Performance Evaluation in Networks

Eric Thierry*
ENS de Lyon

Contents

*eric.thierry@ens-lyon.fr

Part I
Queueing theory

0 Introduction

0.1 Generalities

Performance of computer and communication systems/networks. Objectives:

- Observation
- Prediction
- Control & Optimization

Background needed:

- Not much, try to give always the specification of the system we will use

Systems can be architecture/hardware, code/software , communication network (included distributed systems) and logistic/industrial processes, etc.

The metrics will be:

- Speed, bandwidth, delay, load, losses.
- Worst case, average, ..

Use of either a mathematical analysis of an abstract model or a simulation (math model, scale model). Comparisons with experiments/measures on a real system (statistics).

Objectives:

- Designing and analysing mathematical models (probability assumptions)
- Improve knowledge in probability/statistical tool
- Improve knowledge about communication networks
- practising simulation/statistic with a simulation/statistic software

Why are there so many probabilities?

- Sometimes intrinsic in the system (i.e. noise in the core of the system
- The users themselves introduce probabilities (Ethernet protocol and randomize algorithms)
- Probabilistic is a good way to have a good "rough estimation"
- Usually works well in practice

Use of statistics: Do I need to do more experiments? Do I need to run them longer (asymptotical behaviour)?

Classical questions: Overload? Average waiting time?

Bibliography:

- Performance Evaluation of Computer and Communication systems
- Introduction aux probabilités et à la statistique
- The Art of Computer Systems Performance Analysis
- http://perfeval.epfl.ch/
- http://proweb.

0.2 Refresher about Probability

Cf first year

Usual working hypothesis:

- A list Ω of the outcomes and a measure of their occurrences via the measures of events.

Definition 1. *Ω is the universe, $\mathcal{F}$ a set of subsets of Ω, $\mathbb{P}$ function from Ω to $\mathbb{R}$.*

Vocabulary:

- $\omega \in \Omega$ is an outcome
- $A \in \mathcal{F}$ is an event
- ω realise A if $\omega \in A$

Recalls:

- Conditional probabilities
- Independence of events
- Law of total probabilities
- General/Real Random Variable
- Cumulative distribution function: $F_x(x) = \mathbb{P}(X \leq x)$ ($\lim\limits_{x \to -\infty} F(x) = 0$ and $\lim\limits_{x \to +\infty} F(x) = 1$, F non decreasing, F right continuous)
- Discrete/Continuous r.v.[1]
- Random vectors & Joint distribution
- Expectation
- Composition of discrete/continuous r.v. (if g is Lebesgue integrable, then $\mathbb{E}(g(X)) = \int_x g(x)f(x)dx$)
- Same for vectors
- Moments of a r.v. (generalisation of expectation and standard deviation)

[1] Random Variable

- Generating functions (Probabilities $G_X(s) = \mathbb{E}(s^x)$, moments $M_X(t) = \mathbb{E}(e^{tX})$, characteristic $\Phi_X(t) = \mathbb{E}(e^{itX})$)
- Markov, Bienaymé-Tchebychev, Jensen (convex function), Hölder, Minkowski
- Chernoff and Hoeffding
- Convergence (in law, in proba, almost sure)
- Central limit theorem
- Law of large numbers

1 Random number generation

cf Webpage

2 Markov chains

Recalls

3 Martingales and stuff

4 Continuous time Markov chains

Definition 2. *A process $(X_t)_{t\in\mathbb{R}_+}$ over a space of states E is a homogeneous continuous Markov chain if:*

$$\begin{aligned}\forall t_1 < ... < t_n \in \mathbb{R}_+, \mathbb{P}(X_{t_n} = i_n | X_{t_{n-1}} = i_{n-1}, ..., X_{t_1} = i_1) &= \mathbb{P}(X_{t_n} = i_n | X_{t_{n-1}} = i_{n-1} \qquad \text{(Markov)}\\ &= \mathbb{P}(X_{t_n - t_{n-1}} = i_n | X_0 = i_{n-1}) \qquad \text{(Homogeneous)}\end{aligned}$$

