Principles of Bounded Deferral for Balancing Information Awareness with Interruption

Dimitris Achloptas
Microsoft Research
Redmond, WA 98052
optas@microsoft.com

Eric Horvitz
Microsoft Research
Redmond, WA 98052
horvitz@microsoft.com

Abstract

We present a study of *bounded deferral*, notification policies that defer incoming alerts for a bounded time if a user is busy when alerts arrive. We first review empirical studies that highlight the value of pursuing a deeper understanding of bounded-deferral policies. Then, we present a general decision-theoretic formulation of bounded deferral. We introduce families of functions that are expressive yet enable optimization of deferral times based on the outcome of simple tests.

1 Introduction

Interest has been growing over the last several years on methods for protecting the attentional focus of computer users, by reasoning about the costs and benefits of relaying alerts and the current state of a user's workload. Efforts have focused on the creation of models that can be used to estimate the cognitive load of users [6, 7], and the construction of real-time reasoning platforms for guiding notifications [5]. In user studies, researchers have elucidated the effects of interrupting people in various ways in different situations and have probed the workload and availability of people in office settings user studies of the cost of task switching and disruption [1, 2, 8, 3, 9, 10].

In this paper, we will explore principles and applications of a family of notification policies, referred to as *bounded deferral* [5]. The method hinges on a key concept: A person who is too busy to review a message or information update when it arrives, will likely transition to an available state in time, as described by a probability distribution that can be learned. With a bounded deferral policy guiding notification actions, a notification manager will simply wait until either the user transitions to an available state, or a maximum wait time is reached–the deferral bound–at which point the user is alerted.

As we shall later explore in detail, if a transition occurs before the deferral bound, there is no interruption cost; the only loss is the cost of the delayed review of the message. If the user remains busy up until the deferral bound, the cost receives contributions from the interruption and the maximal delay in seeing the message.

We shall first review efforts to build decision-theoretic notification systems. Then we will explore experimental studies of transitions between states associated with different costs of interruption. We formalize bounded deferral and seek to identify ideal deferral times. We conclude by discussing directions and prototypes.

2 Studies of Costs of Interruption and Delay

Decision-theoretic approaches and prototypes have been developed that continue to balance the cost of disruption, associated with alerting users with informational updates or incoming messages in different settings, with the cost coming with the delayed review of incoming messages [4]. Methods have used decision-analytic assessment techniques, such as assessing for the cost of disruption the dollars someone would be willing to pay to avoid different kinds of alerts in different settings [6].

Machine learning has been used to learn relatively accurate models for predicting the interruptability of users. Models have been developed to predict the cost of interruption of computer users in different settings as a function of multiple observations, including the sensing of ambient context (e.g., proximal conversation detected), desktop activity, and information about meetings drawn from users' calendars [4, 6, 8]. In some of this work, expected costs of disruption have been automatically assigned to different contexts.

Methods have also been developed for representing or inferring the loss of value of information in messages over time. Many messages, including urgent email, traffic updates, and financial alerts have time-dependent value. Taking a decision-analytic perspective, time-dependent losses in information value with delayed review of information can be assessed in terms of a willingness to pay to avoid successively greater delays in reviewing different classes of messages. That is, we can assess from users utility functions that capture, as a function of message properties, the initial value of information in dollars, as well as loss function that captures the loss of value with delayed review.

Prior research has also pursued the automation of the assignment of measures of urgency to messages. Under uncertainty in the value of a message or loss function, we consider the expected cost of delayed review (ECDA) [4]. Beyond the direct assessment of time-dependent utility, as a function of message properties, machine learning has shown promise in the construction of models that map to dollar values or intermediary measures of importance and/or urgency to messages.

To date, decision-theoretic notification systems have continued to do cost-benefit analysis, weighing the cost of delay with the costs of interruption at the current time, or over varying horizons for tractability. We shall introduce a family of policies that takes into consideration the transitions among states of interruptability. promising systems that employ a more global analysis of ideal alerting. We shall start by considering efforts to understand transitions among states of interruptability.

3 User Studies of Transitions in Cost of Interruption

We have performed user studies to build insights about the potential value of bounded deferral. The studies are aimed at probing the probability distributions which describe computer users' transitions among different states of cost of interruption.

