e-Companion

11. Introduction

We want to emphasize that the main paper addresses a challenging problem. Setting aside purely psychological effects, our focus in this paper is on the possible impact of delay announcements on *performance*, through their effect on customer balking (leaving immediately upon arrival) and abandonment (leaving after waiting in queue) decisions. A full treatment of this problem must include several challenging modelling and analysis elements, which include the following:

1. The announcement scheme: What information is provided to waiting customers – and when? What is the format - waiting time estimates, or number of customers in queue? How are waiting time estimates computed? Are they dynamic (customer specific) or static (same to all, based on average queue conditions)?

2. Customer reaction modelling: How does the provided information modify the customer balking and abandonment behavior?

3. Queueing analysis: Exact analysis of relevant performance metrics which takes into account the customer reaction to announced information may become very complex. For example, a state-dependent announcement scheme inevitably introduces state-dependent and correlated abandonment profiles. Appropriate approximations are required in such cases.

4. Equilibrium analysis: The inter-dependence of announced information (and hence customer decisions) on the system performance on the one hand, and of the system performance on customer decisions on the other hand, requires the application of equilibrium equilibrium (or fixed-point) analysis to obtain the actual working point of the system. This adds to the challenge of the overall system analysis.

It is evident that each of these items, let alone their combination, should be the subject of several papers. The present paper provides first steps towards addressing these issues, and their combination within an integrative model.

Organization. This e-companion has six more sections. In §12 we state two basic comparison results for the fluid model, extending §4. In §13 we use the fluid model to study the impact of biased delay announcements, where the announcement is designed to differ from the actual delay. In §14 we briefly discuss the consequence of increasing patience in response to delay announcements. In §15 we discuss iterative techniques to determine the equilibrium delay for the fluid model. In §16 we extend the perturbation analysis in §7 by carrying out perturbation analysis for the general all-exponential model in (5.1) having general functions $\gamma(w)$ and $\delta(w)$. In §17 we display simulated queue-length sample paths associated with DLS announcements for the example in §8 of the main paper with multiple fluid equilibria. The simulations support the conjecture that, unlike for the fluid model, there is a unique limiting steady-state distribution for the queueing model with DLS announcements, independent of the initial conditions. Additional supporting material appears in Armony et al. (2007).

12. Comparisons in the Fluid Model

It is natural to wonder how the equilibrium fluid delay depends on the model elements. The following comparison result provides a partial answer.

Theorem 12.1. (comparison) *Consider two fluid models of the kind specified in* §§3 and 4*, satisfying Condition 4.1. The corresponding equilibria are ordered by* $\tilde{w}_{e,1} < \tilde{w}_{e,2}$ *if and only if* $\rho B_1^c(\tilde{w}_{e,2}) F_1^c(\tilde{w}_{e,2}|\tilde{w}_{e,2}) < 1$*, which in turn holds if and only if* $\rho B_2^c(\tilde{w}_{e,1}) F_2^c(\tilde{w}_{e,1}|\tilde{w}_{e,1}) > 1$*. A sufficient condition is* $\rho B_1^c(w) F_1^c(w|w) < \rho B_2^c(w) F_2^c(w|w)$ *for all* $w > 0$*.*

Proof. Immediate by noting that $\tilde{w}_{e,1}$ and $\tilde{w}_{e,2}$ both satisfy equations of the form (4.2), combined with the assumed monotonicity of $\rho B^c(w) F^c(w|w)$. ∎

A fundamental question is whether or not a delay announcement reduces delays. The following theorem provides conditions for this to be true in the fluid model context.

Theorem 12.2. (conditions for an announcement to reduce delays) *Consider a fluid model satisfying Condition 4.1. We have the ordering* $0 < \tilde{w}_e < \tilde{w}_1$*, where* $\tilde{w}_1$ *is the delay without making an announcement, satisfying* (3.2)*, if and only if* $\rho B^c(\tilde{w}_1) F^c(\tilde{w}_1|\tilde{w}_1) < 1$*, which in turn holds if and only if there exists a* w *with* $0 < w < \tilde{w}_1$ *such that* $\rho B^c(w) F^c(w|w) < 1$*. A sufficient condition is* $\rho B^c(w) F^c(w|w) < \rho F^c(w)$ *for all* $0 \le w \le \tilde{w}_1$*.*

