Sending Timely Status Updates through Channel with Random Delay via Online Learning

Haoyue Tang, Yuchao Chen, Jingzhou Sun, Jintao Wang, Jian Song

Department of Electronic Engineering, Tsinghua University
Beijing National Research Center for Information Science and Engineering, Beijing, China
Research Institute of Tsinghua University in Shenzhen, Shenzhen, China
{thy17@mails, cyc20@mails, sunjz18@mails, wangjintao@mails, jsong@mails}@tsinghua.edu.cn

Abstract—In this work, we study a status update system with a source node sending timely information to the destination through a channel with random delay. We measure the timeliness of the information stored at the receiver via the Age of Information (AoI), the time elapsed since the freshest sample stored at the receiver is generated. The goal is to design a sampling strategy that minimizes the total cost of the expected time average AoI and sampling cost in the absence of transmission delay statistics. We reformulate the total cost minimization problem as the optimization of a renewal-reward process, and propose an online sampling strategy based on the Robbins-Monro algorithm. Denote \( K \) to be the number of samples we have taken. We show that, when the transmission delay is bounded, the expected time average total cost obtained by the proposed online algorithm converges to the minimum cost when \( K \) goes to infinity, and the optimality gap decays with rate \( O(\ln K/K) \). Simulation results validate the performance of our proposed algorithm.

I. INTRODUCTION

Timely status information is crucial for many real-time control systems, e.g., the autonomous vehicular networks and the tactile internet. To measure the timeliness of status update, the metric Age of Information (AoI) [1] has been proposed. By definition, AoI measures the time elapsed since the freshest information stored at the receiver is generated, and a small AoI performance requires the system to possess both high throughput and low transmission delay [1]. However, this is often limited by the transmission resources of the sensor. Moreover, the lack of precise transmission statistics (e.g., channel condition and delay distribution) makes the design of status update system more challenging.

Designing energy efficient sampling and transmission strategies to optimize the timeliness of status update systems have been studied in [2]–[13]. In discrete time scenarios, minimizing the time average AoI performance can be formulated into a Markov decision process. When update packets are generated at will, low complexity algorithms to obtain the AoI minimum are proposed in [12], [13]. Notice that the aforementioned studies require the transmission statistics (e.g., transmission delay or packet-loss probabilities) to be known in advance. When such information is unavailable to the transmitter, reinforcement learning is an efficient tool to learn the optimum policy adaptively based on historical transmissions. Reinforcement learning algorithms such as Q-learning and SARSA have been employed to obtain the AoI minimum sampling and transmission strategies [14], [15]. To reduce the storage and computational complexity, SARSA with tile coding [16], deep Q-Learning [7] and actor-critic algorithm [17] have been used to tackle with the huge state space of AoI minimization problems. The aforementioned RL based algorithms require the transmitter either to store a large table of value functions, or to train a neural network for function approximation. Either exerts high storage and computational burden to the sensor. Moreover, analyzing the convergence rate of the proposed algorithms remains challenging.

Online learning and bandit algorithms provide efficient solutions with a low computational cost for sequential decision making in an unknown environment. When the generation of update packets are controlled by external environment and arrives randomly, AoI minimum adaptive channel selection and link scheduling algorithms based on bandit algorithms have been proposed [18]–[20]. Vishrant et al. [21] model the timeliness at the receiver as a time-varying function of the AoI, and propose online learning algorithms that can adapt to adversarial cost function variations. In multi-user networks, [22] proposes scheduling algorithms that can satisfy the timeliness
constraint of each user with sublinear cumulative utility regret. However, the aforementioned studies deal with discrete time scenarios and assume the transmission of each packet is instantaneous, i.e., the transmission delay is one slot or can be ignored. Although an online sampling algorithm for AoI minimization over a two-way random delay has been proposed in [23], the regret performance is not well understood.

To solve this problem, we consider a point-to-point status update system with a sensor sampling and transmitting information updates to the receiver through a lossless network with a random transmission delay, similar to [13], [24]. We assume each transmission attempt incurs an extra cost, and aim at minimizing the total cost that consists of both the average AoI and transmission cost. The main contributions of the paper are as follows:

- By reformulating the average cost minimization problem as a renewal-reward process optimization, we derive the optimal off-line sampling strategy when the transmission delay distribution is known.
- We propose an online adaptive sampling strategy when the transmission delay distribution is unknown using the Robbins-Monro algorithm [25], [26]. Denote $K$ to be the number of samples we have taken, we show that the gap between the cumulative expected cost by using the proposed online algorithm and the minimum cost decays like $O\left( \ln K / K \right)$, i.e., the proposed online algorithm adaptively learns the optimal policy.

II. PROBLEM FORMULATION

A. System Model

As is depicted in Fig. 1, we consider a sensor observes a time-sensitive process, samples and transmits update information to the receiver through a network interface queue similar to [13], [24]. The network interface serves the update packets on the First-Come-First-Serve (FCFS) basis. An ACK will be sent back to the sensor once an update packet is cleared at the interface, and we assume the transmission duration after passing the network interface is negligible.