3.1 Interruption Workbench Study

In an initial study, we re-examined data that had been collected as part of prior research on learning probabilistic models that can infer a cost of interruption [6]. In the study, five hours of video, shot in one hour segments, was taken of each of two participants in their offices. The video captured the details of work on their computer screens and surrounding office environments. The two subjects participating in the study had cost of interruption.

After the capture of videos, the users used an assessment tool called Interruption Workbench, which allowed the subjects to watch the video that had been taken and to label periods of time on the tape as low, medium, and high cost of interruption. The users had previously assessed a dollar value for each cost of interruption, representing their willingness to pay to avoid interruptions associated with the receipt of alerts when in each state. In the initial work on Interruption Workbench, the tagged states were used to build a case library for learning predictive models of the expected cost of interruption based on information sensed from the computer's applications and operating system, online calendar, and video and acoustical analyses of goings on in the office.

We re-analyzed data from the earlier studies to get a sense for the transitions among states of busyness, as defined by different assessed costs of interruption. As marginals, one participant (a program manager at our organization), spent 0.20, 0.61, and 0.18 of the total time in high, medium, and low cost states respectively. This subject remained in a busy state for a mean time of 21 seconds before transitioning into a lower cost state. The other participant (a software developer), spent 0.29, 0.48, and 0.23 of the total time in high, medium, and low cost states, and remained in a high cost state for a mean time of 202 seconds, before transitioning into a lower cost state.

We can get a sense for the opportunity for employing bounded deferral policy with the Interruption Workbench data. Given a message coming in at a random time while the user is in a busy state, one subject of the study will transition into a lower cost state with a mean time of 11 seconds after the arrival of an alert. The other subject would transition into a lower cost state at a mean time of 101 seconds after the arrival of an alert. Thus, we found that for these two subjects, allowing a relatively small, bounded deferral on the delivery of messages could significantly minimize costly interruptions in return for relatively small delays in message receipt.

3.2 Busy-Context Tool study

The intuitions about the value of deferral policies were supported by another study we carried out. In the study, anonymized data was collected from the logs of users of a prototype context-sensitive telephony system at our organization. The system, named Enhanced Telephone (ET), allows users to specify states of low and high cost of interruption with a *Busy Context* assessment tool. The tool enables the users to build boolean functions that predict being in a high-

or in a low-cost state, based on sensed active computer application and whether conversation is detected in the office. In use, the deterministic policies are used to guide telephone call routing, for example routing telephone calls that come in during busy states to voicemail. A client-side event-sensing system monitors computer activity and compiles a time-stamped event stream in a computer log that is uploaded intermittently to a server.

We investigated the busy versus free situations for 113 users for several weeks. The users included 42 program managers, 25 software developers, 10 administrators, 7 midlevel managers, 2 senior managers, 4 people in sales and marketing, 19 software testers, and 4 research scientists. The participants granted us access to their busy/free definition settings and to their free and busy states. Both the settings and the states were monitored via a server. We analyzed data collected over three sequential business days between 10am and 4pm when users were active at their desktops. We collected 4,803 busy situations.

The graph of Fig. 1 shows the distribution over durations of the monitored busy sessions for the participants. The mean duration of the busy sessions was found to be 43.12 seconds with a standard deviation of 51.79 seconds. The data shows that a great majority of busy situations transition to free situations within 1 to 2 minutes. The graph of Fig. 2 shows the transitions from busy to free for two users, where the total number of busy sessions for each is normalized to 1.0, thus providing probability distributions.

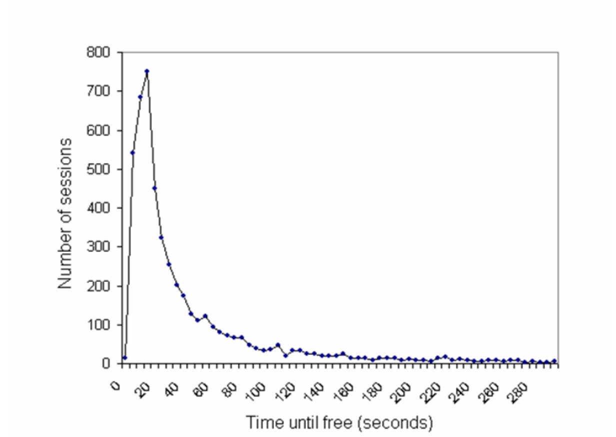

Figure 1: Distribution of the durations of busy situations for 113 users over three business days.