Proof. The first two properties follow directly from the strict monotonicity of $\rho B^c(w) F^c(w|w)$. The sufficient condition follows since $\tilde{w}_1$ is characterized by $\rho F^c(\tilde{w}_1) = 1$. ∎

13. Biased Announcements

We now consider how the fluid model can be used to gain additional insight. In the main paper, we have assumed that, in equilibrium, the anticipated delay announced by the system should be equal to the actual one. In this section we consider an alternative, allowing the announced delay to be larger or smaller than the actual one. This may arise for two reasons: (i) because the system manager purposely chooses to bias the announcements to affect some performance measures of the system, or (ii) because there are inaccurate delay estimates. Announcing a larger delay might reduce the system load and thereby reduce the delay for later customers. Announcing a smaller delay might reduce abandonment. In either case, the deviation from the actual delay should not be too large; otherwise customers may lose confidence in the announced delays. However, a moderate deviation over limited time periods should go unnoticed.

To compare different announcement options, we will make some specific assumptions about the form of the abandonment distribution. The following parallels the notion of information-consistent balking, defined in Definition 3.1.

Definition 13.1. (information-consistent abandonment) *A conditional patience distribution specified by $\rho B^c(w)$ and $F^c(t|w)$ is* ***information consistent*** *if*

$$\rho B^c(w)F^c(t|w) = \rho F^c(w) \quad \text{for} \quad t \le w \quad \textit{and} \quad \rho B^c(w)F^c(t|w) = \rho F^c(t) \quad \text{for} \quad t > w\,. \qquad (13.1)$$

Definition 13.1 requires that, upon hearing a delay announcement of w, all customers who intend to wait no more than w respond by balking (abandoning immediately), while those who intended to wait more than w in the first place are not affected by the announcement; see Figure 4. In particular, this implies that $F^c(t|w) = 1$ for $t \le w$; i.e., there is balking at time 0, but no abandonment at all occurs before w. Since this requirement is somewhat extreme, we also make the following definition.

Definition 13.2. (weak information consistency) *A conditional patience distribution specified by $\rho B^c(w)$ and $F^c(t|w)$ is* ***weakly information consistent*** *if*

$$\rho F^c(t) \ge \rho B^c(w)F^c(t|w) \ge \rho F^c(w) \quad \textit{for} \quad 0 \le t \le w\ ; \qquad (13.2)$$

$$\rho B^c(w)F^c(t|w) \le \rho F^c(t) \quad \textit{for} \quad t > w\ . \qquad (13.3)$$

These definitions are illustrated in Figure 4. For $t \le w$, weak information consistency is a middle ground between the original patience and information-consistent abandonment. It accommodates, for example, a mixture of information-consistent customers with others who are not affected at all by the announcement, or whose abandonment is caused by exogenous events. For $t > w$, customers may be frustrated by the fact that the announced wait was not satisfied, leading to loss of patience and a larger rate of abandonment. Hence the second condition in (13.1) is relaxed to (13.3).