Fig. 1. System model.

Suppose the sensor can generate and submit update packets at any time $t \in \mathbb{R}^+$ at will. Notice that update packets become stale while waiting in the queue, and the busy/idle information of the queue is available to the sensor through the ACK. Therefore, it is better to sample a new update packet after the ACK of the previous submitted packet is received. Denote $S_k$ to be the generation time-stamp of the $k$-th update packet and the time duration spent in the network interface is $D_k$. We assume each $D_k$ is independent and identically distributed (i.i.d.) following probability measure $P_D$.

Assumption 1: The probability measure $P_D$ is absolutely continuous. Moreover, its expectation and second order moment is bounded, i.e.,

$$0 < \bar{D}_{lb} \leq \bar{D} \equiv \mathbb{E}_{P_D}[D] \leq \bar{D}_{ub} < \infty, \quad (1a)$$

$$0 < M_{lb} \leq \mathbb{E}_{P_D}[D^2] \leq M_{ub} < \infty, \quad (1b)$$

where $\bar{D}_{lb}, \bar{D}_{ub}$ are the lower and upper bound of the average transmission delay. The lower and upper bound of the second order moment of transmission delay are denoted by $M_{lb}, M_{ub}$, respectively.

Since the queue is empty when the $k$-th update packet is submitted, packet $k$ will be received at time $S_k + D_k$. Since each update packet $k$ is generated after the reception of packet $k-1$, we denote $W_{k+1} := S_{k+1} - (S_k + D_k)$ as the “waiting” time to take the $k+1$-th sample after receiving the ACK of the $k$-th sample at time $S_k + D_k$.

B. Age of Information

We measure the information freshness of the receiver at time $t$ via the Age of Information [11], namely the time elapsed since the latest information stored at the receiver is generated. Let $i(t) := \arg \max_{k \in \mathbb{N}} \{ S_k + D_k \leq t \}$ be the index of the latest sample received by the destination before time $t$. The AoI at time $t$, denoted by $A(t)$ is:

$$A(t) := t - S_{i(t)}. \quad (2)$$

A sample path of AoI evolution is depicted in Fig. 2.

Fig. 2. Illustration of AoI evolution.

C. Optimization Problem Formulation

Denote $\overline{A}_{\pi,T}$ and $\overline{F}_{\pi,T}$ to be the expected average AoI and sampling frequency by using policy $\pi$ over an interval $(0, T)$, i.e.,

$$\overline{A}_{\pi,T} := \mathbb{E} \left[ \frac{1}{T} \int_{t=1}^{T} A(t) dt \right], \quad \overline{F}_{\pi,T} := \mathbb{E} \left[ \frac{i(T)}{T} \right].$$

We assume an extra cost of $C \geq 0$ is incurred each time when the sensor samples and submits an update packet. Assume that $P_D$ is unknown to the transmitter,
the goal is to design a sampling strategy \( \pi \) represented by a set of waiting times \( \{W_k\}_{k=1}^{\infty} \) to minimize the total sum of expected average AoI and sampling cost based on \( D_b, D_a, M_b, M_a \). Specifically, we focus on the class of causal policies denoted by \( \Pi_{\text{causal}} \), where each policy \( \pi \in \Pi_{\text{causal}} \) selects the waiting time \( W_k \) of the \((k+1)\)-th update packet based on the transmission delay of the \(k\)-th packet \( D_k \) and the historical transmissions denoted by filtration \( \mathcal{F}_{k-1} := \sigma((D_k, W_k))_{k=1}^{\infty} \), where \( \sigma(\cdot) \) represents the \( \sigma \)-field generated by random variables. The overall optimization problem is as follows:

**Problem 1:**

\[
\pi^* = \arg\min_{\pi \in \Pi_{\text{causal}}} h_\pi := (\overline{A}_\pi + CF_\pi),
\]

where \( \overline{A}_\pi := \limsup_{T \to \infty} \overline{A}_{\pi,T} = \limsup_{T \to \infty} \mathbb{E} \left[ \frac{1}{T} \int_{t=1}^{T} A(t) dt \right] \),

\[
F_\pi := \limsup_{T \to \infty} F_{\pi,T} = \limsup_{T \to \infty} \mathbb{E} \left[ \frac{i(T)}{T} \right].
\]

### III. Problem Resolution

In this section, we first reformulate Problem 1 as the optimization of a renewal-reward process. In Subsection-B, we derive the optimum offline policy when the distribution \( P_D \) is known, which provides design insight and is used to measure the convergence rate of the proposed online algorithm. And then provide an adaptive online sampling strategy in Subsection-C when \( P_D \) is unknown. Finally, we analyze the average AoI performance of the proposed online policy in Subsection-D.