Notation: $p_{ij}(t) = \mathbb{P}(X_t = j | X_0 = i) \to P_t = (p_{ij}(t))_{i,j\in E}$

Kolmogorov equations: $P_{t+s} = P_s P_t, \forall s, t \in \mathbb{R}_+$ (same proof)

Discrete case: $P_n = P^n$ with $P = P_1 = \mathbb{P}(X_1 = j | X_0 = i)$ ("kernel")

Continuous case: $P_t \underset{t\to 0}{\to} ?$

Classical assumption: $P_t \underset{t\to 0}{\to} Id$ i.e $\forall i, j, p_{ij}(t) \underset{t\to 0}{\to} 0$ if $i \neq j$ and $p_{ii}(t) \underset{t\to 0}{\to} 1$
("standard") ("continuity") ("differentiability at 0")

$$\begin{aligned}\exists Q, \frac{P_t - Id}{t} &\underset{t\to 0}{\to} Q\\ \text{i.e.} \forall i \neq j, \frac{p_{ij}(t)}{t} &\to q_{ij}\\ \forall i, \frac{p_{ii}(t) - 1}{t} &\to q_{ii}\end{aligned}$$

Q = infinitesimal generator = "kernel"

Remarks about coeff of Q:

- $\forall i \neq j$, $q_{ij} \geq 0$
-
$$\begin{aligned} \forall t \in \mathbb{R}_+, \sum_{j \in E} p_{ij}(t) &= 1 \\ \Rightarrow \forall t > 0, \frac{\sum_{j \in E} P_{ij}(t) - 1}{t} &= 0 \\ (t \to 0) \Rightarrow \sum_{j \in E} q_{ij} &= 0 \\ \Rightarrow q_{ii} &\leq 0 \end{aligned}$$

Vocabulary: q_{ij} = transition *rate* from i to j

Theorem 1 (Kolmogorov)**.** *With previous assumptions,*

$$\forall t \in \mathbb{R}, \frac{\mathrm{d}P_t}{\mathrm{d}t} = QP_t$$

Proof. Use $P_{t+s} = P_t P_s$, more precisely:

$$\begin{aligned} p_{ij}(t+h) - p_{ij}(t) &= \sum_k p_{ik}(h) p_{kj}(t) - p_{ij}(t) \\ &= \sum_{k \neq i} p_{ik}(h) p_{kj}(t) - (p_{ii}(h) - 1) p_{ij}(t) \end{aligned}$$

Divide by h and make h tend to 0:

$$\begin{aligned} \lim_{h \to 0} \frac{p_{ij}(t+h) - p_{ij}(t)}{h} &= \sum_{i \neq k} q_{ik} p_{kj}(t) + q_{ii} p_{ij}(t) \\ &= \text{coeff } (i,j) \text{ of } QP_t \end{aligned}$$

□

Theorem 2 (Kolmogorov bis)**.**
$$\frac{\mathrm{d}P_t}{\mathrm{d}t} = P_t Q$$

Corollary 1. *Since we know that $P_0 = I$ the identity matrix:*

$$P_t = e^{tQ} \underset{def}{=} \sum_{n \in \mathbb{N}} \frac{(tQ)^n}{n!}$$

Remark: all the work on finite matrices apply to those infinite matrices up to defining/using matrix norms (many norms defined by combinations of sup, $\sum$)

Alternate description: "jump process". 1 Trajectory of HMC = alternation of *waiting times* and jumps.

Waiting times: suppose you are at state i at time 0.
Consider

$$\begin{aligned}
W &= \text{the time when you leave } i \\
&= \inf\{t \geq 0 | X_t \neq i\} \\
\mathbb{P}(W > s+t | W > s) &= ? \\
\{W > s\} &= \{\forall 0 \leq u \leq s, X_u = i\} \\
&= \mathbb{P}(\forall \underbrace{0 \leq u \leq s+t}_{s \leq u \leq s+t}, X_u = i | \forall \underbrace{0 \leq u \leq s, X_u = i}_{X_s = i}) \\
&= \mathbb{P}(W > t) \ \rightarrow \ \text{Memoryless law} \\
&= exp(\underbrace{\lambda}_{?}) \\
\mathbb{P}(W > t) &= e^{-\lambda t} \underset{t \text{ small}}{\simeq} p_{ii}(t) = \mathbb{P}(X_t = i | X_0 = i) \\
&\overset{\text{def}}{=} 1 + Q_{ii} t + o(t)
\end{aligned}$$