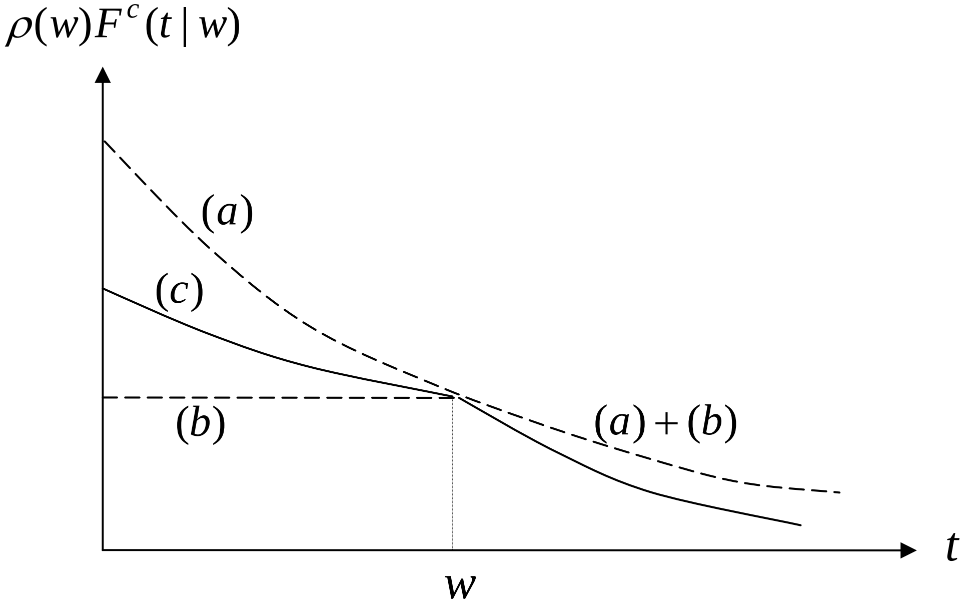

Figure 4: Three cases for the possible effect of a delay announcement on the abandonment distribution: (a) Patience profile without information (b) Information-consistent abandonment (c) Weakly information-consistent abandonment.

For the all-exponential model with constant abandonment rate γ and δ, information consistency is equivalent to $\beta = \theta$, $\gamma = 0$ and $\delta = \theta$, while weak consistency is equivalent to $\beta + \gamma = \theta$ and $\delta \geq \theta$.

We now consider the issue of biased announcements. Let the announced delay be $w_a = w + \Delta$, where w is the actual delay and Δ is a fixed **additive bias**, which may be positive, negative or zero.

Definition 13.3. (equilibrium delay with bias) *A delay w is an* ***equilibrium delay with bias*** *for the fluid model with a fixed* ***additive bias*** *Δ and associated announcement $w + \Delta$ if $d(w + \Delta) = w$, where d is the response delay function in Definition 3.2; i.e., w is an equilibrium delay with bias for the fixed additive bias Δ if either (i) $\rho B^c(\Delta) \leq 1$ and $w = 0$ or (ii) $\rho B^c(\Delta) > 1$ and*

$$\rho B^c(w+\Delta)F^c(w|w+\Delta) = 1 \quad \textit{and} \quad \rho B^c(w+\Delta)F^c(t|w+\Delta) > 1 \quad \textit{for} \quad 0 \leq t < w \ . \quad (13.4)$$

It is readily verified that existence and uniqueness of the equilibrium delay with bias hold under Condition 4.1. In particular, these conditions are satisfied for our simple all-exponential model with $\beta > 0$ and $\gamma > 0$, which we henceforth consider. The equilibrium biased delay w that corresponds to a fixed bias Δ will be denoted by $\tilde{w}_{e,\Delta}$. Hence, $\tilde{w}_{e,0}$ corresponds to the equilibrium fluid delay without bias, $\tilde{w}_e$, considered before.

Consider first the case of $\Delta > 0$. Assume that $\rho e^{-\beta\Delta} > 1$, so that $\tilde{w}_{e,\Delta} > 0$. For $\Delta > 0$ we have $F^c(w|w+\Delta) = e^{-\gamma w}$, and

$$\tilde{w}_{e,\Delta} = \frac{\log\rho - \beta\Delta}{\beta+\gamma} = \tilde{w}_{e,0} - \frac{\beta}{\beta+\gamma}\Delta \ . \tag{13.5}$$

Thus, a positive bias Δ in the announced delay reduces the actual wait by a fraction of this bias. This fraction will be closer to unity when β is large relative to γ. In particular, for information-consistent abandonment (where $\gamma = 0$), we obtain $\tilde{w}_{e,\Delta} = \tilde{w}_{e,0} - \Delta$. The situation is illustrated in Figure 5.

