Capacity of Single-Server Queues

Based on "Bits Through Queues" by Anantharam and Verdu (1996)

6.962 Week 11, 26 April 2001

Tengo Saengudomlert, *tengo@mit.edu*

Brief Summary

This work investigates the fact that information resides not only in the message bits, but also in their arrival times.

Consider a transmission system with one source and one destination. Transmitted messages from the source enter a single-server queue ($\cdot/G/1$) before arriving at the destination. The channel capacity of this system, referred to as the *capacity of the queue*, can be defined as the maximum achievable transmission rate through this queue. If the service time in the queue is deterministic, the capacity of the queue is infinite since the decoder can distinguish between infinite choices of transmission times, yielding the infinite amount of information bits per transmission. When the service time is a random variable (RV) with mean $1/\mu$, intuition suggests that the capacity of the queue is the service rate μ. The authors establish a rather surprising result stating that in some cases the capacity can be higher than μ.

A general upper bound on the capacity of the queue is derived. Special attentions are given to the exponential service time ($\cdot/M/1$ queues), which behaves similarly to the Gaussian noise in the additive-noise channel in two ways. First, there is a closed form expression for the channel capacity. Second, the exponential service time has the *mutual information saddle point* property, i.e. in term of the channel capacity subject to a fixed mean, the exponential probability distribution (PD) yields the best input interarrival time PD, but is also the worst service time PD.

Timing Channel through a $\cdot/G/1$ Queue

The first problem is to find the capacity of a *timing channel*, in which information is transmitted only through the arrival times of message packets, i.e. except for a synchronization packet, each packet contains no information. Assuming no feedback, we can characterize the channel by the conditional joint PD[1]

$$P\{\text{packet departure times}|\text{ packet arrival times}\}.$$

The encoder maps a message into a codeword, which is a sequence of interarrival times. A decoder observes a sequence of interdeparture times and makes a decision. The rate of this code is equal to the logarithm of the number of possible codewords divided by the average transmission time of each codeword.

More specifically, an (n, M, T) code consists of M codewords, each of which is an n-vector of interarrival times $(a_1, ..., a_n)$ such that the n^{th} departure occurs on average within time T. The decoder makes a decision based on the observed n-vector of interdeparture times $(d_1, ..., d_n)$. The rate R of the code is $(\log M)/T$ bits/s. To use an (n, M, T) code, the transmitter sends $n + 1$ packets. The first packet, denoted packet 0, is sent for synchronization at the very beginning of the codeword transmission period. The other n packets are sent based on the n-vector $(a_1, ..., a_n)$ associated with the codeword being transmitted.

Based on the standard definition, a rate R is achievable if there is a sequence of (n, M, T) codes each with rate $\geq R$, and the maximum probability of error $P_e^{\max}$ approaches 0 as T (as well as n and M) goes to infinity. The capacity C of the queue is the supremum of all achievable rates.

A rate R is achievable *at the output rate* λ if there is a sequence of $(n, M, n/\lambda)$ codes each with rate $\geq R$, and $P_e^{\max}$ approaches 0 as n (as well as M) goes to infinity. Let $C(\lambda)$ denote the capacity of the queue *at the output rate* λ, which is the supremum of all achievable rates at the output rate λ. We shall only consider stable queues, and assume that the departure rate is smaller than the service rate, i.e. $\lambda < \mu$.

The definition of $C(\lambda)$ is stricter than that of C since we require the output rate to be equal to λ for every member of the code sequence. Therefore, for any $\lambda < \mu$, we have that $C \geq C(\lambda)$. The authors show that in fact $C = \sup_{\lambda<\mu} C(\lambda)$ (theorem 1). Thus, there is no loss of optimality in working with a sequence of $(n, M, n/\lambda)$ codes instead of a sequence of (n, M, T) codes. A general expression for the upper bound on $C(\lambda)$ is given by

[1]Given the same arrival times, departure times are the same for all nonpreemptive service disciplines such as FIFO and LIFO. Therefore, all nonpreemptive disciplines yield the same capacity of the queue.

Theorem 2: *For a $\cdot/G/1$ queue with the mean service time $1/\mu$, let a RV S denote a service time, if $\lambda < \mu$, then*

$$C(\lambda) \;\le\; \lambda \sup_{W \ge 0,\; E[W] \le \frac{1}{\lambda} - \frac{1}{\mu}} I(W; W+S),$$

where W is independent of S.

The above RV W refers to the idle time RV, i.e. the time period between two arrivals in which the server is idle. When the service time is exponential, the above upper bound in theorem 2 has a closed form expression and is tight. In order to prove so, the following mutual information saddle point property of the exponential PD is useful.

Theorem 3:[2] *Let $a, b \ge 0$ be fixed. Let $\overline{S}$ be exponentially distributed with mean b. Define a RV $\overline{W}$ to be independent of $\overline{S}$ with mean a and the distribution described by the mixture of a point mass and the exponential PD below*[3]

$$\begin{aligned} P\{\overline{W} = 0\} &= \frac{b}{a+b}, \\ P\{\overline{W} > w \mid \overline{W} > 0\} &= e^{-w/(a+b)}. \end{aligned}$$

1. $I(\overline{W}; \overline{W} + \overline{S}) = \log\left(1 + \frac{a}{b}\right)$.