Jump probability: supp. that you are at state i at time 0.
Consider an an interval $[s, s+t[$ such that $s \leq W < s+t$, and suppose that the chain jumps only in this interval,

$$\mathbb{P}(\text{jump to } i | \text{it jumps}) \underset{t \text{ small}}{\simeq} \frac{p_{ij}(t)}{1 - p_{ii}(t)} = \frac{q_{ij} t + o(t)}{-q_{ii} t + o(t)} \overset{t \to 0}{\rightarrow} \frac{q_{ij}}{-q_{ii}}$$

Evolution/asymptotic behaviour of HMC

Invariant measure/distribution: $\pi \in \mathbb{R}_+^E$ satisfying $\forall t \leq 0, \pi P_t = \pi$ (and proba dist, i.e. $\sum_{i \in E} \pi_i = 1$)

Theorem 3.

$$\forall t \geq 0, \pi P_t = \pi \Leftrightarrow \pi Q = 0$$

Proof.

$$\begin{aligned}
\pi Q = 0 &\Leftrightarrow \forall n \in \mathbb{N}^*, \pi Q^n = 0 \\
&\Leftrightarrow \forall t \in \mathbb{R}_+, \pi \sum_{n=0}^{+\infty} \frac{t^n Q^n}{n!} = \pi
\end{aligned}$$

□

Irreducible HMC: rate transition graph

$$= \begin{cases} \text{vertices} & \text{states } E \\ \text{directed edges } i \to j & \text{if } q_{ij} > 0 \end{cases}$$

irreducible $\overset{\text{def}}{=}$ graph strongly connected.

Theorem 4 (Convergence result)**.** *Suppose that an HMC is irreducible and admits an invariant proba distribution π.*

$$\lim_{t \to \infty} P_{ij}(t) = \pi_j$$

Moreover, let $f : E \to \mathbb{R}$ such that $\sum_{i \in E} \pi_i |f(i)| < \infty$.
Then

$$\lim_{t \to +\infty} \frac{1}{t} \int_0^t f(X_s) \mathrm{d}s = \sum_{i \in E} \pi_i f(i)$$

5 Queues

Queue = buffer + server

Classical questions:

- Average waiting time for a client
- Proportion of time the server is busy
- Distribution of nb of clients waiting in the queue

Kendall notation (for single queue)

$$\underbrace{A}_{\substack{\text{law of interarrival}\\\text{of clients}}} \;/\; \underbrace{B}_{\substack{\text{law of service}\\\text{time (time to}\\\text{serve one client)}}} \;/\; \underbrace{S}_{\substack{\text{nb of}\\\text{process}}} \;(/\; \underbrace{N}_{\substack{\text{nb of}\\\text{buffers}}}\;)$$

All are usually assumed independent.

Classical laws:

- M (Markov) = exponential law
- D (Deterministic) = constant time
- E (Erlang) = Erlang law
- G (General) = arbitrary law

Example: analysis of the $M(\lambda)/M(\mu)/1$ queue.

$$X_t = \text{ size of the queue at time } t$$

Different behaviour:

- $X_t = 0 \rightarrow$ new client will arrive within time $\sim exp(\lambda)$
- $X_t \geq 1 \rightarrow$ one client is being saved within time $\sim exp(\mu)$ and one new client will arrive within time $\sim exp(\lambda)$

5.1 M/M/1 queue

$(N_t)_{t\in\mathbb{R}_+}$ is a continuous time HMC.