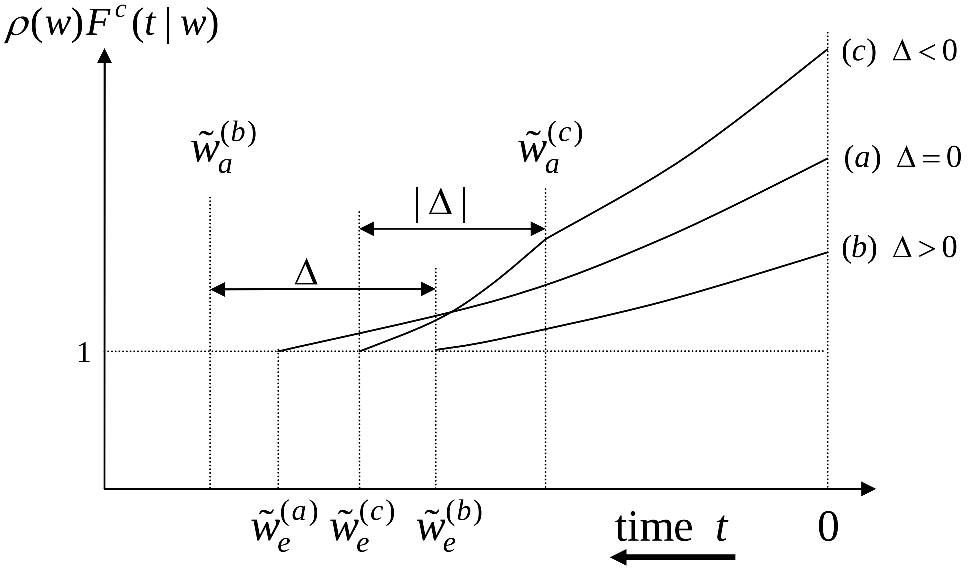

Figure 5: The effect of positive and negative bias in the announced delay on the equilibrium point.

In contrast, for $\Delta < 0$, $F^c(w|w+\Delta) = e^{-\gamma(w-|\Delta|)}e^{-\delta|\Delta|}$, so that

$$\tilde{w}_{e,\Delta} = \tilde{w}_{e,0} - \frac{\delta-(\beta+\gamma)}{\beta+\gamma}|\Delta|\,. \tag{13.6}$$

Interestingly, for information-consistent abandonment, $\delta = \beta+\gamma$, so that $\tilde{w}_{e,\Delta} = \tilde{w}_{e,0}$. For weak consistency, $\delta \geq \beta+\gamma$, so that $\tilde{w}_{e,\Delta} \leq \tilde{w}_{e,0}$. Thus the delay may decrease even in the case of negatively biased announcements.

14. Increasing Patience

So far, our analysis has shown how waiting decreases in response to delay announcements, compared to making no announcement at all. However, that analysis is based on the assumption that, with a delay announcement of w, customers are more likely to instantly balk upon arrival and then later abandon at a higher rate after time w. However, customer behavior might actually be different. Indeed, data analysis from a banking call center, related to Brown (2005), indicates that customer patience may increase in response to a delay announcement, even after the announced delay time; Feigin (2006).

Upon reflection, this customer behavior is intuitively reasonable, because the delay announcement may serve to reduce the customer's sense of uncertainty and ambiguity. The customer may be willing to wait provided he understands the situation. The delay announcements may improve the customer's feelings about the contact center. At any rate, it is interesting to consider the consequences of increasing patience in response to a delay announcement in system overload. The delay announcement may actually increase the overload. In this section we point out that the fluid model can be used to quantify that phenomenon, provided that we are able to quantify customer behavior..

For simplicity, we consider the all-exponential model. In that context, the key is to recognize that the abandonment rate θ with no announcement may actually exceed the balking and abandonment rates β, γ and δ. Given larger θ, these parameters might naturally be ordered as

$$\theta > \delta > \beta > \gamma \ .$$

Now we no longer have $\tilde{w}_e < \tilde{w}_1$. Indeed, the inequality is reversed. Thus, starting from no announcement, the iterations increase from $\tilde{w}_1$ to $\tilde{w}_e$.

15. Iterations and Convergence

In this section, we investigate iterative schemes for the fluid model. In particular, we assume that we make an initial delay announcement w_0. Then we observe the actual steady-state delay w_1 of those customers who are served. We then make the latter our delay announcement, and see the actual steady-state delay w_2 of those customers served with announcement w_1. We continue in this way, looking at the actual steady-state delay w_{k+1} of those customers served, given delay announcement w_k, for $k \geq 0$. This iteration scheme is a natural way to compute the equilibrium delay, but it does not actually correspond to a natural evolution of the system over time, unless there is a substantial time between successive iteration steps. Otherwise, the system would not be able to reach steady state before adjustment.