2. *For any nonnegative RVs W (independent of $\overline{S}$) and S (independent of $\overline{W}$) with mean a and b respectively,*

$$I(W; W+\overline{S}) \;\le\; I(\overline{W}; \overline{W}+\overline{S}) \;\le\; I(\overline{W}; \overline{W}+S).$$

As an immediate consequence, for a $\cdot/M/1$ queue, the upper bound on $C(\lambda)$ from theorem 2 is obtained by setting $b = 1/\mu$ and assigning W to have the same PD as that of $\overline{W}$ in theorem 3 with mean $a = 1/\lambda - 1/\mu$. It follows that (theorem 4)

$$\begin{aligned} C(\lambda) &\le \lambda \log \frac{\mu}{\lambda}, \;\; \lambda < \mu, \\ C = \max_{\lambda < \mu} C(\lambda) &\le \frac{\mu}{e} \text{ nats/s}, \end{aligned}$$

[2] To ease the discussion, only parts of the original theorem are stated.

[3] Considering the RV W as idle time, this PD states that between given two arrivals, no idle time occurs with probability $\frac{b}{a+b}$. On the other hand, if the server is left idle, the idle time PD is exponential with mean $a + b$.

where the maximum $C(\lambda)$ is achieved at $\lambda = \mu/e$.

More generally, for a $\cdot/G/1$ queue, let P_S denote the service time PD and e_μ denote the exponential PD with mean $1/\mu$, then we can express the upper bound on $C(\lambda)$ as (theorem 5)

$$C(\lambda) \;\leq\; \lambda \log \frac{\mu}{\lambda} + \lambda D(P_s || e_\mu).$$

Roughly speaking, the more "different" the service time PD is from the exponential PD, the higher the capacity can be. Moreover, the fact that the exponential PD is the worst noise is not surprising, since, for a fixed mean, the exponential PD has the highest differential entropy, i.e. it is the most random noise source.

To show that the upper bound on C for the $\cdot/M/1$ queue is achievable, we can use *random coding*. As a reminder, to use an $(n, M, n/\lambda)$ code, the transmitter sends $n+1$ packets the first of which (packet 0) is used to denote the start of the new codeword. The decoder makes a decision based on n observed interdeparture times. Because of the exponential service time, the maximum likelihood decision rule corresponds to computing the service times from the observed interdeparture times and the arrival times for each candidate codeword, and then selecting the codeword with nonnegative service times and the minimum total service time.

The authors show that for the $\cdot/M/1$ queue, the optimal packet arrivals form a Poisson process of rate λ, i.e. packet interarrival times are IID exponential RVs with mean $1/\lambda$, and the system contains an $M/M/1$ queue.[4] Therefore, to generate M codewords in the codebook, we can generate M independent realizations of the Poisson process with rate μ/e (optimal value for λ from theorem 4) over an interval of T seconds. Equivalently, we can generate a Poisson RV N with mean $\mu T/e$ and then distribute N arrivals uniformly over the interval of T seconds. In conclusion, this coding strategy achieves the upper bound on the capacity of the $\cdot/M/1$ queue in theorem 4, yielding the following result.

[4]Note that the result is consistent with the PD of $\overline{W}$ in theorem 3 and the derivation of theorem 4, since for a particular packet, the probability of having no idle time is

$$P\{\text{an arrival finds the server busy}\} \;=\; \frac{\lambda}{\mu} \;=\; \frac{b}{a+b}, \quad \text{where } a = \frac{1}{\lambda} - \frac{1}{\mu} \text{ and } b = \frac{1}{\mu}.$$

Moreover, by the memoryless property of the exponential PD, given that the server is idle (the next arrival did not occur before the last departure), the PD of the idle time is exponential with mean $1/\lambda = a + b$.

Theorem 6: *For the $\cdot/M/1$ queue with the mean service time $1/\mu$,*

$$
\begin{aligned}
C(\lambda) &= \lambda \log \frac{\mu}{\lambda}, \ \lambda < \mu, \\
C = \max_{\lambda < \mu} C(\lambda) &= \frac{\mu}{e} \text{ nats/s},
\end{aligned}
$$

where the maximum $C(\lambda)$ is achieved at $\lambda = \mu/e$.

The authors show that the same random coding strategy works for a general $\cdot/G/1$ queue. Thus in general, $C(\lambda) \geq \lambda \log(\mu/\lambda), \lambda < \mu$, and $C \geq \mu/e$ nats/s (theorem 7), confirming the fact that the exponential service time is the worst noise.