Search for invariant proba distribution π $\pi Q = \pi \Rightarrow \forall n \in \mathbb{N}, \pi_n = \left(\frac{\mu}{\lambda}\right)^n \pi_0$ An invariant proba measure exists iff $\sum_{n=0}^{+\infty} \pi_n = 1 \Leftrightarrow \frac{\mu}{\lambda} < 1$. Then $\pi_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right)$ (geometric law).
under the assumption that the system is at the permanent/stationary state/regime, i.e. N_0 follows the invariant distribution π, let's study some performance parameters.

1. Average nb of packets in the system (we assume that we have the distribution π at time 0)

$$\mathbb{E}_\pi(N_t) = \sum_{n=0}^{+\infty} n\pi_n = \frac{\frac{\mu}{\lambda}}{1 - \frac{\mu}{\lambda}} \leftarrow \text{ mean for geometric law}$$

2. Let W = waiting time before being served

   Two cases:

   - arriving in an empty queue $\Leftrightarrow$ being served immediately
     $\mathbb{P}(W=0) = \mathbb{P}_\pi(N_t = 0) = \pi_0 = 1 - \frac{\mu}{\lambda}$
   - arriving in a non-empty queue: if there is a packet (with probability π_n) you must wait or sum of n independent exponential laws $Exp(\lambda)$
     $\rightarrow$ density $f(t)$ for W:

$$f(t) = \sum_{n=1}^{+\infty} \underbrace{\pi_n}_{\left(\frac{\mu}{\lambda}\right)^n\left(1-\frac{\mu}{\lambda}\right)} \times \underbrace{\text{density of a sum of independent } Exp(\lambda)}_{e^{-\lambda t}\frac{\lambda^n t^{n-1}}{(n-1)!}}$$
$$= e^{-\lambda t}\left(1-\frac{\mu}{\lambda}\right)\mu \underbrace{\sum_{n=1}^{+\infty}\frac{(\mu t)^{n-1}}{(n-1)!}}_{e^{\mu t}}$$
$$= e^{-(\lambda-\mu t}(\lambda-\mu)\frac{\mu}{\lambda}$$

We have the law of W, let's compute $\mathbb{E}_\pi(W)$

$$\mathbb{E}_\pi(W) = \int_{t=0}^{+\infty} tf(t)\mathrm{d}t = \int_0^{+\infty} te^{-(\lambda-\mu)t}(\lambda-\mu)\frac{\lambda}{\mu}\mathrm{d}t$$
$$= \frac{\mu}{\lambda}\int_0^{+\infty} t \underbrace{(\lambda-\mu)e^{-(\lambda-\mu)t}}_{\text{density for } Exp(\lambda-\mu} \mathrm{d}t$$
$$= \frac{\mu}{\lambda}\frac{1}{\lambda-\mu}$$

Average time before leaving the system $\overline{T}$ i.e. spent in the system $= \mathbb{E}_\pi(W) + \underbrace{\frac{1}{\lambda}}_{\substack{\text{average}\\ \text{waiting time}}} = \frac{\mu}{\lambda}\frac{1}{\lambda-\mu} + \frac{1}{\lambda} =$

$\frac{1}{\lambda-\mu}$

Remark

$$\underbrace{\overline{N}}_{\substack{\text{average size}\\ \text{of the queue}}} = \underbrace{\mu}_{\text{input rate}} \underbrace{\overline{T}}_{\substack{\text{average time}\\ \text{spent in the queue}}} \qquad \text{(Little's law)}$$

5.2 M/M/∞ queue

Behaviour of N_t = number of packet in the system?

Search for an invariant distribution π Substitution:

$$\pi_n = \frac{1}{n!}\left(\frac{\mu}{\lambda}\right)^n \pi_0$$

There is always an invariant distribution because $\sum_{n=0}^{+\infty}\pi_n = \pi_0 \sum_{n=0}^{+\infty}\frac{1}{n!}\left(\frac{\mu}{\lambda}\right)^n = \pi_0 e^{\frac{\mu}{\lambda}}$

$$\Rightarrow \forall n \in \mathbb{N}, \pi_n = e^{-\frac{\mu}{\lambda}} \frac{1}{n!} \left(\frac{\mu}{\lambda}\right)^n$$

Under the stationary assumption

1. $\mathbb{E}_\pi(N_t) = \overline{N} =$ mean value for Poisson laws of parameter $\frac{\lambda}{\mu} = \frac{\lambda}{\mu}$
2. $\mathbb{E}_\pi(\underbrace{\text{time spent in the system}}_{\text{service time}}) =$ mean value for $Exp(\lambda) = \frac{1}{\lambda}$.