There are some difficulties in general: First, we may have $d(w) > \tilde{w}_e$ for $w < \tilde{w}_e$ and, in the event that occurs, we may fail to have strict monotonicity of the two-stage iteration $d^{(2)}(w) \equiv d(d(w))$, and consequently the announced delay sequence might oscillate or diverge. We illustrate by next giving an example in which the two-stage iteration operator $d^{(2)}(w)$ has multiple fixed points.

Example 15.1. (multiple fixed points for the two-stage iteration operator) This example shows cycling around the equilibrium delay $\tilde{w}_e$ instead of convergence to it. In particular, we show that the two-stage iteration operator $d^{(2)}(w) \equiv d(d(w))$ has multiple fixed points. This example uses linear functions with slope -1. In particular, $\rho F^c(t) = \rho B^c(t) = \rho - t$ for $0 \le t \le \rho$. We also have $\rho B^c(w) F^c(t|w) = \rho - 2t$. The cyclic behavior is shown in Figure 6. In this example, $w_{2k} = w_0$ and $w_{2k-1} = w_1 = d(w_0)$ for all $k \ge 1$. Such cycling will occur in this linear example (with lines of slope -1) for each announced delay w with $0 < w < \tilde{w}$ except for the equilibrium delay $\tilde{w}_e = \tilde{w}/2$, where here $\tilde{w}$ without subscript denotes the delay without an announcement.

We now consider iteration and convergence for the all-exponential model, stating the result without proof. We show that there is bad oscillating behavior when $\delta = \gamma < \beta$ in part (b). let $[x]^+ \equiv \max\{x, 0\}$.

Theorem 15.1. (iteration and convergence for the all-exponential model) *Consider the fluid model associated with the simple all-exponential model in §5 of the main paper.*

(*a*) *Assume that $\delta \ge \beta + \gamma$. Then the delay associated with announcement w is*

$$d(w) = \frac{\log \rho + (\delta - \gamma - \beta) w}{\delta} \quad \text{for} \quad 0 < w \le \tilde{w}_e \ , \tag{15.1}$$

which has the property that $w < d(w) < \tilde{w}_e = \log \rho/(\beta + \gamma)$, while $d(w) = 0$ for $w \ge \tilde{w}$, $0 < d(w) < \tilde{w}_e$ for $\tilde{w}_e \le w \le \tilde{w}$ and $d(0) = \tilde{w}_1 = \log \rho/\delta < \tilde{w}_e$. As a consequence,

$$\tilde{w}_e > w_{k+1} \equiv d(w_k) > w_k \ge \tilde{w}_1 > 0 \quad \text{for all} \quad k \ge 2 \tag{15.2}$$

and $w_k \equiv d^{(k)}(w_0) \to \tilde{w}_e$ as $k \to \infty$.

(*b*) *Assume that $\delta = \gamma < \beta$. Then*

$$d(w) = \left[\frac{\log(\rho)}{\gamma} - \left(\frac{\beta}{\gamma}\right) w\right]^+ = \left[\tilde{w}_e - \left(\frac{\beta}{\gamma}\right)(w - \tilde{w}_e)\right]^+ \ , \tag{15.3}$$

Cycling Around the Equilibrium Delay

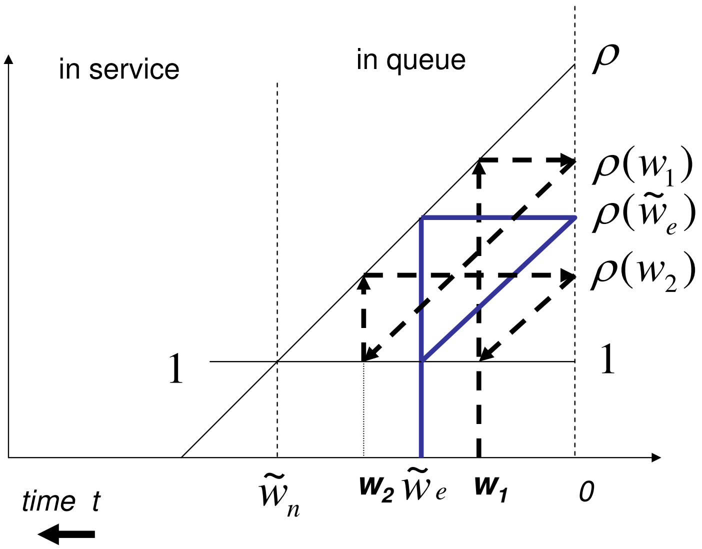

Figure 6: Cycling around the equilibrium delay: The announced delay becomes the actual delay every other iteration. The delay without an announcement is $\tilde{w}$.

