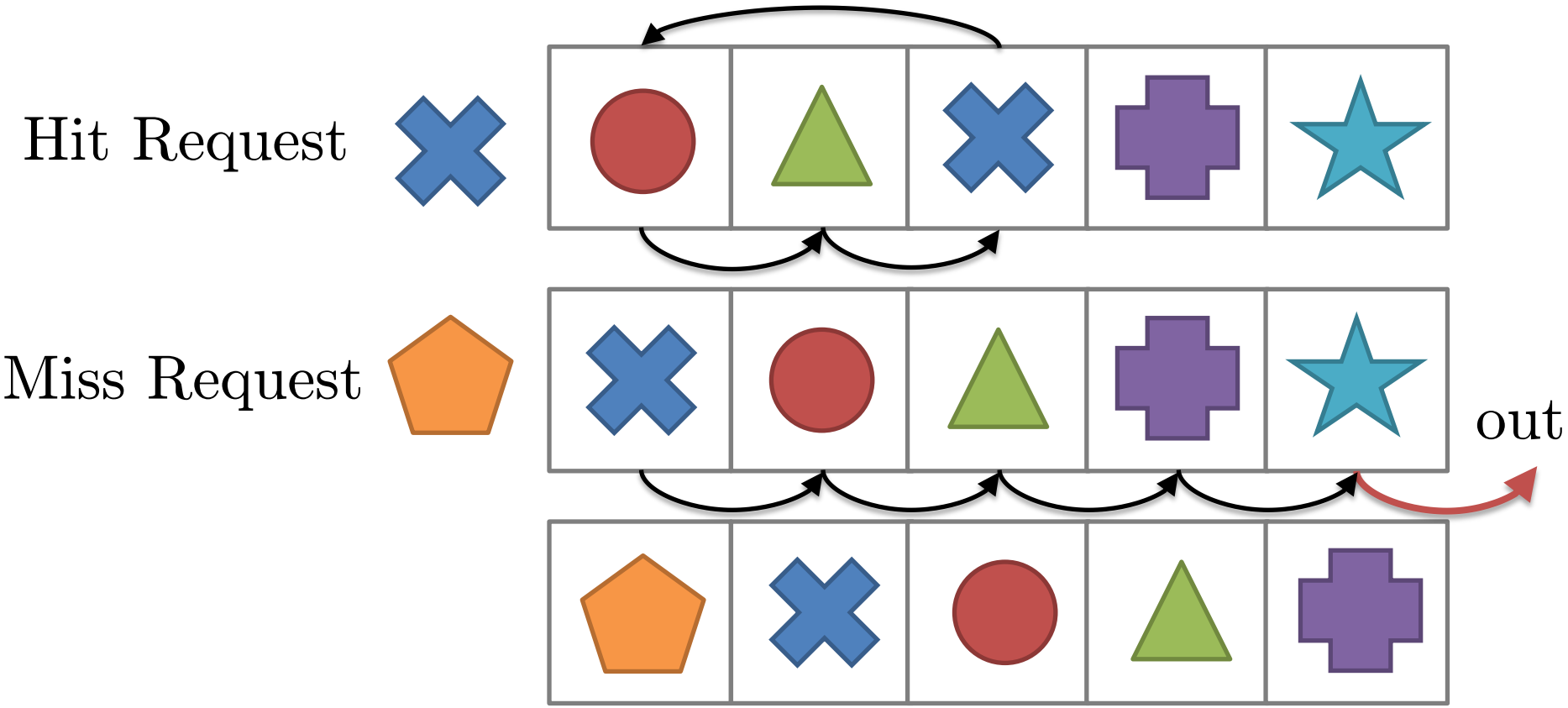
- **Miss: The document is not in cache.**

Then the cache:

- Must download this document from the central server.
- Uploads this document to the user.
- Places this document at the front of the list.
- Shifts down one slot all documents in cache except the last one which is eliminated.



LRU Cache Policy



Catalog Arrivals and User Requests

The document arrivals and the user requests are modeled by a

Poisson cluster point process

that has recently been independently proposed and studied heuristically in[2] and in.[3]

This point process on $\mathbb{R}$ is defined as follows.

---

[2] Stefano Traverso et al. (2013). "Temporal locality in today's content caching: why it matters and how to model it". In: *ACM SIGCOMM Computer Communication Review* 5.

[3] Felipe Olmos, Bruno Kauffmann, et al. (2014). "Catalog Dynamics: Impact of Content Publishing and Perishing on the Performance of a LRU cache". In: *26th International Teletraffic Congress (ITC)*. IEEE.

