Bounds and Approximations: heavy-traffic approximation, various bounds on $W(t)$

John C.S. Lui

Department of Computer Science & Engineering
The Chinese University of Hong Kong
www.cse.cuhk.edu.hk/∼cslui

Outline

Outline

Heavy-Traffic Approximation

- Goal: derive the waiting time distribution for $G/G/1$ when $\rho \approx 1$ (but remains less than 1 for stability).
- Recall from $G/G/1$, we have the following result

$$A^*(-s)B^*(s) - 1 = \frac{\Psi_+(s)}{\Psi_-(s)} \tag{1}$$

where

- where for $Re(s) > 0$, $\Psi_+(s)$ must be an analytic function of s that contains no zeros in this half plane,
- for $Re(s) < D$, $\Psi_-(s)$ must be analytic function of s and be zero-free (where $D > 0$),
- require for $|s|$ approaching infinity that $\Psi_+(s) \approx s$ for $Re(s) > 0$ and $\Psi_-(s) \approx -s$ for $Re(s) < D$.

Heavy-Traffic Approximation (continue)

- The spectrum factorization for $\Psi_+(s)$ is

$$\Phi_+(s) = \frac{K}{\Psi_+(s)} \quad \text{where } K = \lim_{s \to 0} \frac{\Psi_+(s)}{s}, \tag{2}$$

  $\Phi_+(s)$ is the Laplace transform of the PDF for $W(y)$ (waiting time).
- Consider the Taylor series expansion of $B^*(s), A^*(-s)$:

$$B^*(s) = \sum_{k=0}^{\infty} \frac{s^k}{k!} B^{*(k)}(0),$$

  and we know that $B^{*(k)}(0) = (-1)^k \overline{x^k}$.

Heavy-Traffic Approximation (continue)

- Using this and considering that $B^*(s)|_{s\to 0}$, we have[a]:

$$B^*(s) = 1 - \overline{x}s + \frac{\overline{x^2}s^2}{2!} + o(s^2).$$

- Similarly

$$A^*(-s) = 1 + \overline{t}s + \frac{\overline{t^2}s^2}{2!} + o(s^2).$$

- We are interested in large waiting times. Large values of y for $W(y)$ is governed by pole of $\Phi_+(s)$ which has the smallest $Re(s)$ in absolute value. Therefore, we need to find zero near $s = 0$.

[a] where $o(x)$ is any function which goes to zero faster than x, or $\lim_{x\to 0}[o(x)/x] = 0$

Heavy-Traffic Approximation (continue)

- Putting the two Taylor's expansions in Eq. (1), we have:

$$\begin{aligned} A^*(-s)B^*(s)-1 &= \left(1-\bar{x}s+\frac{\overline{x}s^2}{2}\right)\left(1+\bar{t}s+\frac{\overline{t}s^2}{2}\right)-1+o(s^2) \\ &= 1+s(\bar{t}-\bar{x})+s^2\left(\frac{\overline{x^2}}{2}+\frac{\overline{t^2}}{2}-\bar{x}\bar{t}\right)-1+o(s^2) \\ &= s\left[\bar{t}-\bar{x}+s\left(\frac{\overline{x^2}}{2}+\frac{\overline{t^2}}{2}-\bar{x}\bar{t}\right)\right]+o(s^2) \end{aligned} \tag{3}$$

- We see that we have a root at $s=0$. To find the 2^{nd} root near $s=0$, we note that:

$$\frac{\overline{x^2}}{2}+\frac{\overline{t^2}}{2}-\bar{x}\bar{t}=\frac{\sigma_b^2+\sigma_a^2}{2}+\frac{(\bar{x}-\bar{t})^2}{2} \tag{4}$$

Heavy-Traffic Approximation (continue)