#### A. A Renewal-Reward Process Reformulation

To formulate Problem 1 as a renewal-reward process optimization problem, we define "frame" \( k \) to be the interval of time during the generation of the \( k \)-th and the \((k+1)\)-th update packet. The length of the \( k \)-th frame is:

\[
L_k := D_k + W_k.
\]

According to Fig. 2, the cumulative AoI in frame \( k \), denoted by \( X_k := \int_{t=S_{k-1}}^{S_k} A(t) dt \), is the sum of the area of a parallelogram and a triangle, i.e.,

\[
X_k = (D_{k-1} + W_{k-1})D_k + \frac{1}{2}L_k^2.
\]

For ease of exposition, we denote \( D_0 = 0 \) and \( W_0 = 0 \). We consider the time interval \((0, T)\) with \( T = S_{K+1} \), i.e., the time when the \((K+1)\)-th update packet is sampled. Let \( Y_k = X_k + C \) be the sum of cumulative AoI and sampling cost in frame \( k \), then with probability 1, the total cost in (3a) can be rewritten as follows:

\[
h_\pi = \lim_{K \to \infty} \frac{\mathbb{E}[\sum_{k=1}^{K} X_k]}{\mathbb{E}[\sum_{k=1}^{K} L_k]} + \lim_{K \to \infty} \frac{C \cdot K}{\mathbb{E}[\sum_{k=1}^{K} L_k]}
\]

\[
= \lim_{K \to \infty} \frac{\mathbb{E}[\sum_{k=1}^{K} Y_k]}{\mathbb{E}[\sum_{k=1}^{K} L_k]}
\]

**Definition 1:** Let \( \Pi_{\text{SD}} \subset \Pi_{\text{causal}} \) be the set of stationary deterministic policies. A stationary deterministic policy \( \pi \in \Pi_{\text{SD}} \) chooses the waiting time \( W_k \) in frame \( k \) based on a deterministic function parameterized by policy \( \pi \) and the current delay, i.e., \( W_k = f_\pi(D_k) \), \( \forall k \).

**Theorem 1:** If the distribution \( P_D \) is known, there exists a stationary policy that is optimal to Problem [1]

**Proof:** The proof is similar to [13] Theorem 2 and Theorem 3 and is omitted due to space limitations. \( \blacksquare \)

We then focus on searching for the optimum stationary policy \( \pi \in \Pi_{\text{SD}} \). With slight abuse of notations, we denote \( \pi(D) \) to be the waiting time selected by a policy \( \pi \) upon observing transmission delay \( D \). To facilitate further analysis, denote

\[
R_k := \frac{1}{2}L_k^2 + C.
\]

Therefore the total cost in frame \( k \) can be written as

\[
Y_k = R_k + L_{k-1}D_k
\]

Then, the optimization objective (6) can be rewritten as follows:

\[
h_\pi = \lim_{K \to \infty} \frac{\mathbb{E}[\sum_{k=1}^{K} \left( \frac{1}{2}L_k^2 + L_{k-1}D_k + C \right) \bigg]}{\mathbb{E}[\sum_{k=1}^{K} L_k]}
\]

\[
= \lim_{K \to \infty} \frac{\mathbb{E}[\sum_{k=1}^{K} R_k] + \mathbb{E}[\sum_{k=1}^{K} L_{k-1}D_k]}{\mathbb{E}[\sum_{k=1}^{K} L_k]}
\]

\[
\approx \lim_{K \to \infty} \frac{\mathbb{E}[\sum_{k=1}^{K} R_k] + \mathbb{E}[\sum_{k=1}^{K} L_{k-1}D_k]}{\mathbb{E}[\sum_{k=1}^{K} L_k]}
\]

\[
(9)
\]

where \((a)\) holds because \( D_k \) is i.i.d.

We treat \( R_k \) as the "reward" in frame \( k \). For stationary deterministic policy \( \pi \in \Pi_{\text{SD}} \), the reward \( R_k \) and length \( L_k \) of frame \( k \) only depend on the current delay \( D_k \). Since the delay \( D_k \) of each packet is i.i.d, the reward and frame length \( (R_k, L_k) \) is independent of \((R_{k'}, L_{k'})\) in other frames and can be modeled as a renewal-reward process. Moreover, the expectation \( \mathbb{E}[L_k] \) is bounded. Therefore, according to the renewal theory, denote \( \mathbb{E}[R] = \mathbb{E}[\frac{1}{2}(D + \pi(D))^2 + C] \) and \( \mathbb{E}[L] = \mathbb{E}[D + \pi(D)] \), with probability 1. Problem (1) can be reformulated as follows:

**Problem 2 (Renewal-Reward process reformulation of Problem 1):**

\[
\pi^* = \arg\min_{\pi \in \Pi_{\text{SD}}} \left( \frac{\mathbb{E}[\frac{1}{2}(D + \pi(D))^2 + C]}{\mathbb{E}[D + \pi(D)]} + \overline{T} \right).
\]

**B. Optimal Offline Algorithm based on \( P_D \)**

We first analyze the optimum stationary policy when the delay distribution \( P_D \) is known. The optimum offline policy will provide important insight to the design of online algorithm.