Poisson Cluster Point Process

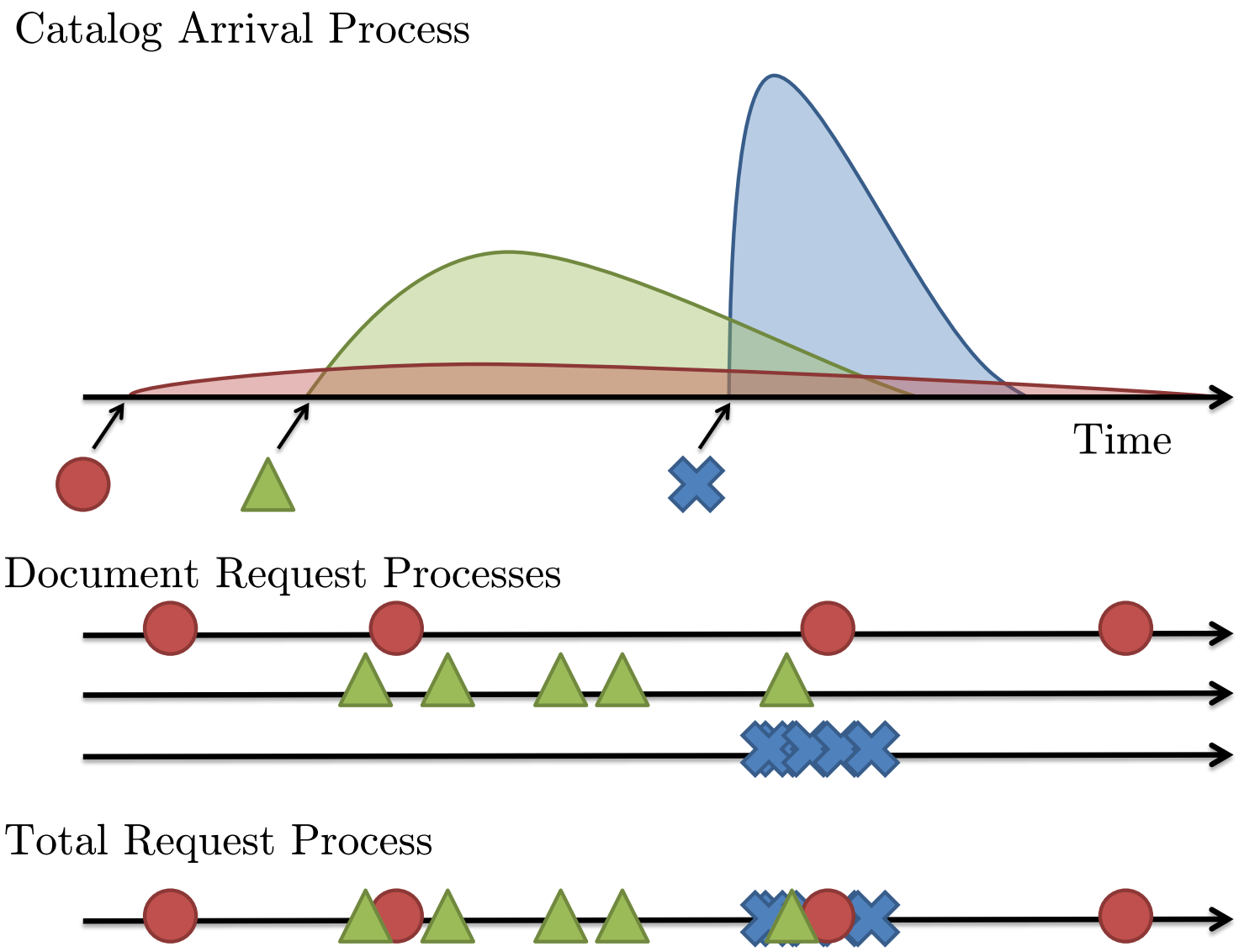
- The arrival of new documents in the catalog follows a Poisson process Γ^g on $\mathbb{R}$ of rate $\gamma > 0$.
- Each such arrival at an instant a of Γ^g triggers a user request process ξ_a which is a Cox process with random intensity function (popularity)

$$\lambda_a : t \in \mathbb{R} \mapsto \lambda_a(t) \in \mathbb{R}_+ .$$

- Given Γ^g, the λ_a for a in Γ^g are independent.
- The λ_a are causal: if $t < a$ then $\lambda_a(t) = 0$.
- The λ_a are stationary: for a canonical intensity function λ corresponding to an arrival at time 0,

$$\lambda_a(\bullet) \overset{\text{law}}{=} \lambda(\bullet - a) .$$

Catalog Arrival and User Request Model



Mean Function, Complementary Mean Function

Let a be a generic arrival instant of Γ^g.
The mean function of ξ_a is given by

$$\Lambda_a(t) = \int_{-\infty}^{t} \lambda_a(s)\,\mathrm{d}s = \int_{a}^{t} \lambda_a(s)\,\mathrm{d}s\,, \quad t \in \mathbb{R}\,.$$

We assume that $\Lambda_a(\infty) < \infty$, a.s., and define
the complementary mean function

$$\overline{\Lambda}_a(t) = \Lambda_a(\infty) - \Lambda_a(t) = \int_{t}^{\infty} \lambda_a(s)\,\mathrm{d}s\,, \quad t \in \mathbb{R}\,.$$

We denote by λ, Λ and $\overline{\Lambda}$ a generic instance of these processes corresponding to an arrival at time 0.

Poisson Point Process, Request Process

Let $\mathscr{M}^{\#}(\mathbb{R})$ denote the space of point processes on $\mathbb{R}$.
The document arrival process Γ^g marked with the document request processes ξ_a constitutes a Poisson point process

$$\widetilde{\Gamma} = \sum_{a \in \Gamma^g} \delta_{(a, \xi_a)} \quad \text{on} \quad \mathbb{R} \times \mathscr{M}^{\#}(\mathbb{R}) \,.$$

The total request process by the users is the superposition

$$\Gamma = \sum_{a \in \Gamma^g} \xi_a \,.$$

We assume that the expected numbers of points of Γ on each $[s, t]$ is finite:

$$\gamma \int_{-\infty}^{t} \mathbb{E}\left[\, 1 - \mathrm{e}^{-(\Lambda_a(t) - \Lambda_a(s))} \,\right] \mathrm{d}a < \infty, \quad \forall -\infty < s < t \leq +\infty \,.$$

This is necessary and sufficient for Γ to be locally finite, a.s.