- Since $\rho \approx 1$, we drop the last term of Eq. (4), or the squared difference of the first moment is negligible compared to the sum of variances.
- To find the 2^{nd} root which we denote as s_0, we have:

$$\bar{t} - \bar{x} + s_0 \frac{\sigma_b^2 + \sigma_a^2}{2} \approx 0$$

which yields

$$s_0 \approx -\frac{2\bar{t}(1-\rho)}{\sigma_a^2 + \sigma_b^2} \tag{5}$$

- Thus, the approximation near the origin is

$$A^*(-s)B^*(s) - 1 = s(s - s_0)\frac{(\sigma_a^2 + \sigma_b^2)}{2}$$

Heavy-Traffic Approximation (continue)

- Doing spectrum factorization, we have $\Psi_+(s) \approx s(s - s_0)C$ where $C = \Psi_-(0)[\sigma_a^2 + \sigma_b^2]/2$.
- To proceed to our solution of $\Psi_+(s)$, we need to find K where $K = \lim_{s \to 0}(s - s_0)C = -s_0 C$, and this yields $\Psi_+(s) \approx \frac{-s_0}{s(s-s_0)}$
- Doing partial fraction expansion, we have

$$\Psi_+(s) \approx \frac{1}{s} - \frac{1}{s - s_0}.$$

- Since $\Psi_+(s)$ is the Laplace transform of $W(y)$ (for $\rho \approx 1$):

$$W(y) \approx 1 - \exp\left(-\frac{2\bar{t}(1-\rho)}{\sigma_a^2 + \sigma_b^2} y\right) \quad (6)$$

$$\overline{W} = \frac{(\sigma_a^2 + \sigma_b^2)}{2(1-\rho)\bar{t}} \quad (7)$$

Outline

Derivation of Upper Bound on $\overline{W}$

- Goal: derive not an approximation, but a firm upper bound on $\overline{W}$.
- Recall $\tilde{u} = \tilde{x} - \tilde{t}$, and we have the following relationship:

$$\tilde{w} = (\tilde{w} + \tilde{u})^+$$

- Assuming the following moments exist, we must have

$$E[(\tilde{w})^k] = E\{[(\tilde{w} + \tilde{u})^+]^k\} \tag{8}$$

- For a random variable X, we introduce the following definition

$$(X)^- = -\min[0, X] \tag{9}$$

Recalling that $(X)^+ = \max[0, X]$, we have the simple relationships

$$X = (X)^+ - (X)^- \tag{10}$$

$$(X)^+(X)^- = 0 \tag{11}$$

Derivation of Upper Bound on $\overline{W}$ (continue)

- Squaring Eq. (10) and using Eq. (11), we have

$$X^2 = [(X)^+]^2 + [(X)^-]^2 \tag{12}$$

- We may form expectations in Eq. (10) to yield:

$$\overline{X} = \overline{(X)^+} - \overline{(X)^-} \tag{13}$$

Likewise, from Eq. (12), we have

$$\overline{X^2} = \overline{[(X)^+]^2} + \overline{[(X)^-]^2}$$

- Since $\sigma_X^2 = \overline{X^2} - (\overline{X})^2$, we use the above relationships to yield:

$$\sigma_X^2 = \sigma_{(X)^+}^2 + \sigma_{(X)^-}^2 + 2\overline{(X)^+(X)^-} \tag{14}$$

and the above equality is true for any random variable X.

Derivation of Upper Bound on $\overline{W}$ (continue)

- Now taking $X = \tilde{w} + \tilde{u}$, we see from Eq. (13) that $\overline{X} = \overline{w} + \overline{u}$ is given by

$$\overline{w} + \overline{u} = \overline{(\tilde{w} + \tilde{u})^+} - \overline{(\tilde{w} + \tilde{u})^-} \tag{15}$$

- However, from Eq (8) (with $k = 1$), we have $\overline{w} = \overline{(\tilde{w} + \tilde{u})^+}$, and so Eq. (15) can be rewritten as

$$\overline{u} = -\overline{(\tilde{w} + \tilde{u})^-}$$

- Furthermore, from Eq. (8), we have that

$$\sigma^2_{\tilde{w}} = \sigma^2_{(\tilde{w}+\tilde{u})^+} \tag{16}$$

Derivation of Upper Bound on $\overline{W}$ (continue)