Recall that \( h_{\pi^*} \) is the minimum cost any policy \( \pi \in \Pi_{\text{SD}} \) can achieve, i.e.,

\[
\mathbb{E}[\frac{1}{2}(D + \pi(D))^2 + C] \geq h_{\pi^*}, \forall \pi \in \Pi_{\text{SD}}.
\]
Denote $\gamma^* = h_{\gamma^*} - \mathcal{D}$. Multiplying $E[\pi(D + \pi(D))]$ on both sides of inequality \([11]\), we have:

$$
\frac{1}{2}[E[(D + \pi(D))^2 + C] - \gamma^*E[D + \pi(D)] \geq 0, \forall \pi \in \Pi_{\text{SD}}.
$$

(12)

For simplicity, denote function $g(\pi, \gamma) := \frac{1}{2}E[(D + \pi(D))^2 + C] - \gamma^*E[D + \pi(D)]$. When $\gamma = \gamma^*$, inequality (12) implies $g(\pi, \gamma^*) = 0$ if and only if $\pi = \pi^*$. Therefore, if the $\gamma^*$ is known, the optimum sampling policy that achieves $h_{\gamma^*}$ can be obtained by solving the following constrained optimization problem:

**Problem 3 (Functional Optimization Problem):**

$$
\min_{\pi \in \Pi_{\text{SD}}} g(\pi, \gamma^*) := E\left[\frac{1}{2}(D + \pi(D))^2 + C - \gamma^*(D + \pi(D))\right].
$$

(13)

To search for the optimum policy $\pi^*$, we present the following lemmas and corollaries:

**Theorem 2:** Denote $\tilde{\pi}_\gamma^* := \arg\min_{\pi \in \Pi_{\text{SD}}} g(\pi, \gamma)$ as the optimum stationary deterministic policy that minimizes the function $g(\pi, \gamma)$. Policy $\tilde{\pi}_\gamma^*$ specifies the waiting time upon observing transmission delay $D$ as follows:

$$
\tilde{\pi}_\gamma^*(D) = (\gamma - D)^+.
$$

Proof: The strict mathematic proof is similar to and is omitted due to space limitations \([13]\).

**Corollary 1:** The optimum ratio $\gamma^* = \frac{1}{2}E[(D + \pi^*(D))^2 + C - \gamma^*(D + \pi^*(D))]$ and can be lower and upper bounded by:

$$
\gamma_{\text{lb}} := \frac{1}{2}D_{\text{lb}} \leq \gamma \leq \frac{1}{2}M_{\text{ub}} + C =: \gamma_{\text{ub}}.
$$

(14)

**Corollary 2:** Define mapping $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ to be:

$$
\mathcal{T}(\gamma) := \frac{E[\frac{1}{2}(D + \tilde{\pi}_\gamma^*(D))^2 + C]}{E[D + \tilde{\pi}_\gamma^*(D)]}.
$$

(15)

Let $\mathcal{T}(\gamma) \triangleq \mathcal{T} \circ \cdots \circ \mathcal{T}(\gamma)$. Then $\lim_{\gamma \to \infty} \mathcal{T}(\gamma) = \gamma^*$, $\forall \gamma \in [\gamma_{\text{lb}}, \gamma_{\text{ub}}]$.

Proofs for Corollary 1 and 2 are provided in Appendix A and B respectively.

We then present our offline algorithm that find $\gamma^*$ iteratively using Corollary 2 if the delay distribution $P_{\mathcal{D}}$ is known:

- We initialize $\gamma_0$ by uniformly choosing from interval $[\gamma_{\text{lb}}, \gamma_{\text{ub}}]$.
- In each iteration $j \geq 1$, we compute the corresponding mapping $\gamma_j = \mathcal{T}(\gamma_{j-1})$. The iteration stops in iteration $J$ when the absolute value is below a certain threshold $|\gamma_{j} - \gamma_{j+1}| \leq \delta$.
- We use the policy $\tilde{\pi}_{\gamma_j}$ in the last epoch, i.e., when observing transmission delay $D$, we wait for $W = (\gamma_j - D)^+$ before taking the next sample.

**C. Proposed Online Algorithm**

We then provide an online AoI minimization algorithm in the absence of distribution $P_{\mathcal{D}}$. The key is to maintain a sequence $\{\gamma_k\}$ that reflects our guess of the optimum $\gamma^*$ using the Robbins-Monro algorithm \([25]\). We initialize $\gamma_0 \in [\gamma_{\text{lb}}, \gamma_{\text{ub}}]$ randomly in the first frame $k = 1$. For frame $k \geq 2$, the algorithm operates as follows:

- We observe the transmission delay $D_k$ and choose waiting time:

$$
W_k = (\gamma_k - D_k)^+.
$$

(16a)

We then compute the frame length $L_k = D_k + W_k$ and $R_k = \frac{1}{2}L_k^2 + C$.