Remark

$$\overline{N} = \mu \overline{T}$$

A useful relation: Little's law

In many cases,

$$\text{average nb of guys in the system} = \text{average arrival rate} \times \text{average time spend in the system}$$
$$\overline{N} = \mu \overline{T}$$

Examples

1. M/M/1 queue in the stationary state
2. M/M/∞ in the stationary state
3. General deterministic case with finite horizon

Definition 3 (Pseudo-inverse).

$$f^{(-1)}(n) = \inf\{t \mid f(t) \geq n\}$$

Definition 4 (Definitions).

- $A(t) =$ *number* $n \in \mathbb{N}$ *of packets entering between time 0 and time* $t \in \mathbb{R}_+$
- $B(t) =$ *number* $n \in \mathbb{N}$ *of packets leaving between time 0 and time* $t \in \mathbb{R}_+$
- $N(t) = A(t) - B(t) =$ *number of packet* in *the system at time* t
- $T(n) = B^{(-1)}(n) - A^{(-1)}(n) =$ *time spent in the system by the* n*-th packet*

"Averages:"

- *"finite horizons"* $=$ *on* $[0, t]$
- $\mu = \frac{A(t)}{t}$ *(arrival rate)*
- $\overline{T} = \frac{1}{A(t)} \sum_{i=1}^{A(t)} T(i)$ *(average time spent in the system)*
- $\overline{N} = \frac{1}{t} \int_{x=0}^{t} N(x) \mathrm{d}x$ *(average number of packets).*

Theorem 5 (Little's theorem (1961)). *If* $A(t) = B(t)$

$$\overline{N} = \mu \overline{T}$$

Proof. Graphical proof, interpreting integrals as areas.

$$\underbrace{\frac{1}{t}\int_{x=0}^{t} N(x)\mathrm{d}x}_{\overline{N}} = \underbrace{\frac{A(t)}{t}}_{\mu} \underbrace{\frac{1}{A(t)}\sum_{i=1}^{A(t)} T(i)}_{\overline{T}}$$

□

Part II
Statistics

1 Framework

From data samples(s), statistical tools aims at:

1. Clarify/Sum up/Compress the information (indicators, graphics) $\rightarrow$ *descriptive statistics*
2. Modelize the "randomness" underlying the generation of the data sample(s) $\rightarrow$ *inferential statistics*

2 Descriptive statistics

Example of indicators (to describe/summarize data)

- *Position:* give a "central" value
- *Dispersion:* dispersion around a "central value
- *Shape:* how data is distributed up to shifting the "central" values

Two versions "statistical", for a sample sorted $x_1 < ... < x_n \in \mathbb{R}$; and "probabilistic", for a real random variable X.

Position	Stat version	Proba version
Mean μ	$\frac{1}{n}(x_1 + ... + x_n)$	$\mathbb{E}(X)$
Median m	$x_{\lfloor \frac{n+1}{2} \rfloor}$	$\mathbb{P}(X < m) leq \frac{1}{2},\ \mathbb{P}(X > m) \leq \frac{1}{2}$
Mode	largest value	value maximizing the law

Dispersion	Stat version	Proba version
α-quantile q_α ($\alpha \in [0,1]$)	$x_{\lfloor \alpha(n+1) \rfloor}$	$\mathbb{P}(X < q_\alpha) leq \alpha,\ \mathbb{P}(X > q_\alpha) \leq 1 - \alpha$
Variance σ^2	$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$	$\mathbb{E}((X - \mathbb{E}(X))^2)$

Link between the two versions: For instance for data $\in \mathbb{R}$, data $= x_1, ..., x_n \ \rightarrow$ "empirical" distribution/law discrete with possible repetitions with mass function $f(x) = \frac{card\{i \mid x_i = i\}}{n}$

Example of graphic item

- box plot ("boîte à moustache")
- histogram

Objective: Avoid hiding some unpleasant bad cases by only analysing the mean of some results.