Point of View of a Document: Palm theory

For this

Poisson cluster point process request model,

we use Palm theory in order to construct a probability space on which we may analyze mathematically

a tagged document

w.r.t.

the rest of the request process.

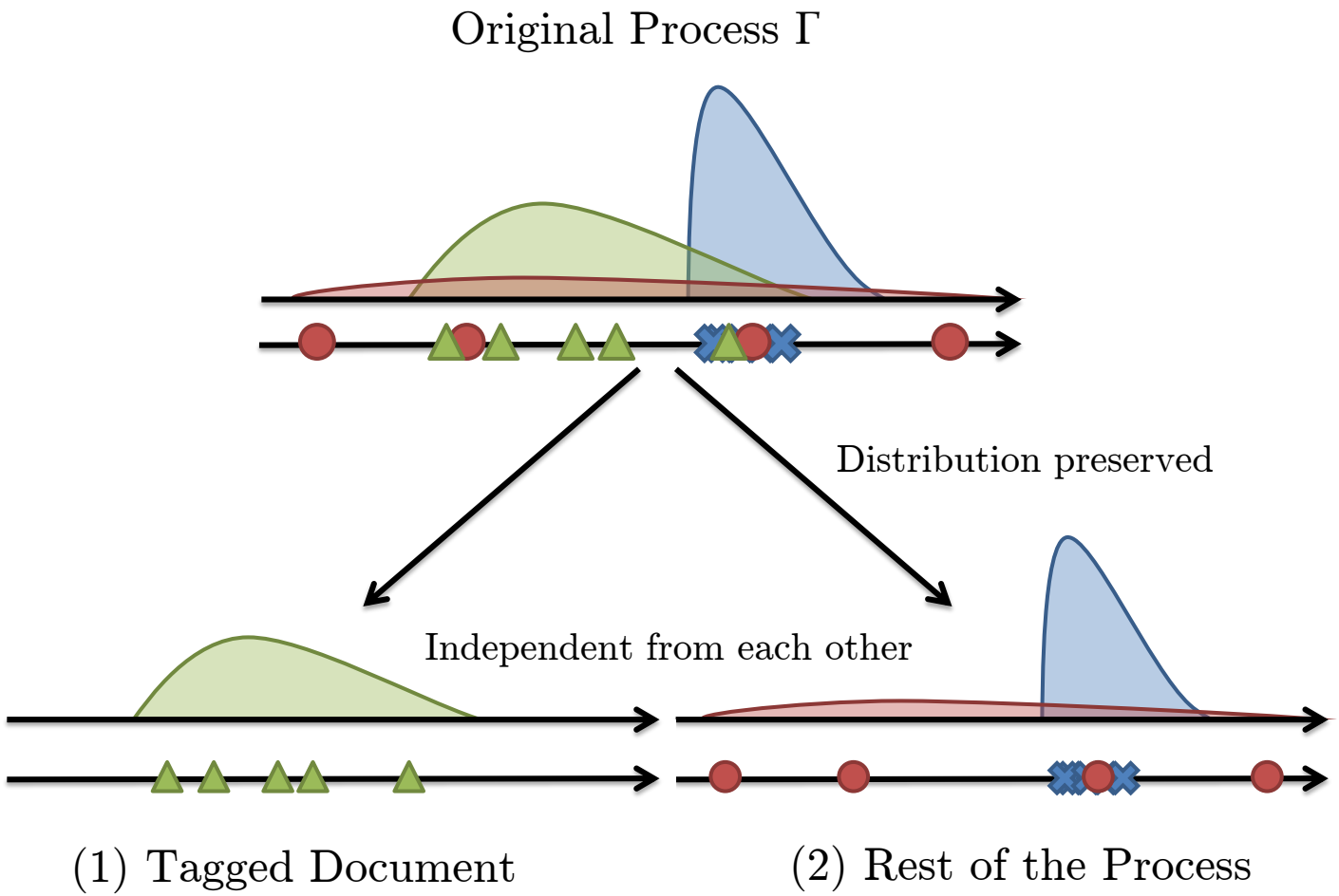
A simple decomposition follows from the fact that

$\widetilde{\Gamma}$ is a Poisson point process.

Point of View of a Document: Decomposition

The conditioned process Γ is decomposed independently into:

- The tagged document. Its arrival time is assumed to be 0.
- The rest of the request process, which has same law as Γ.



Miss Probability

We consider a LRU cache of size $C \geq 1$,
and a tagged document arriving at time 0 in the catalog.

Let N be the number of requests of this document, and μ_C the number of misses. The miss probability is defined by

$$p_C = \frac{\mathbb{E}[\,\mu_C\,]}{\mathbb{E}[\,N\,]}\,.$$

This is also the average per-document miss-ratio under the size-biased distribution of N.