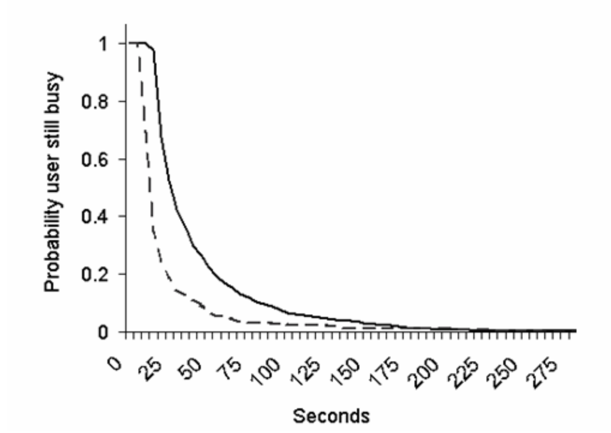

Figure 2: Probability distribution of the time of transition of participants from busy to free situations.

4 General Functional Analysis of Ideal Deferral Polices

Given motivating data from studies about users, we have pursued a formal analysis of bounded deferral. Our goal is to develop policies that employ ideal maximal deferral times, t^* for incoming notifications of different types in varying contexts. We wish to identify the ideal deferral policy for incoming messages, given the context-sensitive probability distribution over a recipient's transition from busy to free states, and the time-dependent loss of value with delay of the incoming message at hand.

For simplicity, we shall assume that users are either busy or free; this assumption can be relaxed into multiple states of cost of interruption. We further assume that if a user is busy, interruptions with an incoming message or informational update, are associated with an assessed cost of interruption, else interruptions are cost free. Again, this can be easily generalized. Finally, we assume a one shot analysis at the time a notification arrives. This assumption is violated by the opportunity to continue to optimize given changing probability distributions over transitions.

Let us start with a general analysis. Let the loss of value of reviewing an incoming message or alert at increasing times t after the arrival of that message be $f(t)$ and the cost of interruption of alerting the user when the user is in a busy state to be c. We take the probability that the user is free after time t as

$$p(t) = 1 - g(t)$$

where $g : [0, \infty) \to [0, 1]$ is an *arbitrary* non-increasing function.

Fig. 3 displays key relationships under consideration

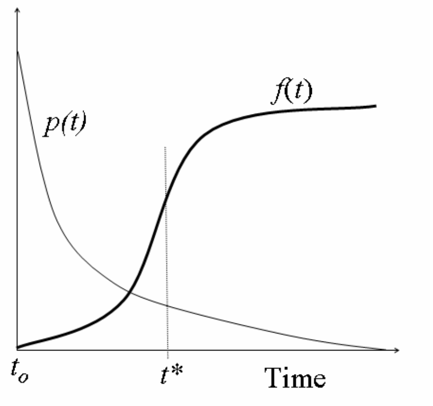

Figure 3: Key influences. Functions considered in computing ideal bounded-deferral policies are the probability distribution over time until free and the cost of delayed review.

in computing ideal bounded-deferral policies. We consider the time-dependent loss of the value of information in an incoming message, $f(t)$, and the probability distribution, $p(t)$, describing the likelihood that a busy user will transition to being free with increasing deferral times.

The best bound on deferral t^*, is the time that minimizes the overall expected cost to the user, including the expected cost the cost of disruption and the loss with delayed review of message information under uncertainty. The total cost takes into consideration two cases: (1) the case where the user remains busy up to the deferral bound, and (2) the case where the user becomes free at some time t before the maximal deferral time. For the case where the user remains busy, we need to consider the contributions to total expected cost of the losses from the disruption of being alerted when busy in addition to delay. In the case where the user becomes free before the bound on deferral time, only the cost of delay until the transition to the free state contributes to the total expected costs.

To compute a total expected cost associated with a selected deferral time, we combine the influence of the case where the deferral bound is reached, and the case where the user becomes free at some time before the bound. Thus, the total cost, $W(t)$, is

$$W(t) \quad = \quad (c + f(t))g(t) + \int_0^t -g'(s)f(s)ds$$

We seek to probe the existence of a global maxima by examining the derivative, $W'(t)$,

$$W'(t) \quad = \quad f'(t)g(t) + cg'(t) \qquad (1)$$

Given an instantiation of (1) with concrete functions f, g, let $T = \{t_0, t_1, \ldots\}$ be the set of all t such that $W'(t) = 0$ and $W''(t) < 0$. Then the expected total cost W is minimized for some $t \in \{0, T, \infty\}$, i.e., by one of the following policies: i) not waiting at all (immediate notification), ii) waiting until the time $t \in T$ that minimizes W, or iii) simply waiting until the user is free, no matter how long it takes.