- Once again, taking $X = \tilde{w} + \tilde{u}$, we see the term $\sigma^2_{(X)^+}$ from Eq. (14) is equal to $\sigma^2_{\tilde{w}}$ due to relationship in Eq. (16). Since $\tilde{w}$ and $\tilde{u}$ are independent, we have $\sigma^2_{(\tilde{w}+\tilde{u})} = \sigma^2_{\tilde{w}} + \sigma^2_{\tilde{u}}$, and so Eq. (14) finally takes the form

$$\sigma^2_{\tilde{w}} + \sigma^2_{\tilde{u}} = \sigma^2_{\tilde{w}} + \sigma^2_{(X)^-} + 2\overline{(\tilde{w} + \tilde{u})^+(\tilde{w} + \tilde{u})^-} \tag{17}$$

- For the last term above, we already established that $\overline{(\tilde{w} + \tilde{u})^+} = \overline{w}$ and $\overline{(\tilde{w} + \tilde{u})^-} = -\overline{u}$; using this and canceling the variance of $\tilde{w}$ from both sides of the last equation, we have

$$\sigma^2_{\tilde{u}} = \sigma^2_{(X)^-} - 2\overline{w}\,\overline{u} \tag{18}$$

Derivation of Upper Bound on $\overline{W}$ (continue)

- By definition, $\tilde{u} = \tilde{x} - \tilde{t}$ and $\overline{u} = \overline{t}(\rho - 1)$. Since $\tilde{x}$ and $\tilde{t}$ are independent, it must be that $\sigma^2_{\tilde{u}} = \sigma^2_{\tilde{t}} + \sigma^2_{\tilde{x}} = \sigma^2_{\tilde{a}} + \sigma^2_{\tilde{b}}$. Now we can solve for $\overline{w}$ (which is denoted as $\overline{W}$) in Eq. (18) as:

$$W = \frac{\sigma_a^2 + \sigma_b^2}{2\overline{t}(1-\rho)} - \frac{\sigma^2_{(X)^-}}{2\overline{t}(1-\rho)}$$

- Since variance is always non-negative, we drop the last term of the above equation to create an upper bound:

$$W \leq \frac{\sigma_a^2 + \sigma_b^2}{2\overline{t}(1-\rho)} \qquad \text{for } 0 \leq \rho < 1. \tag{19}$$

It means the heavy-traffic approximation forms a *strict upper bound* on W for $G/G/1$.

Outline

Deriving the tail of the waiting time

- Waiting time of an arriving customer is equal to the service time of all customers he finds in the queue upon his arrival, plus the residual service time of customer in the service center.
- Our working equation $w_{n+1} = \max[0, w_n + u_n]$. For $y \geq 0$, we may write

$$P[w_{n+1} \geq y] = P[w_n + u_n \geq y]$$

- Conditioning on the value of v_n and $P[w_n \geq 0] = 1$, we have

$$\begin{aligned} P[w_{n+1} \geq y] &= \int_{-\infty}^{\infty} P[w_n \geq y - u] dC(u) \\ &= \int_{-\infty}^{y} P[w_n \geq y - u] dC(u) + 1 - C(y). \quad (20) \end{aligned}$$

Derivation (continue)

- Consider $C^*(-s) = E[e^{su_n}]$ where s is taken to be a real (rather than a complex) variable.
- We know that s must lie in a restricted range if this transform is to remain bounded. In particular, if there exists a real s' such that $B^*(-s) = E[e^{s'\tilde{x}}] < \infty$, then a permissible range for s is $0 \leq s \leq s'$.
- Furthermore, there will be a range in which $C^*(-s) \leq 1$. For example, in this stable case, $C^*(0) = 1$ and for $s = 0$, $dC^*(-s)/ds = \overline{u} < 0$. Thus identifying a neighborhood in this range.
- We let s_0 denote the largest value for s such that this remains true.