Since

$$\mathbb{E}[\,N\,] = \mathbb{E}[\,\mathbb{E}[\,N \mid \Lambda\,]\,] = \mathbb{E}[\,\Lambda\,],$$

it is left to study μ_C and then $\mathbb{E}[\,\mu_C\,]$.

Number of Misses

Let $(\Theta_r)_{r=1}^N$ be the request times for the tagged document. Then

$$\mu_C = \mathbb{1}\{ N \geq 1 \} + \mathbb{1}\{ N \geq 2 \} \sum_{r=2}^{N} \mathbb{1}\{ \text{request at time } \Theta_r \text{ is a miss} \}.$$

Under the LRU policy, a document requested at time s
will next exit the cache at the

first time after s
that C distinct other documents
have been requested without interruption
by a request for this first document.

The Distinct Document Counting Process I

For s in $\mathbb{R}$ let $X^s := (X_t^s)_{t \geq s}$ be defined by

$$X_t^s = \#\{ \text{Distinct documents in the rest of the process on } [s, t] \}.$$

If $F^s(\xi_a)$ denotes the first arrival time of ξ_a in $[s, +\infty)$, then

$$X_t^s = \#\{ (a, \xi_a) \in \widetilde{\Gamma} - \{(0, \xi_0)\} : F^s(\xi_a) \leq t \}.$$

The first hitting time of level C by X^s is defined by

$$T_C^s = \inf\{ t \geq s : X_t^s = C \},$$

and

$$\mu_C = \mathbb{1}\{ N \geq 1 \} + \mathbb{1}\{ N \geq 2 \} \sum_{r=2}^{N} \mathbb{1}\{ \Theta_r > T_C^{\Theta_{r-1}} \}.$$

The Distinct Document Counting Process II

Lemma

For each s in the real line, $X^s := (X_t^s)_{t \geq s}$ is an inhomogeneous Poisson process with intensity function

$$\Xi^s(t) = \mathbb{E}[\, X_t^s \,] = \gamma \int_{-\infty}^{t} \mathbb{E}\left[\, 1 - \mathrm{e}^{-(\Lambda_a(t) - \Lambda_a(s))} \,\right] \mathrm{d}a\,, \quad t \geq s\,.$$

Proof.

Use the Poisson point process properties of $\widetilde{\Gamma}$. □

Notation: $X \triangleq X^0$ and $T_C \triangleq T_C^0$ and $\Xi \triangleq \Xi^0$.
In particular,

$$T_C^s \overset{\text{law}}{=} s + T_C\,.$$

Expected Number of Misses

This framework allows us to derive rigorously the following.

Theorem (Integral Formula for Expected Misses)

The expected number of misses of the tagged document satisfies

$$\mathbb{E}[\,\mu_C\,] = \mathbb{E}[\,m(T_C)\,],$$

where T_C is the exit time of a document requested at time 0, and

$$m(t) = \mathbb{E}\left[\int_0^\infty \lambda(u)\mathrm{e}^{-(\Lambda(u+t)-\Lambda(u))}\,\mathrm{d}u\right], \quad t \geq 0\,.$$

Recall that the miss probability is given by $p_C = \mathbb{E}[\,\mu_C\,]/\mathbb{E}[\,\Lambda\,]$.

Elements of Proof

The proof uses the following.

Lemma (Functionals of Holding Times)

Let ξ be an inhomogeneous Poisson process on $\mathbb{R}_+$ with deterministic intensity function λ.
Assume that $\Lambda(\infty) < \infty$ and denote by $(\Theta_r)_{r=1}^N$ the instants of ξ.
Then, for any $F : \mathbb{R}_+ \to \mathbb{R}$,

$$\mathbb{E}\left[\mathbb{1}\{N \geq 2\} \sum_{r=2}^{N} F(\Theta_r - \Theta_{r-1})\right]$$
$$= \int_0^\infty \int_0^\infty F(w)\lambda(u)\lambda(u+w)\mathrm{e}^{-(\Lambda(u+w)-\Lambda(u))}\,\mathrm{d}u\,\mathrm{d}w\,.$$

Interpretation in Terms of TTL Cache

By integration by parts

$$\begin{aligned}
&\int_0^\infty \lambda(u)\mathrm{e}^{-(\Lambda(u+t)-\Lambda(u))}\,\mathrm{d}u \\
&\quad = \mathrm{e}^{-(\Lambda(u+t)-\Lambda(u))}\Big]_{u=0}^{\infty} + \int_0^\infty \lambda(u+t)\mathrm{e}^{-(\Lambda(u+t)-\Lambda(u))}\,\mathrm{d}u \\
&\quad = 1 - \mathrm{e}^{-\Lambda(t)} + \int_t^\infty \lambda(u)\mathrm{e}^{-(\Lambda(u)-\Lambda(u-t))}\,\mathrm{d}u \\
&\quad = \int_0^\infty \lambda(u)\mathrm{e}^{-(\Lambda(u)-\Lambda(u-t))}\,\mathrm{d}u\,.
\end{aligned}$$

Interpretation in Terms of TTL Cache

Hence

$$m(t) = \mathbb{E}\left[\int_0^\infty \lambda(u) \mathrm{e}^{-(\Lambda(u+t)-\Lambda(u))}\,\mathrm{d}u\right]$$
$$= \mathbb{E}\left[\int_0^\infty \lambda(u) \mathrm{e}^{-(\Lambda(u)-\Lambda(u-t))}\,\mathrm{d}u\right]$$

and since

$$\mathrm{e}^{-(\Lambda(u)-\Lambda(u-t))} = \mathbb{P}[\,\xi([u-t,u]) = 0 \mid \Lambda\,],$$

we can interpret $m(t)$ as the average number of misses for a t-TTL (time to live) cache, in which a document is evicted after any interval of time of length t with no request for it.