$$d^{(2n)}(w) = \begin{cases} 0, & w \leq \tilde{w}_e(1 - (\gamma/\beta)^{2n}) \ , \\ \frac{\log(\rho)}{\gamma}, & w \geq \tilde{w}_e(1 + (\gamma/\beta)^{2n-1}) \ , \\ \tilde{w}_e + (w - \tilde{w}_e)\left(\frac{\beta}{\gamma}\right)^{2n}, & t > w \ , \end{cases} \tag{15.4}$$

Consequently, for all $w < \tilde{w}_e$,

$$d^{(2n)}(w) = 0 \quad \textit{and} \quad d^{(2n+1)}(w) = \frac{\log(\rho)}{\gamma} \tag{15.5}$$

for all n sufficiently large; for all $w > \tilde{w}_e$,

$$d^{(2n+1)}(w) = 0 \quad \textit{and} \quad d^{(2n)}(w) = \frac{\log(\rho)}{\gamma} \tag{15.6}$$

for all n sufficiently large.

In order to avoid oscillations, it may be desirable to use a **damped iteration**. We can let the successive announced delays be defined recursively by

$$w_{k+1} = pd(w_k) + (1-p)w_k = w_k + p(d(w_k) - w_k) \tag{15.7}$$

for some constant p with $0 < p \leq 1$. In Armony et al. (2007) we establish convergence results for damped iterations under the condition that p be small enough. We also establish other results about iteration there.

16. Perturbation Analysis for More General All-Exponential Models

In this section we carry out perturbation analysis for the more general all-exponential model in (5.1) of the main paper, i.e., with

$$F^c(t|w) = \begin{cases} e^{-\gamma(w)t}, & 0 \leq t \leq w \ , \\ e^{-\gamma(w)w}e^{-\delta(w)(t-w)}, & t > w \ , \end{cases} \tag{16.1}$$

where $\gamma(w)$ and $\delta(w)$ are two component abandonment-rate functions, assumed to be positive and nondecreasing in the announced delay w, which was defined in Section 5 of the main paper.

We see that the fluid model should perform well whenever $\delta(w) = \gamma(w)$, even if these functions are not constant. We now consider that specific case here. For that model, we determine the response to an announcement $\tilde{w}_e + \epsilon$. Since $\delta(w) = \gamma(w)$, the response does not depend upon the sign of ϵ. We also let the balking rate depend on w, so we have the function $\beta(w)$. We assume that the functions $\gamma(w)$ and $\beta(w)$ both are smooth having three continuous derivatives.

Let

$$\gamma^{(k)} \equiv \gamma^{(k)}(\tilde{w}_e) \equiv \frac{d^k}{dw^k}\gamma(w)|_{w=\tilde{w}_e} \tag{16.2}$$

and

$$\beta^{(k)} \equiv \beta^{(k)}(\tilde{w}_e) \equiv \frac{d^k}{dw^k}\beta(w)|_{w=\tilde{w}_e} \ . \tag{16.3}$$

Then it is elementary to see that

$$d(\tilde{w}_e + \epsilon) = \tilde{w}_e - A\epsilon + B\epsilon^2 + O(\epsilon^3) \quad \text{as} \quad \epsilon \downarrow 0 \ , \tag{16.4}$$

where

$$A \equiv A(\beta, \gamma, \tilde{w}_e) \equiv \frac{\beta(\tilde{w}_e) + (\beta^{(1)} + \gamma^{(1)})\tilde{w}_e}{\gamma(\tilde{w}_e)} \tag{16.5}$$

and

$$B \equiv B(\beta, \gamma, \tilde{w}_e) \equiv \frac{(\beta(\tilde{w}_e)\gamma^{(1)} - \gamma(\tilde{w}_e)\beta^{(1)}) + \tilde{w}_e[\gamma^{(1)}(\beta^{(1)} + \gamma^{(1)}) - \gamma(\tilde{w}_e)(\gamma^{(2)} + \beta^{(2)})/2]}{\gamma(\tilde{w}_e)^2} \ . \tag{16.6}$$

Thus, if the actual DLS delay is distributed as $N(\tilde{w}_e, \sigma_e^2)$, then

$$E[d(N(\tilde{w}_e, \sigma_e^2))] \approx \tilde{w}_e + B\sigma_e^2 \tag{16.7}$$

for B in (16.6). Since we have assumed that $\delta(w) = \gamma(w)$, it is reasonable to expect that the actual equilibrium delay will indeed be approximately normally distributed with mean $\tilde{w}_e$. We thus get a refined analysis of the impact of stochastic fluctuations in the "symmetric" case. We also illustrate how to proceed in other models.