Derivation (continue)

- We can now express

$$e^{-s_0 y} \geq e^{-s_0 y} C^*(-s_0) = e^{-s_0 y} \int_{-\infty}^{\infty} e^{s_0 u} dC(u) = \int_{-\infty}^{\infty} e^{-s_0 (y-u)} dC(u) \tag{21}$$

- Since $s_0 > 0$, for the range $u \geq y$, it must be $e^{-s_0(y-u)}] \geq 1$, we can express Eq. (21) as

$$\begin{aligned} e^{-s_0 y} &\geq \int_{-\infty}^{\infty} e^{-s_0(y-u)} dC(u) + \int_{y}^{\infty} dC(u) \\ &= \int_{-\infty}^{y} e^{-s_0(y-u)} dC(u) + 1 - C(y) \quad \text{for } y > 0 \end{aligned} \tag{22}$$

Derivation (continue)

- Let us assume w_0 (an *initial* customer's waiting time) is chosen so that $P[w_0 \geq y] \leq e^{-s_0 y}$. We wish to prove that this hypothesis carries over for all w_n. We prove this by induction.
- Assume that it is true for n, or $P[w_n \geq y] \leq e^{-s_0 y}$, then

$$P[w_{n+1} \leq y] \leq \int_{-\infty}^{\infty} e^{-s_0(y-u)} dC(u) + 1 - C(y)$$

- But this right-hand side is exactly the expression we bounded in Eq. (22). So we have $P[w_{n+1} \geq y] \leq e^{-s_0 y}$.
- We have established the following exponential bound on the tail of the waiting time distribution (by letter $n \to \infty$):

$$P[\tilde{w} \geq y] \leq e^{-s_0 y} \tag{23}$$

where s_0 can be found from $s_0 = \sup\{s > 0 : C^*(-s) \leq 1\}$

Derivation (continue)

- It is possible to prove that this tail has a lower bound, with combined result of

$$\gamma e^{-s_0 y} \leq 1 - W(y) \leq e^{-s_0 y} \tag{24}$$

where $W(y) = P[\tilde{w} \leq y]$ and γ must satisfy:

$$\gamma \leq \frac{1 - C(y)}{\int_y^\infty e^{-s_0(y-u)} dC(u)} \quad \text{for } y > 0 \tag{25}$$

Therefore, γ is the smallest value that the ratio of the above equation takes on.

- From these bounds on the distribution function, it is trivial to show the bounds on the mean waiting time as

$$\frac{\gamma}{s_0} \leq W \leq \frac{1}{s_0} \tag{26}$$

Outline

Discrete $G/G/1$

- Instead of finding an approximation to our problem, we attempt an exact solution for an approximation of the problem.
- For $G/G/1$, we have the following recurrence relationship:

$$w_{n+1} = \max[0, w_n + u_n] \tag{27}$$

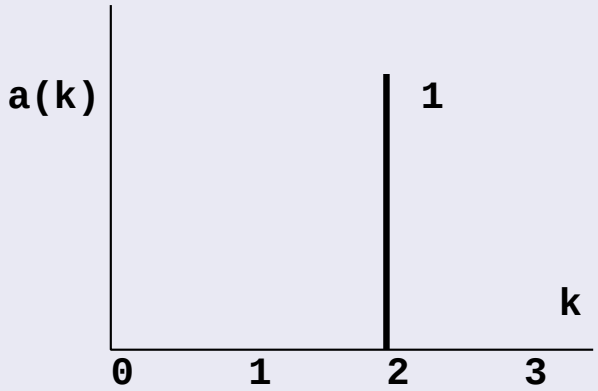
- When the interarrival time and service time are both discrete random variables, whose only nonzero values occur at the instants $k\tau$ ($k = 0, 1, \ldots$) and τ is a basic time unit, the iterative application of the above equation is quite simple.
- Let us assume that we can approximate the given continuous interarrival time and service time to discrete random variables (this requires some thinking). We illustrate this method via an example.