Relation between Ξ and m

Recall that $\Xi := (\Xi(t))_{t\geq 0}$ is the intensity function of the Poisson process $X := (X_t)_{t\geq 0}$ counting the number of distinct documents in the rest of the request process after time 0.

Theorem

The functions $\Xi := (\Xi(t))_{t\geq 0}$ *and* $m := (m(t))_{t\geq 0}$ *satisfy*

$$\Xi'(t) = \gamma m(t), \qquad \Xi(t) = \gamma M(t), \quad M(t) \triangleq \int_0^t m(s)\,\mathrm{d}s\,.$$

Proof.

Use integration by parts and change of variables as before. □

Relation between Ξ and m

Hence, considering left-continuous inverses,

$$\Xi(t) = x \Leftrightarrow M(t) = \frac{x}{\gamma}, \quad \Xi^{-1}(x) = M^{-1}\left(\frac{x}{\gamma}\right), \qquad x \geq 0\,.$$

The exit time T_C can be expressed as

$$T_C = \Xi^{-1}(\widehat{T}_C)$$

where $\widehat{T}_C$ is the first hitting time of C by a unit Poisson process. Thus

$$\mathbb{E}[\,\mu_C\,] = \mathbb{E}[\,m(T_C)\,] = \mathbb{E}\left[\,m\left(M^{-1}\left(\frac{\widehat{T}_C}{\gamma}\right)\right)\right].$$

Scaling Limit

Since $\widehat{T}_C$ is the sum of C i.i.d. exponential $\mathscr{E}(1)$ random variables, the strong law of large numbers (SLLN) implies that

$$\frac{\widehat{T}_C}{C} \xrightarrow[C\to\infty]{\text{a.s.}} 1$$

(and we have great convergence estimates). This and

$$\mathbb{E}[\,\mu_C\,] = \mathbb{E}\left[m\left(M^{-1}\left(\frac{\widehat{T}_C}{\gamma} \right) \right) \right]$$

lead us to consider the natural scaling limit in which the cache size C and the arrival rate γ go to infinity in proportion:

$$C, \gamma \to \infty\,, \quad \text{with} \;\; C = \gamma\theta \;\; \text{for some fixed} \;\; \theta > 0\,.$$

Little's Law

The fixed parameter θ has a simple interpretation as the

average sojourn time in cache for a document,

since by Little's law

$$C = \gamma \mathbb{E} \int_0^\infty \mathbb{1}\{\text{a given document is in cache at time } t\}\, \mathrm{d}t\,.$$

Justification of the Che Approximation

In this scaling limit, the SLLN yields by dominated cv. that

$$\mathbb{E}[\,\mu_C\,] = \mathbb{E}\left[m\left(M^{-1}\left(\frac{\widehat{T}_C}{\gamma} \right) \right) \right] \xrightarrow[C\to\infty]{} m(t_\theta)\,, \qquad t_\theta \triangleq M^{-1}(\theta)\,.$$

In this context, the heuristic Che approximation consists in:

- Replacing the exit time T_C by the deterministic time $\tilde{t}_C \triangleq \Xi^{-1}(C)$, called the characteristic time.
- Replacing the expected number of misses $\mathbb{E}[\,\mu_C\,] = \mathbb{E}[\,m(T_C)\,]$ by $m(\tilde{t}_C)$.

A rigourous justification is brought here since

$$\tilde{t}_C \triangleq \Xi^{-1}(C) = M^{-1}\left(\frac{C}{\gamma} \right) = M^{-1}(\theta) \triangleq t_\theta\,.$$

Expansion in powers of $1/C$

A probabilistic asymptotic analysis will yield our main result.

Theorem (Expansion in Powers of $1/C$)

Assume that m is $\mathscr{C}^2$ on $(0, +\infty)$. Consider the scaling limit $C \to \infty$ with $\gamma = C/\theta$ for some fixed $\theta > 0$. Let $t_\theta \triangleq M^{-1}(\theta)$. Then

$$\mathbb{E}[\,\mu_C\,] = m(t_\theta) + e(t_\theta)\frac{1}{C} + o\left(\frac{1}{C}\right)$$

where

$$e(t_\theta) \triangleq \frac{\theta^2}{2m(t_\theta)^2}\left(m''(t_\theta) - \frac{m'(t_\theta)^2}{m(t_\theta)}\right).$$

This expansion can be computed from the system parameters. It quantifies and improves the accuracy of the Che approximation, which corresponds to its 0th order.

