Learning with Abandonment

Ramesh Johari
Stanford University
rjohari@stanford.edu

Sven Schmit
Stanford University
schmit@stanford.edu

Abstract

Consider a platform that wants to learn a personalized policy for each user, but the platform faces the risk of a user abandoning the platform if she is dissatisfied with the actions of the platform. For example, a platform is interested in personalizing the number of newsletters it sends, but faces the risk that the user unsubscribes forever. We propose a general thresholded learning problem that models scenarios like this, and discuss the structure of optimal policies. We discuss salient features of optimal policies under various assumptions on the distribution of the thresholds and feedback the platform receives.

1 Introduction

We study the problem of a platform that wants to learn a personalized policy for each user. The distinctive feature in this work is that the platform faces the risk of a user abandoning the platform if she is dissatisfied with the actions of the platform.

There are many examples of such settings. In the near future, smart energy meters are able to throttle consumer's energy consumption to increase efficiency of the power grid during peak demand, e.g. by turning the A/C up or down. This can lead to cost savings for both utility companies and consumers. However, if the utility company is too aggressive in its throttling of energy, a user might abandon the program. Newsletter creators face a similar problem. There is value in sending more emails, but each email also risks the recipient unsubscribing, taking away any opportunity of the creator to interact with the user again. Yet another example is that of smartphone app notifications. These can be used to improve user engagement and experience. However if the platform sends too many notifications, a user can turn off the notifications.

In all of the above scenarios, we face a decision problem where 'more is better', however there is a threshold beyond which the user abandons and no further rewards are gained.

Results We propose a general model for learning problems with the risk of abandonment. We first focus on the setting without any additional feedback and consider two special cases: independent thresholds and a single threshold across time. In both cases, we show the optimal policy is a constant policy, that is, the platform does not learn about the user. We then give approximation bounds of constant policies for threshold models with small deviations from the two special cases. Finally, we discuss (implicit) user feedback, allowing the platform to better personalize its policy. We show that this leads to partial learning, and that depending on the level of feedback, the optimal policy can be more aggressive or more conservative compared to the model without feedback.

Related work The work by [Lu et al., 2017] considers the same setting, but their model considers only a risky and a safe action. Furthermore, our work has similarities with the mechanism design literature [Myerson, 1981] and in particular there is an interesting connection to the dynamic pricing problem with strategic agents [Rothschild, 1974, Farias and Van Roy, 2010, Pavan et al., 2014, Lobel and Paes Leme, 2017]. In revenue management the focus is less on strategic agents, but the main focus is on selling a finite inventory [Gallego and Van Ryzin, 1994]. Finally, there is work on safe

reinforcement learning [García and Fernández, 2015, Berkenkamp et al., 2017] where a learner needs to avoid catastrophes. In that setting, the learner usually has access to additional feedback, for example it is given a safety region. We also do not insist on avoiding such catastrophic events.

2 Model

We consider a setting where a single user interacts with a platform at discrete time steps indexed by t. The user is characterized by sequence of hidden thresholds $\{\theta_t\}_{t=0}^{\infty}$ drawn from some known distribution. At every time t, the platform selects an action $x_t \in \mathbf{X} \subset \mathbb{R}_+$ from a given closed set $\mathbf{X}$. Based on the chosen action x_t, the platform obtains the random reward $R_t(x_t) \geq 0$. The expected reward of action x is given by $r(x) = \mathbb{E}(R_t(x)) < \infty$. However, when the action exceeds the threshold at time t, the process stops. More formally, let T be the stopping time that denotes the first time the x_t exceeds the threshold c:

$$T = \min\{t : x_t > \theta_t\}, \tag{1}$$

and let $\gamma \in (0, 1)$ denote the discount factor. The goal is to find a sequence of actions $\{x_t\}_{t=0}^{\infty}$ that maximizes

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t R_t(x_t)\right]. \tag{2}$$

We note that this expectation is well defined even if $T = \infty$.

3 Analysis

In this section we characterize the optimal policy under specific assumptions on the distribution of thresholds. We start by the simple case of independent thresholds, and then consider the other extreme; a single threshold. Finally, we discuss an additive noise model that interpolates between these two.

Independent thresholds First, consider the case where the thresholds θ_t are drawn independently from the same distribution F. Since there is no correlation between time steps, the optimal policy must be a constant policy that maximizes the instantaneous reward.

Proposition 1 *Suppose the function*

$$\frac{r(x)(1 - F(x))}{1 - \gamma(1 - F(x))} \tag{3}$$

has a unique optimum $x^* \in \mathbf{X}$. Then the optimal policy under the independent threshold assumption is $x_t = x^*$ for all t.

Single threshold Now we consider the other extreme, where the threshold is drawn once and used for all time steps, that is $\theta_t = \theta \sim F$. Surprisingly, in this setting the optimal policy is also a constant policy which corresponds to a complete lack of learning.