- We then update $\gamma_k$ via the Robbins-Monro algorithm \([25]\) as follows:

$$
\gamma_{k+1} = \gamma_k + \eta_k (R_k - \gamma_k L_k)^{\gamma_{\text{ub}}},
$$

(16b)

where $[\gamma_k^n] = \min\{b, \max\{\gamma, a\}\}$ and $\{\eta_k\}$ is a set of diminishing step sizes that is selected to be:

$$
\eta_k = \begin{cases} \frac{1}{2\eta_{\text{ub}}}, & k = 1; \\ \frac{1}{(k+2)^2\eta_{\text{ub}}}, & k \geq 2. \end{cases}
$$

(16c)

**D. Theoretical Analysis**

The average AoI regret performance of the proposed policy is hard to analyze in general. As an alternative, recall that the total AoI and the sampling cost over the first $K$ frames can be computed by $\sum_{k=1}^{K} Y_k$, where $Y_k = R_k + L_k - D_k$ is the total cost in frame $k$ as pointed out by \([8]\). The total length of the first $K$ frames is $\sum_{k=1}^{K} L_k$. Therefore the ratio $\frac{\gamma_{k}}{\gamma_{k}} \triangleq \frac{E[\Sigma_{k=1}^{K} Y_k]}{E[\Sigma_{k=1}^{K} L_k]}$ reflects the average AoI obtained by the online policy. Denote $\pi_K$ as the waiting strategy used in the $K$-th frame, i.e., $\pi_K(D) = (\gamma_K - D)^+$. We measure the performance of the proposed algorithm via the convergence rate of gap $h_{\gamma^*} - h_{\gamma}$ and $h_{\pi_K} - h_{\gamma^*}$, and the main results are as follow:

**Theorem 3:** If the transmission delay $D$ is upper bounded, i.e., $D < B < \infty$, we have:

$$
\pi_K - h_{\gamma^*} \leq \frac{(L_{\text{ub}} + C)^2}{D_{\text{lb}}^2} \times \frac{1 + \ln K}{K}. 
$$

(17a)

where $L_{\text{ub}} = B + \gamma_{\text{ub}}$. The expected difference between the expected average cost by using policy $\pi_K$ in frame $K$ and $h_{\gamma^*}$ can be upper bounded by:

$$
E[h_{\pi_K} - h_{\gamma^*}] \leq \frac{(L_{\text{ub}} + C)^2}{D_{\text{lb}}^2} \times \frac{1}{K}. 
$$

(17b)

Proof of Theorem 3 can be found in Appendix C.

**IV. SIMULATIONS**

In this section we simulate three sampling policies:

1. Zero-wait policy that takes a new sample immediately when the ACK of the last sample is received, i.e., $\pi_{\text{ZW}}(d) = 0, \forall d$;
2. The optimum off-line policy


The proposed algorithm for $C > 0$ of transmission delay is small. The online algorithm converges faster when the variance $\sigma^2$ is small. The expectation $\mu^*$ of the average AoI obtained by our online algorithm converges to the expected average AoI obtained by the optimal offline policy. As is illustrated in Fig. 3, the time average AoI obtained by our algorithm converges faster than the second sub-plots in Fig. 3 our online algorithm converges faster when the variance of transmission delay is small.

Next we study time average cost obtained by the proposed algorithm for $C > 0$. Recall that $A_{x,t}$ is the cumulative AoI in frame $K$ and $C$ is the cumulative cost. We plot the average cost $\pi_K = \mathbb{E}\sum_{k=1}^{K}(L_k + C)$ up to frame $K$ by using different algorithms in Fig. 4. The transmission delay $P_D$ follows the log-normal distribution parameterized by $(\mu, \sigma) = (1, 1.5)$.

V. CONCLUSIONS

We investigated sampling strategies to minimize the sum of average AoI and sampling cost for status update system with random transmission delay. We reformulate the problem as the optimization of a renewal-reward process and find the optimal policy when the delay distribution is known. We then propose an online sampling strategy that adaptively learns the optimum offline policy using the Robbins-Monro algorithm. It has been demonstrated that the optimality gap between the average cost of the online algorithm and the minimum cost obtained by the optimal offline policy diminishes when the number of samples goes to infinity.

APPENDIX A

PROOF OF COROLLARY

Proof: First we derive the lower bound of $\gamma^*$. Consider any stationary deterministic policy $\pi \in \Pi_{SD}$, we have:

$$\mathbb{E}\left[\frac{1}{2}(D + \pi(D))^2 + C\right] = \mathbb{E}[\pi^2 + D^2 + \pi(D)] = \mathbb{E}[D + \pi(D)] + \frac{1}{2} E[D^2] + C \geq \frac{1}{2} M_{ub} + C.$$

where inequality (a) is because $C \geq 0$ and the Cauchy-Schwarz inequality implies $\mathbb{E}[(D + \pi(D))^2] \geq \mathbb{E}[D + \pi(D)]^2$.