Elements of Proof

Let

$$f_\theta : x \in \mathbb{R}_+ \mapsto f_\theta(x) = m(M^{-1}(\theta x)) = m(t_{\theta x})$$

so that

$$\mathbb{E}[\,\mu_C\,] = \mathbb{E}[\,f_\theta(X_C)\,], \qquad X_C \triangleq \frac{\widehat{T}_C}{C}\,.$$

The Taylor formula yields, for some r.v. Y_C in $[1, X_C] \cup [X_C, 1]$,

$$\begin{aligned} f_\theta(X_C) &= f_\theta(1) + f'_\theta(1)(X_C - 1) + \frac{1}{2} f''_\theta(Y_C)(X_C - 1)^2 \\ &= f_\theta(1) + f'_\theta(1)(X_C - 1) + \frac{1}{2} f''_\theta(1)(X_C - 1)^2 \\ &\qquad + \frac{1}{2} (f''_\theta(Y_C) - f''_\theta(1))(X_C - 1)^2\,. \end{aligned}$$

Recall that $f_\theta(1) = m(t_\theta)$.

Elements of Proof

Since

$$\mathbb{E}[\, X_C - 1 \,] = 0\,, \qquad \mathbb{E}[\, (X_C - 1)^2 \,] = \mathrm{Var}\left[\frac{\widehat{T}_C}{C}\right] = \frac{1}{C}\,,$$

it holds that

$$\mathbb{E}[\, \mu_C \,] = f_\theta(1) + \frac{1}{2C} f_\theta''(1) + \frac{1}{2}\mathbb{E}[\, (f_\theta''(Y_C) - f_\theta''(1))(X_C - 1)^2 \,]$$

and it remains to control the last remainder term and express $f_\theta''(1)$ in terms of m and t_θ to finish the proof.

This control is technical.
It uses in particular the exponential upper bound
in Cramer's large deviation theorem.

Elements of Proof

Another proof is to write explicitly the integral

$$\mathbb{E}[\,\mu_C\,] = \mathbb{E}\left[f_\theta\left(\frac{\widehat{T}_C}{C}\right)\right] = \frac{C^C}{\Gamma(C)}\int_0^\infty \mathrm{e}^{-C(w-\ln w)}\,\frac{f_\theta(w)}{w}\,\mathrm{d}w$$

and use Laplace's asymptotic method and the Stirling formula.

This method is actually more complicated: it involves the expansion of both numerator and denominator in powers of $\sqrt{C}$, and compensations between these powers.

With both methods, the expansion can be continued to the order n if m is in $\mathscr{C}^{2n}$.

Numerical experiments

We illustrate numerically this expansion by comparing it to values obtained from system simulation. For the popularity we use

$$\lambda(t) = R\mathbb{1}\{\, 0 \le t \le L \,\}, \qquad t \in \mathbb{R},$$

where the request rate R and lifespan L of a document are independent and have Pareto-Lomax density

$$\frac{\alpha\,\beta^{\alpha}}{(x+\beta)^{\alpha+1}}\mathbb{1}\{\, x > 0 \,\}$$

with resp. parameters $\alpha = 1.9$, $\beta = 22.5$ and $\alpha = 1.7$, $\beta = 0.07$. Then $\mathbb{E}[\, R \,] = 25$ and $\mathbb{E}[\, L \,] = 0.1$ with no variance.