17. Multiple Equilibria

In this section we complement §8 in the main paper by displaying sample paths of the queue-length stochastic process using DLS announcements with the nonlinear abandonment rate function

$$\gamma(w) = \begin{cases} 4.0, & 0 \le w < 0.10 \ , \\ 7.5 - 35w & 0.10 \le w < 0.20 \ , \\ 0.5, & t > 0.20 \ . \end{cases} \tag{17.1}$$

We have constructed $\gamma(w)$ to be constant over the two subintervals $[0, 0.10)$ and $[0.20, \infty)$, linear and decreasing in the interval $[0.10, 0.20)$ and continuous overall. It is elementary to see that the fluid model has three equilibria, with one in each region: The three fluid equilibria are $\tilde{w}_e = 0.0672$, $\tilde{w}_e = 0.193$ and $\tilde{w}_e = 0.224$. The abandonment rates at these three equilibria are, respectively, $\gamma(0.0672) = 4.0$, $\gamma(0.193) = 0.7395$ and $\gamma(0.224) = 0.5$. The associated fluid queue contents are $q(0.672) = 0.077$, $q(0.193) = 0.180$ and $q(0.224) = 0.237$. One may multiply by $s = 100$ to get the associated approximating queue lengths.

Here we display the sample path of the queue-length process estimated from simulation using DLS announcements. First, in Figure 7 we display a queue-length sample path for the abandonment-rate function in (17.1). Then in Figures 8 and 9 we plot this sample path again together with a sample path of the queue length process when $\gamma(w)$ is constant, first at 0.5 and then at 4.0.

For this example, the simulation supports our conjecture that there should be a well-defined unique steady-state for the DLS announcements, even though there are three separate equilibria for the associated fluid model. We see that the sample path of the queue-length process visits the regions of both of the queue-length processes with fixed delay announcements, without getting stuck in either.

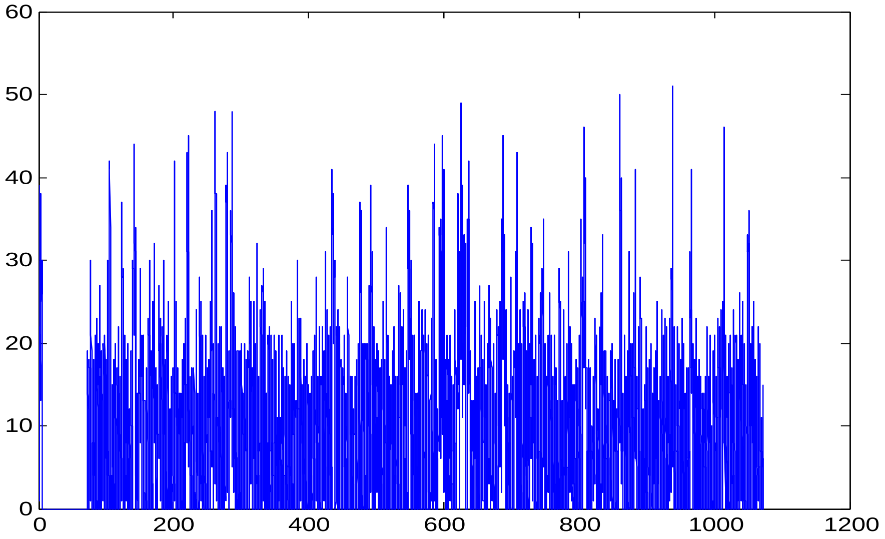

Figure 7: A sample path of the queue-length process for the all-exponential model with $\delta(w) = \gamma(w)$ for all w, with the nonlinear $\gamma(w)$ in (17.1).