Proposition 2 *Suppose the function and the function $x \to r(x)(1 - F(x))$ has a unique optimum $x^* \in \mathbf{X}$. Then, the optimal policy is $x_t = x^*$ for all t.*

Note also that the optimal policy is independent of the discount factor γ. One way to prove this is by casting the problem as a dynamic program then using induction on the value iteration. A different way to explain the result is as follows. First, it is clear that the optimal policy is not decreasing. Now suppose the optimal policy is increasing, then there exists some t such that $x_t = y$ and $x_{t+1} = z > y$. Now we can show that in fact it must be better to play z at time t. First, suppose $\theta < y$, then playing y or z at time t both lead to abandonment. Now suppose $\theta > y$, by the optimality of the increasing policy, it is optimal to play z. Hence, in this case it is better to play z at time t as well. This contradicts the optimality of the original policy.

Additive noise model We can interpolate between the two threshold models by considering an additive noise model. In this case, the threshold at time t is given by$\theta_t = \theta + \varepsilon_t$, where $\theta \sim F$ is drawn once, and the noise terms are drawn independently. In general, the optimal now is increasing and intractable because the posterior over θ now depends on all previous actions. However, there exists constant policies that are close to optimal in the case the noise terms are either small or large.

First consider the case where the noise terms are small. In particular, suppose the error distribution has an arbitrary distribution over a small interval $[-y, y]$.

Proposition 3 *Suppose $\varepsilon_t \in [-y, y]$ and the reward function r is L-Lipschitz. Then there exists a constant policy with value V_c such that*

$$V^* - V_c \leq \frac{2yL}{1-\gamma} \tag{4}$$

where V^* is the optimal policy for the noise model, and x^* is the optimal constant policy for the noiseless case.

Similarly, when the noise level is sufficiently large with respect to the threshold distribution F there also exists a constant policy that is close to optimal. The formal statement requires us to develop more notation than space permits, so we state the following result informally.

Proposition 4 *(Informal) Under appropriate conditions, there exists a constant policy with (expected) value V_c such that*

$$V^* - V_c \leq \frac{2\eta B}{1-\gamma} + (1-\eta)\frac{L_v w}{2} \tag{5}$$

where η, w, and L_v depend on F, r, the noise distribution and the discount factor γ.

4 Feedback

As mentioned in the introduction, often the platform is able to capture feedback before abandonment. As example, consider optimizing the number of push notifications. When a user receives a notification, she may decide to open the app, or decide to turn off notifications. However, often her most likely action is to ignore the notification. The platform can interpret this as a signal to improve the policy.

To incorporate such feedback, we expand the model such that when the current action x_t level exceeds the threshold $x_t > \theta_t$, with probability p we receive no reward but the process continues, and with probability $1-p$ the user abandons. Furthermore, we restrict our attention to the single threshold model, where θ is drawn once and then fixed for all time periods. The optimal policy cannot be found in closed form, but two interesting phenomena occur which we discuss now.

Partial learning First, we show that the optimal policy exhibits *partial learning*. That is, the optimal policy in consists of two phases. Initially, the policy learns about the threshold using a bisection-type search. However, at some point (dependent on the user's threshold), further learning is too risky and the optimal policy switches to a constant policy. We note that this happens even when there is no risk of abandonment at all ($p = 1$).

Aggressive and conservative policies Another salient feature of the structure of optimal policies in the feedback model is the aggressiveness of the policy. In particular, we say a policy is *aggressive* if the first action x_0 is larger than the optimal constant policy x^* in the absence of feedback (corresponding to $p = 0$), and *conservative* if it is smaller.

When there is low risk of abandonment, i.e. $p \approx 1$, then the optimal policy is aggressive, as the optimal policy immediately targets high-value users, without risking abandonment of users with lower thresholds.

However, when the risk of abandonment is large ($p \approx 0$) and the discount factor is sufficiently close to one, the optimal policy is more conservative than the optimal constant policy when $p = 0$.

References

Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. *CoRR*, abs/1705.08551, 2017.

Vivek F Farias and Benjamin Van Roy. Dynamic pricing with a prior on market response. *Operations Research*, 58(1):16–29, 2010.

Guillermo Gallego and Garrett Van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management science*, 40(8):999–1020, 1994.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16:1437–1480, 2015.

Ilan Lobel and Renato Paes Leme. Dynamic mechanism design under positive commitment. 2017.

Jiaqi Lu, Yash Kanoria, and Ilan Lobel. Dynamic decision making under customer abandonment risk. *MSOM*, 2017.

Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

Alessandro Pavan, Ilya Segal, and Juuso Toikka. Dynamic mechanism design: A myersonian approach. *Econometrica*, 82(2):601–653, 2014.

Michael Rothschild. A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9 (2):185–202, 1974.