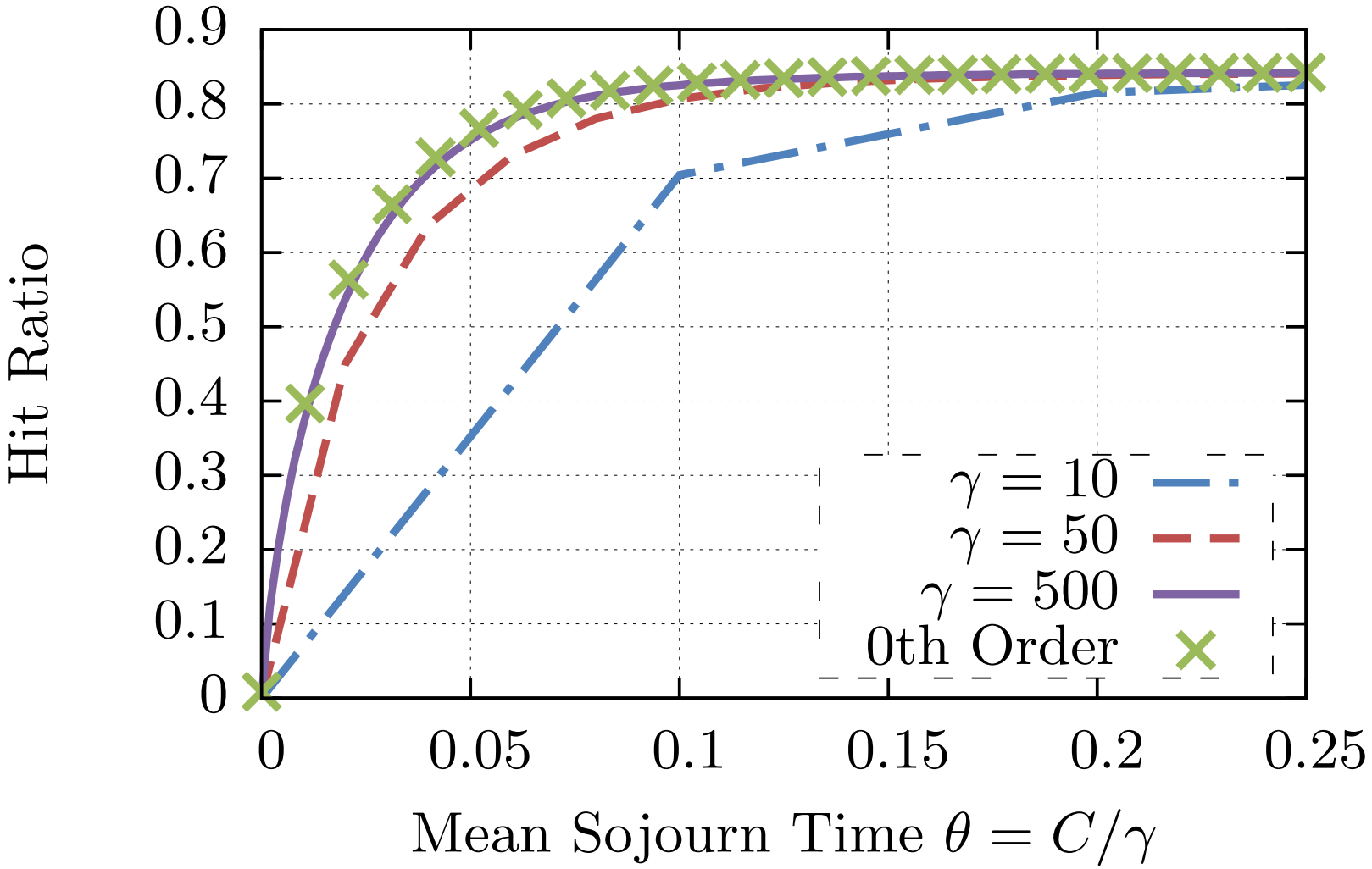
Numerical experiments

We use numerical integration and numerical inversion to compute the expansion.

We use the stable-law central limit theorem to ensure that the Monte-Carlo simulations are quite accurate.

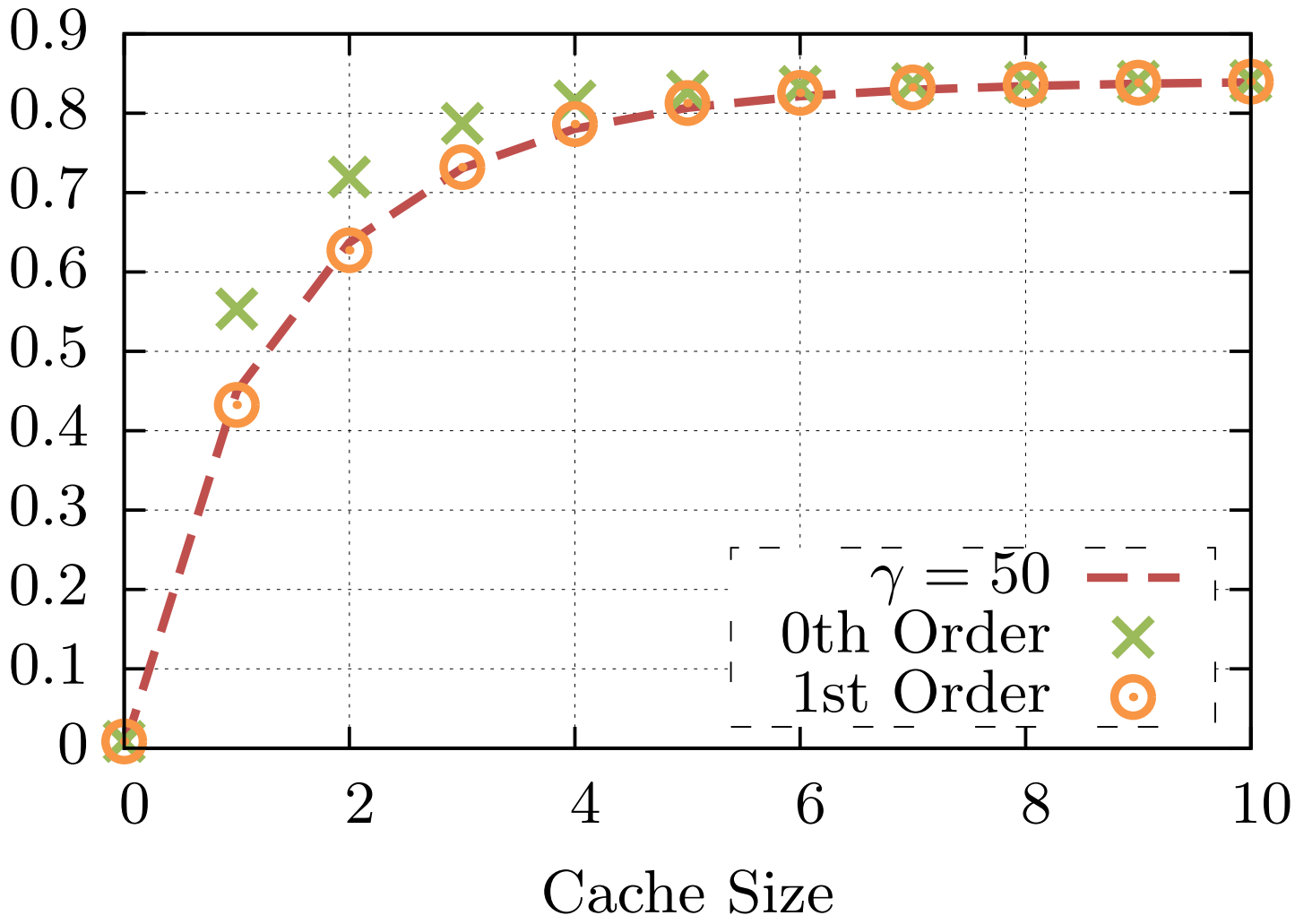
We give some examples and comments on the results.