Complexity of computing indicators (input data not sorted) "easy", linear time and space for the classical indicators

3 Inferential statistics

3.1 General question

- Data: sample$(x_1, ..., x_n) \in E^n$
- Models:
  - Parametric: models from a family of *laws* parametrized by one or several parameters θ
  - Non-parametric: no restriction on the family of laws
- Question: finding the model(s) *fitting the best* the *generation* of the sample

3.2 An example

A device producing valid messages (0) or invalid ones (1).
$n = 20$ messages $\rightarrow$ experimental sample:

00010 01100 01010 00000

Modelization test: consider that the validity of each message is independent of the other, and only depends on the device which generates invalid messages with probability p.
$\rightarrow$ model: product of independent Bernoulli laws with some parameter p

Question: what is the value of p?

Formalize "the generation of the sample" $\rightarrow$ "the sample is a realization for a law of the model", i.e. $(x_1, ..., x_n) = (X_1(\omega), ..., X_n(\omega)$ for $\omega \in \Omega$, $X_1, ..., X_n$ iid Bernoulli(p).

- intuitively, $p < \frac{1}{2}$ because the number of 1 < number of 0
- law of large numbers:
$$\underbrace{\frac{1}{n}(X_1 + ... + X_n)}_{\mathbb{P}(\omega\in\Omega \mid \frac{1}{n}(X_1(\omega),...)\rightarrow p)=1} \overset{\text{a.s.}}{\rightarrow} \mathbb{E}(X_1)$$
(We know $\mathbb{E}(X_1) \leq \infty$) $\Rightarrow$ looking at $\frac{1}{n}(x_1 + ... + x_n)$ should approximate p.
- Weak law of large numbers states something about convergence speed, or Bienaymé-Chebytchev $\mathbb{P}(|\frac{1}{n}(X_1 + ... + X_n) - p| > \epsilon) \leq \frac{\mathbb{V}X_1}{n\epsilon^2} \rightarrow p(1-p)$
- $p \neq 0, p \neq 1 \rightarrow$ otherwise no 1, no 0
- Let $0 < p < 1 \rightarrow$ the experimental sample may occur with proba > 0.
With B-C inequality:
$$\mathbb{P}(\frac{1}{n}(X_1 + ... + X_n) - \epsilon < p, \frac{1}{n}(X_1, ..., X_m) + \epsilon) \geq 1 - \frac{p(1-p)}{n\epsilon^2}$$

I would like to say: “p lies in this interval with good probability”

$$\mathbb{P}(p \in [I_n^-, I_n^+]) \geq \alpha_n$$

(Several versions exist depending the choice of ϵ). Beware that I_n^+ and I_n^- are random variables !
If $(x_1, ..., x_n)$ is a realisation of $(X_1, ..., X_n)$, one can compute $\frac{1}{n}(x_1 + ... + x_n) = \frac{5}{20} = 0.25$. Choosing ϵ to get an interesting statement from B-C is not easy here, because n is rather small (B-C is a bit too loose here). It is easier e.g. for the sample $n = 100$ =concatenation of the previous sample 5 times
$\alpha_n = 1 - \frac{1}{4n\epsilon^2} = 1 - \frac{1}{400 \times 0.001} = 0.75$ if $\epsilon = 0.1$
$\frac{1}{n}(x_1 + ... + x_n) - \epsilon = 0.25 - 0.1 = 0.15$
$\frac{1}{n}(x_1 + ... + x_n) + \epsilon = 0.25 + 0.1 = 0.35$

WARNING Do *never* write

$$"\mathbb{P}(p \in [0.15, 0.35]) \geq 0.75"$$

Because p is *not* a random variable, it is a fixed value (unknown).
A proper way to state something is: “The interval $[0.15, 0.35]$ is a confidence interval for p with confidence/guarantee 0.75”