Telephone Signalling Channel

Imagine trying to send some information for free through a telephone line using the following procedure. The transmitter places the i^{th} call, waits for S_i seconds until the first ring is heard, hangs up, waits for W_{i+1} seconds (where W_{i+1} depends on the message and $S_i, S_{i-1}, ..., S_1$), and then places the $(i+1)^{\text{th}}$ call. The receiver decodes the message based on the observed intervals between calls, denoted by $D_1, D_2, ...$

This *telephone signalling* channel can be considered as an additive-noise memoryless channel with feedback as follow

$$
D_i = W_i + S_i, \quad \text{for all } i.
$$

Using the same definition of the channel capacity as before, denote the capacity of this telephone signalling channel by C_F, and the capacity *at the output rate* λ by $C_F(\lambda)$. Since feedback cannot increase the capacity of a memoryless channel, we have that $C_F(\lambda)$ is equal to the capacity of the additive-noise channel without feedback, which is shown to be (theorem 8)

$$
\begin{aligned}
C_F &= \sup_{\lambda \leq \mu} C_F(\lambda) \\
C_F(\lambda) &= \lambda \sup_{W \geq 0,\ E[W] \leq \frac{1}{\lambda} - \frac{1}{\mu}} I(W; W+S)
\end{aligned}
$$

Intuitively, the capacity achieving waiting time RV is nonnegative, has the mean no greater than $1/\lambda - 1/\mu$ (or else the arrival rate cannot be as high as λ), and maximizes the above mutual information. Since the capacity of the telephone signalling channel has the same form as the capacity of the $\cdot/G/1$ queue, we can use previous results to establish that the capacity C_F of the telephone signalling channel with the mean service time $1/\mu$ is $\geq \mu/e$ nats/s, with equality if and only if the service time is exponential (corollary to theorem 8).

The telephone signalling channel is equivalent, at least in term of its capacity, to the single-server queueing channel with instantaneous and noiseless feedback. This is because the observed packet departures remain the same between any given strategy and its modified version in which packet arrivals are delayed until the queue is empty (possible thanks to instantaneous and noiseless feedback). When queueing is not allowed, the single-server queueing channel becomes an additive-noise memoryless channel with feedback (which we have just discussed). Since $C_F = C$ for an $\cdot/M/1$ queue, we conclude that feedback does not increase the capacity of an $\cdot/M/1$ queue.

Queues with Information Bearing Packets

A straightforward extension of the results allows information to be stored in packets. Let C_0 denote the capacity of the *information channel*, as opposed to the timing channel, in bits or nats per packet transmission. Since errors in packet transmission across the information channel are independent to service times in the queue, we can view the channel as two independent parallel channels. Thus, for a packet arrival rate λ, the capacity is $C(\lambda) = C(\lambda) + \lambda C_0$. It follows that the capacity of the information bearing queue with mean service time $1/\mu$ is (theorem 9)

$$C_I \;=\; \sup_{\lambda \leq \mu} \left[C(\lambda) + \lambda C_0 \right].$$

Applying previous results on the expression of $C(\lambda)$ (theorems 7 and 9), the method of Lagrange multipliers yields the following result.

Theorem 10: *For an information bearing $\cdot/G/1$ queue with the mean service time $1/\mu$,*

$$C_I \;\geq\; \begin{cases} \mu e^{C_0}/e, & 0 \leq C_0 \leq 1 \text{ nat/packet transmission}, \\ \mu C_0, & C_0 \geq 1 \text{ nat/packet transmission}, \end{cases}$$

with equality if and only if the service time is exponential.

Notice that if the information channel capacity C_0 is comparatively high, it is better not to use the timing channel. For the special case of a noiseless DMC with the set of channel alphabets $\mathcal{A}$, the capacity of the information bearing $\cdot/M/1$ queue is (in bits/s)

$$C_I = \begin{cases} \mu/e \ \log_2 e, & |\mathcal{A}| = 1, \\ 2\mu/e \ \log_2 e, & |\mathcal{A}| = 2, \\ \mu \log_2 |\mathcal{A}|, & |\mathcal{A}| > 2. \end{cases}$$

Note that for $|\mathcal{A}| = 2$, the capacity is $2\mu/e \times \log_2 e \approx 1.0615\mu$ bits/sec, which is rather surprising since it is higher than the queue service rate μ. Although in this case the timing channel is worth using when each packet contains at most one bit of information, for other service time PDs, a capacity higher than the service rate may also be achieved for multiple-bit packets.

Related Readings

The idea that the arrival times of information packets contain information was mentioned before in [Gal76], which points out that, in addition to data bits, extra information bits are required to faciliate information exchanges between multiple source and destination pairs. Gallager calls these extra bits *protocol information*.

Analogous results for discrete-time queueing channels are derived in [BA98].

References

[AV96] V. Anantharam and S. Verdu, "Bits Through Queues," *IEEE Transactions on Information Theory*, January 1996.

[Gal76] R.G. Gallager, "Basic Limits on Protocol Information in Data Communication Networks," *IEEE Transactions on Information Theory*, July 1976.

[BA98] A.S. Bedekar and M. Azizoglu, "The Information-Theoretic Capacity of Discrete-Time Queues," *IEEE Transactions on Information Theory*, March 1998.