Convergence to the Limit: Che Approximation



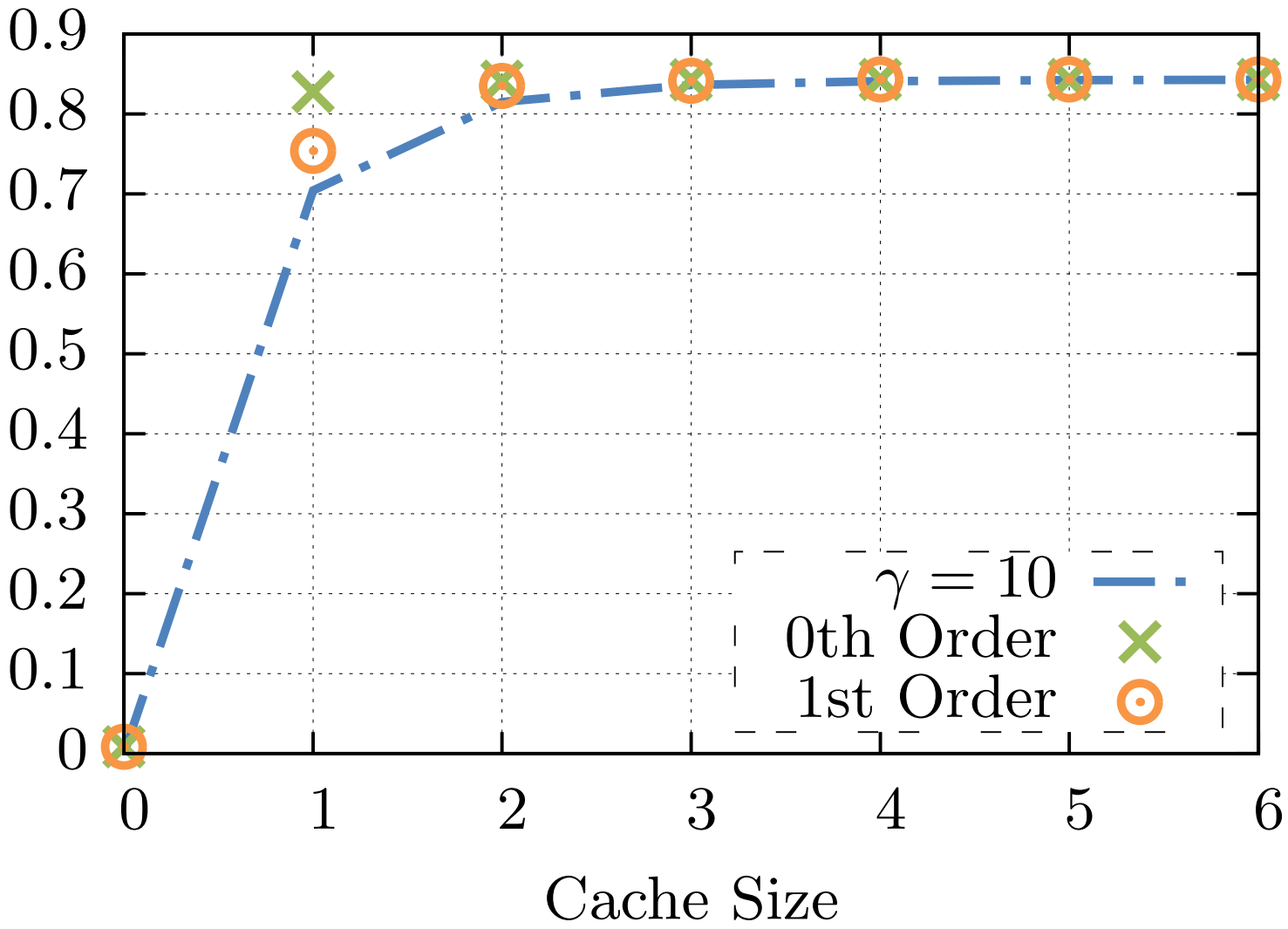
0th order approximation (limit). Quite accurate for $\gamma = 500$.

Che Approximation vs. First Order



The 0th and 1st order approximation for $\gamma = 50$.

First Order Imperfection



The 0th and 1st order approximation for $\gamma = 10$.
A higher order expansion could be needed.

Bibliography

Thank you for your attention!

- Olmos, Felipe, Carl Graham, and Alain Simonian (2015). "Cache miss estimation for non-stationary request processes". In: *ArXiv:1511.07392*.
- Olmos, Felipe, Bruno Kauffmann, et al. (2014). "Catalog Dynamics: Impact of Content Publishing and Perishing on the Performance of a LRU cache". In: *26th International Teletraffic Congress (ITC)*. IEEE.
- Traverso, Stefano et al. (2013). "Temporal locality in today's content caching: why it matters and how to model it". In: *ACM SIGCOMM Computer Communication Review* 5.