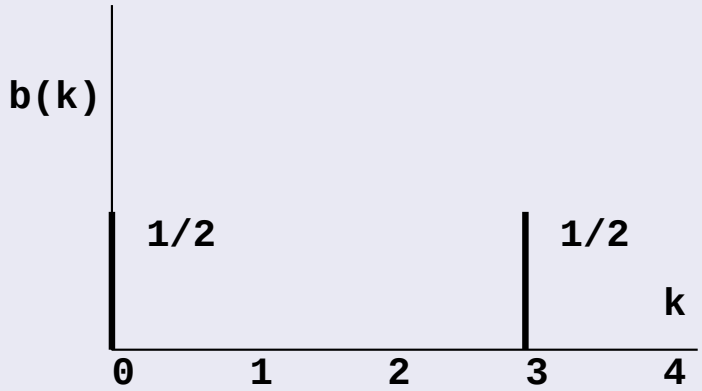
Discrete $G/G/1$: example

- Consider the following discrete random variables (CDF):

$$A(t) = \begin{cases} 0 & t < 2\tau \\ 1 & t \geq 2\tau \end{cases} \quad ; \quad B(x) = \begin{cases} 0 & x < 0 \\ 1/2 & 0 \leq x < 3\tau \\ 1 & 3\tau \leq x \end{cases}$$

- Illustrate pdfs of $a(t)$ (for $A(t)$), and $b(x)$ (for $B(x)$).





- The average interarrival time is 2τ, average service time is $1/2 \times 0 + 1/2 \times 3\tau = 1.5\tau$. So $\rho = \frac{1.5\tau}{2\tau} = 0.75 < 1$, so the system is stable.

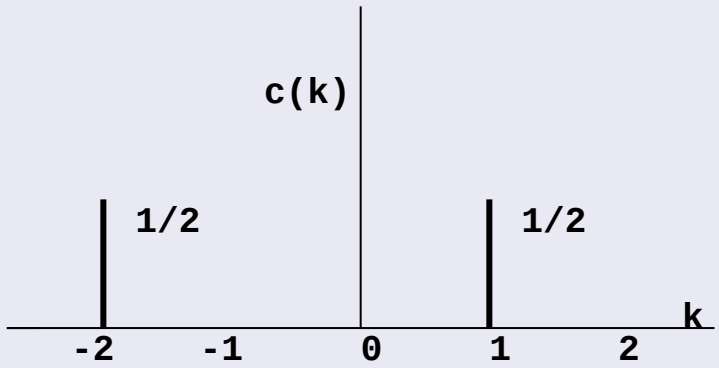
Example (continue)

- Define the pdf of u_n: $c(k) = P[u_n = k\tau]$, since $u_n = x_n - t_{n+1}$:

$$c(k) = a(-k) \otimes b(k) = \sum_{i=-\infty}^{\infty} a(-k+i)b(i).$$

Carrying out this convolution, we have

$$c(k) = \begin{cases} 1/2 & k = -2 \\ 1/2 & k = 1 \\ 0 & \text{otherwise} \end{cases} \tag{28}$$



Example (continue)

- Assume the *initial waiting time* in the system is $w_0 = 0$, now we can apply the recursion.
- Define $p_n(k) = P[w_n = k\tau]$ and apply Eq. (27).
- Procedure is:
  1. Assume we have $p_n(k)$.
  2. We need to find the pdf of $w_n + u_n$, which means we need to *convolve* $p_n(k)$ and $c(k)$.