For general f, g we need to solve,

$$W'(t) = 0 \Longleftrightarrow f'(t) = -c\frac{g'(t)}{g(t)} \ . \qquad (2)$$

Writing

$$g(t) = \exp(-h(t))$$

for some arbitrary increasing function t, we can rewrite (2) as

$$W'(t) = 0 \Longleftrightarrow f'(t) = ch'(t) \ .$$

Thus, the analysis of deferral times associated with minimal cost boils down to a consideration of how many times f' and ch' cross. Without further assumptions on the relationship between f, g, though, not much more can be said. One exception is that if for all t either

$$f'(t) < ch'(t) \qquad \text{or} \qquad f'(t) > ch'(t)$$

then the optimal policy is either "wait forever" or "act now" depending on whether $W(0) < W_\infty$ or not. Intuitively, if one rate dominates the other at all times, there is never a "critical" time such that it is reasonable to wait until then, but not more.

Let us now turn to examine a family of bounded deferral policies.

5 Analysis of an Expressive Family

We shall now introduce an expressive family of bounded-deferral problems that is amenable to optimization via a set of simple tests. The family, which we refer to as being in the SIGEX class of bounded deferral problems, is an instance captured by Equation 2. With this family, the probability that the user is free after time t is

$$p(t) = 1 - e^{-\lambda t}$$

where, in Equation 2, $h(t) = \lambda t$.

We also assume that the cost of interruption is some constant c.

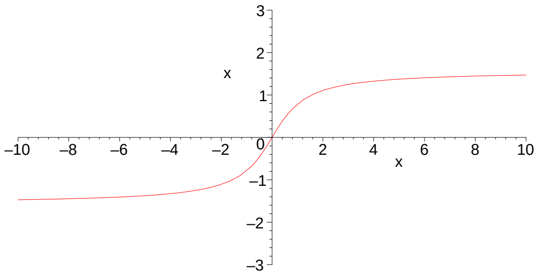

Figure 4: $f(t) = \arctan(\mathrm{t})$

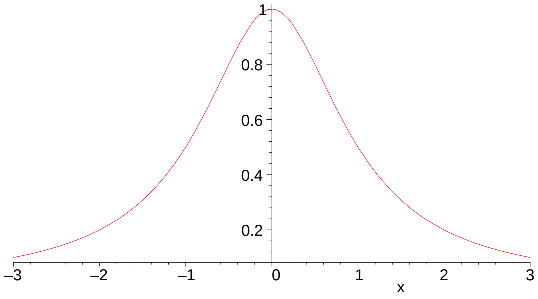

Figure 5: $f'(t) = \dfrac{1}{1+t^2}$

The total expected cost of waiting until time t before interrupting is

$$W(t) = (c + f(t))e^{-\lambda t} + \int_0^t \lambda e^{-\lambda s} f(s)ds$$

5.1 Analysis

Since

$$W'(t) = e^{-\lambda t}\left(f'(t) - \lambda c\right)$$

we see that if $t_0 > 0$ is to be an "optimal weighting time" we must have

$$\begin{aligned} f'(t_0) &= \lambda c \\ f''(t_0) &> 0 \ , \end{aligned}$$

where the second condition ensures that $f(t_0)$ is a minimum.

For the SIGEX family, we assume that the cost associated with the delayed receipt of information, $f(t)$ behaves like a sigmoid. We model this by requiring that

f' is a unimodal function having a unique maximum.

For example, the derivative of the arctan sigmoid function is the pulse:

Overlaying the horizontal line $g(t) = \lambda c$ with the plot of $f'(t)$ we see that the equation

$$W'(t) = 0 \Leftrightarrow f'(t) = \lambda c$$

can have at most two solutions, since f' can cut the line at no more than two points.

5.2 Case Analysis

Armed with the above, we can now distinguish cases.

- If $f'(t) < \lambda c$ for all t, i.e., the horizontal line is above the plot of f', then $W'(t) < 0$ for all t and therefore one should "Wait forever".
- If $f'(t) = \lambda c$ at a unique point t_0, i.e., the line "kisses" the maximum of f', then t_0 is *not* a local minimum of W since $W'(t) < 0$ for all $t \neq t_0$. It is just a point where W "bends".
- If $t_0 < t_1$ are the two points where $f'(t) = \lambda c$ then t_0 is a local minimum and t_1 is a local maximum (to see this consider the sign change of W').