To find the upper bound of $\gamma^*$, we consider zero wait policy $\pi_{zw}$ that selects $W = 0, \forall D$. Since policy $\pi_{zw}$ may not be the optimum policy, we have:

$$\gamma^* \leq \mathbb{E}\left[\frac{1}{2}(D + \pi_{zw}(D))^2 + C\right] \mathbb{E}[D + \pi_{zw}(D)]$$

$$= \frac{1}{2} \mathbb{E}[D^2] + C \leq \frac{1}{2} M_{ub} + C : = \gamma_{ub}. \quad (19)$$

APPENDIX B

PROOF OF COROLLARY

Proof: Recall that the mapping $\mathcal{T}(\cdot): \mathbb{R} \to \mathbb{R}$ is:

$$\mathcal{T}(\gamma) := \frac{\mathbb{E}\left[\frac{1}{2}(D + \pi_{zw}^*(D))^2 + C\right]}{\mathbb{E}[D + \pi_{zw}^*(D)]}.$$
and according to Theorem 2, the waiting time obtained by policy \( \tilde{\pi} \) is:

\[
\tilde{\pi}^*_t(d) = (\gamma - d)^+. 
\]

Therefore, \( d + \tilde{\pi}^*_t(d) = \max\{d, \gamma\} \), we can rewrite mapping \( T(\gamma) \) as follows:

\[
T(\gamma) = \frac{E\left[\frac{1}{2}(\max\{D, \gamma\})^2 + C\right]}{E[\max\{D, \gamma\}]}.
\] (20)

Since distribution \( P_D \) is absolutely continuous, \( T \) is continuous. Next, we will show that for any \( \gamma \in [\gamma_b, \gamma_u] \), \( T(\gamma) \) is bounded via the following inequality:

\[
T(\gamma) \leq \frac{1}{D_b} \left( \frac{1}{2}M_{\gamma_b} + \mathcal{D}_{\gamma_b} + \gamma^2_{\gamma_b} + C \right) =: T_{\gamma_b}. 
\] (21)

Apparently, \( T_{\gamma_b} \geq \gamma_{\gamma_b} \) and the stationary point \( \gamma \in [\gamma_b, T_{\gamma_b}] \).

We will then show mapping \( T(\gamma) \) has a unique stationary point \( \gamma^* \). Suppose \( \gamma^* \) is a stationary point, we have:

\[
E\left[\frac{1}{2}(\max\{D, \gamma^*\})^2 + C\right] = \gamma^*E[\max\{D, \gamma^*\}]. 
\] (22)

The proof is divided into two cases:

- If \( \gamma > \gamma^* \), we will show \( T(\gamma) \) is a sub-contraction mapping:

  \[
  T(\gamma) - \gamma^* = E\left[\frac{1}{2}(\max\{D, \gamma^*\})^2 + C\right] - \gamma^* 
  \leq \frac{1}{2}(\max\{D, \gamma^*\})^2 - \gamma^* \max\{D, \gamma\} + C 
  \leq (\gamma - \gamma^*) \frac{1}{2}(\max\{D, \gamma^*\})^2 - \gamma^* \max\{D, \gamma\} + C 
  \leq (\gamma - \gamma^*) (1 - \frac{\max\{D, \gamma^*\}}{\max\{D, \gamma\}}) \leq (\gamma - \gamma^*),
  \]

  where inequality (a) is because policy \( \pi(D) = (\gamma - D)^+ \) is the optimum policy to minimize function \( g(\pi, \gamma) \); equality (b) is obtained because of (22).

  We will then show \( T(\gamma) - \gamma^* \) is positive:

  \[
  T(\gamma) - \gamma^* = E\left[\frac{1}{2}(\max\{D, \gamma\})^2 + C\right] - \gamma^* 
  \leq \frac{1}{2}(\max\{D, \gamma\})^2 - \gamma^* \max\{D, \gamma\} + C 
  \]

  where inequality (c) is because policy \( \pi(\gamma, D) = (\gamma - D)^+ \) is the optimum policy to minimize \( g(\gamma, \pi) \) and equality (d) is obtained because of (22).