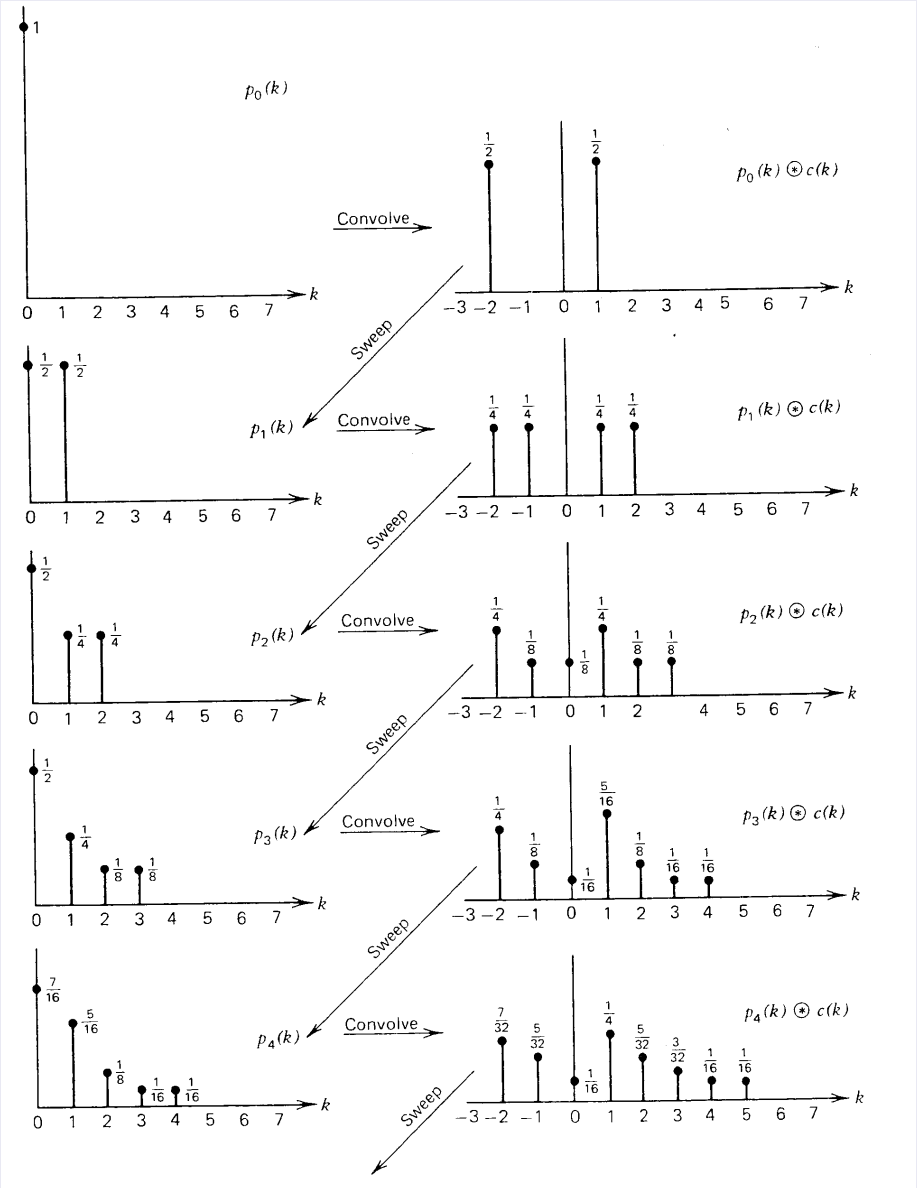
$$p_n(k) \otimes c(k).$$

  3. Then we apply:

$$\pi(p_n(k) \otimes c(k),$$

     where π means sweeping the probability in the negative half-line up to the origin, this gives $p_{n+1}(k)$.
  4. Repeat until converges, or $p(k) = \lim_{n\to\infty} p_n(k)$. (or $p_{n+1}(k) = p_n(k)$).

Example (illustration)



Comment

- The above is really a *numerical method*.
- The important thing is how to approximate the continuous random variables of $A(t)$ and $B(x)$ to discrete random variables.
- One may consider ways to match the first k moments as "approximation".