Now, to determine whether t_0 is a *global* minimum we first observe that $W(t_0) < W(0)$ as $W'(0) < 0$. Therefore, the only thing we need to check is whether

$$\begin{aligned} W(t_0) &< \lim_{t\to\infty} W(t) && (3) \\ &= \int_0^\infty \lambda e^{-\lambda s} f(s)ds \ , && (4) \end{aligned}$$

i.e., it must be that the expected cost of waiting up to t_0 is smaller than the expected cost of waiting forever.

At the intuitive level, in the first case (since $h'(t)$ is constant), waiting forever is not the optimal policy only if initially f increases slowly, but there exists some time t_0 around which f increases dramatically. In such a case, we want to not wait until that point.

5.3 Algorithm for Determining Strategy

Given a bounded-deferral problem in the SIGEX family, we can identify the optimal deferral policy at each point in time with the following algorithm:

1. Solve the equation

   $$f'(t) = \lambda c \ .$$

   If the equation has fewer than two roots, then "Wait forever".

2. Otherwise let $t_0 < t_1$ be the two roots. Determine the cost of "unbounded delay", i.e.,

   $$W_\infty = \int_0^\infty \lambda e^{-\lambda s} f(s)ds \ .$$

3. If $W_\infty < W(t_0)$ then "Wait forever", else "Wait until time t_0".

Assuming a cost of delay that has a unimodal f' requires us to only check one place. We note that that we can generalize beyond SIGEX problems by considering multimodal functions for representing the cost of delayed review. In such cases, we need to check all the places where

$$f'(t) = c\lambda \ .$$

6 Conclusions and Directions

We pursued an analysis of bounded deferral policies. We first reviewed two user studies that demonstrate the potential value of bounded deferral. The studies identified examples of real-world probability distributions that describe how users may transition from busy to available states. Then, we formalized bounded deferral and performed a functional analysis of the task of identifying the optimal deferral policy at the time a notification arrives. After presenting a general analysis, we focused on a specific promising family of bounded deferral problems. We introduced an algorithm for testing whether notifications should be passed immediately to users, should wait until the user is free, or should wait for a specific amount of time for delivery. We are currently pursuing bounded deferral strategies that take into consideration the potential changes in the probability distribution over transitions from busy to free, as the time a user has remained busy grows. We are also pursuing the integration and evaluation of bounded deferral polices in several real-world applications.

References

[1] E. Cutrell, M. Czerwinski, and E. Horvitz. Notification, disruption and memory: Effects of messaging interruptions on memory and performance. In *Human-Computer Interaction–Interact '01, 2001.*, pages 263–269, 2001.

[2] M. Czerwinski, E. Cutrell, and E. Horvitz. Instant messaging and interruption: Influence of task type on performance. In *Proceedings of OZCHI 2000 of OZCHI 2000*, pages 99–100, 2000.

[3] T. Gillie and D. Broadbent. What makes interruptions disruptive? A study of length, similarity and complexity. *Psychological Research*, 50:243–250, 1989.

[4] E. Horvitz, A. Jacobs, and D. Hovel. Attention-sensitive alerting. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 305–313. AUAI, Morgan Kaufmann, August 1999.

[5] E. Horvitz, C. Kadie, T. Paek, and D. Hovel. Models of attention in computing and communication: From principles to applications. *Communications of the ACM*, 46(3), 2003.

[6] Eric Horvitz and Johnson Apacible. Learning and reasoning about interruption. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 20–27, 2003.

[7] Eric Horvitz, Paul Koch, and Johnson Apacible. Busybody: Creating and fielding personalized models of the cost of interruption. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW'04)*, pages 20–27, 2004.

[8] S. Hudson, J. Fogarty, C. Atkeson, J. Forlizzi, S. Kiesler, J. Lee, and J. Yang. Predicting human interruptibility with sensors: A wizard of oz feasibility study. In *Proceedings of CHI 2003*, pages 207–214, 2003.

[9] D.C. McFarlane. Coordinating the interruption of people in human-computer interaction. In *Human Computer Interaction - INTERACT'99*, pages 295–303. IOS Press, Inc., 1999.

[10] D.C. Mcfarlane. Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human-Computer Interaction*, 17(1):63–139, 2002.