  Since the stationary point \( \gamma^* \) exists, combining (22) and (24) leads to the conclusion that \( T(\gamma) \) is a sub-contraction mapping:

  \[
  |T(\gamma) - \gamma^*| < (\gamma - \gamma^*), \forall \gamma > \gamma^*.
  \]

  Finally, consider that the stationary point exists and satisfies \( \gamma^* < \gamma_u < \infty \), the image \( 0 < T(\gamma) < \gamma_u < \infty \) belongs to a compact set. According to [27], we have:

\[
\lim_{\tau \to \infty} T_T(\gamma) = \gamma^*.
\] (25)

- If \( \gamma < \gamma^* \), we will first show \( T(\gamma) - \gamma^* > (\gamma - \gamma^*) \):

  \[
  T(\gamma) - \gamma^* = E\left[\frac{1}{2}(\max\{D, \gamma\})^2 + C\right] - \gamma^* 
  \geq \frac{1}{2}(\max\{D, \gamma\})^2 - \gamma^* \max\{D, \gamma\} + C 
  \geq (\gamma - \gamma^*) (1 - \frac{\max\{D, \gamma\}}{\max\{D, \gamma^*\}}) \geq (\gamma - \gamma^*),
  \]

  where inequality (e) is because policy \( \pi(D) = (\gamma - D)^+ \) is the optimum policy to minimize \( g(\pi, \gamma) \); inequality (f) is because policy \( \pi(D) = (\gamma^* - D)^+ \) is the optimum policy to minimize function \( g(\gamma, \pi^*) \); equality (g) is obtained because of (22); inequality (h) is obtained because \( \min\{\max\{D, \gamma\} \} < 1 \) and \( \gamma < \gamma^* \) implies

\[
E[\max\{D, \gamma\}] < E[\max\{D, \gamma^*\}] 
\]

For the case that \( \gamma < \gamma^* \), denote \( \tau \) be the stopping time that:

\[
\tau(\gamma) = \arg \min_{\tau \in \mathbb{N}^+} \{T_T(\gamma) \geq \gamma^*\}.
\]

If \( \tau(\gamma) < \infty \), then \( T_T(\tau(\gamma)) > \gamma^* \), and according to (25), we have

\[
\lim_{\tau \to \infty} T_T(\tau(\gamma)) = \gamma^*.
\]
Otherwise, if \( \epsilon(\gamma) = \infty \), due to (25), the mapping is subgradient. Since the stationary point exists, \( \lim_{\tau \to \infty} T^{(\gamma)}(\gamma) = \gamma^* \).

**APPENDIX C**

**PROOF OF THEOREM 3**

**Proof:** First, recall that the ratio \( \gamma_k \) used in any frame \( k \) is upper bounded by \( \gamma^* \), since the transmission delay is bounded by \( D_k \), the frame length \( L_k \) and the reward \( R_k \) can be upper bounded by:

\[
L_k = \frac{1}{2} L_k^2 + C \leq L_k^* + C.
\]

Denote \( T^k := \mathbb{E}[D + \pi^*(D)] \) and \( \bar{T} := \mathbb{E}[\frac{1}{L}(D + \pi^*(D))^2 + C] \) to be the expected frame length and the reward in each frame using the optimum policy \( \pi^* \). To proceed, we provide Lemma 1 and Lemma 2, the proofs are provided in Appendix D and E respectively.

**Lemma 1:** The expected frame length \( E[L_k|F_{k-1}] \) and the expected reward \( \mathbb{E}[R_k|F_{k-1}] \) of frame \( k \) satisfies:

\[
\mathbb{E}[R_k - \gamma_k L_k|F_{k-1}] \leq (\gamma^* - \gamma_k) \bar{T}^k,
\]

\[
\mathbb{E}[R_k - \gamma^* L_k|F_{k-1}] \leq -(\gamma^* - \gamma_k) \left( \mathbb{E}[L_k|F_{k-1}] - \bar{T}^k \right).
\]

**Lemma 2:** Recall that \( Y_k = R_k + L_k - D_k \) is the sum of cumulative AoI and extra cost within frame \( k \) according to (3). Denote

\[
\Delta_K := \mathbb{E} \left[ \sum_{k=1}^{K} (Y_k - \gamma^* L_k) \right].
\]

Then, \( \Delta_K \) can be upper bounded by:

\[
\Delta_K \leq \mathbb{E} \left[ \sum_{k=1}^{K} (\gamma^* - \gamma_k)^2 \right].
\]

Notice that the average cost deviation \( \bar{T}_k - h_{\pi^*} = \mathbb{E}\left[ \sum_{k=1}^{K} \Delta_K \right] \) is sufficent to study the convergence behavior of \( \mathbb{E}\left[ \sum_{k=1}^{K} (\gamma^* - \gamma_k)^2 \right] \). The next lemma bound the expected difference between \( \gamma_k \) and \( \gamma^* \) in frame \( k \), whose proof is provided in Appendix F.

**Lemma 3:** The expected difference between \( \gamma_k \) and \( \gamma^* \) can be upper bounded by:

\[
\mathbb{E}[\gamma_k - \gamma^*] \leq \frac{1}{K} \left( \frac{L_{ab}^2 + C}{D_k^2} \right).
\]

With the above lemmas, we can verify the two conclusions in Theorem 3 respectively:

**Proof of (17a):** Plugging (30) into inequality (29) from Lemma 2 we can first upper bound \( \Delta_K \) by:

\[
\Delta_K \leq \left( \frac{L_{ab}^2 + C}{D_k^2} \right) \left( \sum_{k=1}^{K} \frac{1}{E_k} \right).
\]

\[
= \left( \frac{L_{ab}^2 + C}{D_k^2} \right) \left( 1 + \int_{k=1}^{K} \frac{1}{k} \right)
\]

\[
= \left( \frac{L_{ab}^2 + C}{D_k^2} \right) \left( 1 + \ln K \right),
\]

where inequality (a) is because \( \frac{1}{k} \leq \int_{k-1}^{k} \frac{1}{k} \).

Notice that:

\[
\bar{T}_K - h_{\pi^*} = \frac{1}{K} \mathbb{E} \left[ \sum_{k=1}^{K} L_k \Delta_K \right]
\]

\[
\leq \frac{\left( \frac{L_{ab}^2 + C}{D_k^2} \right)^2 (1 + \ln K)}{K},
\]

where inequality (b) is because

\[
\mathbb{E} \left[ \sum_{k=1}^{K} L_k \right] \geq \mathbb{E} \left[ \sum_{k=1}^{K} D_k \right] \geq K \bar{T}^k.
\]

**Proof of (17b):** Recall that the average cost using policy \( \pi_K \) can be computed by

\[
h_{\pi_K} = \frac{1}{K} \left[ \frac{1}{2} (\gamma_k - D)^2 + C \right] + \bar{T}^k,
\]

which is smaller or equal to the minimum cost, i.e., \( h\pi_K \geq h_{\pi^*} \). Therefore for each frame \( K \), we can upper bound the gap between policy \( \pi_K \) and \( \pi^* \) by:

\[
h_{\pi_k} - h_{\pi^*} = \mathbb{E}\left[ \frac{1}{2} ((\gamma_k - D)^2 + C) \right] + \bar{T}^k,
\]

\[
\mathbb{E}\left[ \frac{1}{2} ((\gamma_k - D)^2 + C) \right] - \gamma^*
\]

\[
= \mathbb{E}\left[ \frac{1}{2} ((\gamma_k - D)^2 + C - \gamma_k ((\gamma_k - D)^2 + D) \right]
\]

\[
+ (\gamma_k - \gamma^*)
\]

\[
\leq \mathbb{E}\left[ \frac{1}{2} ((\gamma_k - D)^2 + C - \gamma_k ((\gamma_k - D) + D) \right]
\]

\[
+ (\gamma_k - \gamma^*)
\]

\[
\leq \mathbb{E}[\gamma_k \gamma_k (\gamma_k - D)^2 + D] \mathbb{E}[\gamma_k (\gamma_k - D)^2 + D] + (\gamma_k - \gamma^*)
\]

\[
= \frac{1}{D} (\gamma_k - \gamma^*)^2.
\]

Plugging (30) from Lemma 3 into the above equation, we have

\[
\mathbb{E}[h_{\pi_k} - h_{\pi^*}] \leq \left( \frac{L_{ab}^2 + C}{D_k^2} \right) \left( \frac{1}{K} \right).
\]

And this completes the proof of Theorem 3.

**APPENDIX D**

**PROOF OF LEMMA 1**

**Proof:** Notice that in each frame \( k \), the waiting time \( W_k \) is chosen to minimize the objective function (13), therefore we have:

\[
\mathbb{E}[R_k - \gamma_k L_k|F_{k-1}] \leq (\bar{T}^k - \gamma_k \bar{T}^k).
\]
where inequality (a) is because policy $\pi_k$ is a stationary policy. Equality (b) is obtained because for the stationary point $\gamma^*$ we have $R = \gamma^*$.

This verifies the first inequality in Lemma 1. Then, adding $(\gamma_k - \gamma^*)E[L_k | F_{k-1}]$ to both sides of (34), we have:

$$
E[R_k - \gamma^* L_k | F_{k-1}] \leq (\gamma_k - \gamma^*)E[L_k - \gamma^* | F_{k-1}].
$$

(35)

which verifies the second inequality in Lemma 1.

APPENDIX E
PROOF OF LEMMA 2

Proof: To find the upper bound of $\Delta_k$, first we add $E[L_{k-1} D_k | F_{k-1}]$ on both sides of (38b). By replacing $R_k + L_{k-1} D_k$ with $Y_k$, we then have the following inequality:

$$
E[Y_k - \gamma^* L_k | F_{k-1}] \leq (\gamma_k - \gamma^*)E[L_k + D_k | F_{k-1}]
$$

(d) \leq (\gamma_k - \gamma^*)^2 + E[L_{k-1} | F_{k-1}] D_k,

(36)

where inequality (d) is because

$$
E[L_k - \gamma^* | F_{k-1}] = E[(\gamma_k - D)^+ - (\gamma^* - D)^+] \leq |\gamma_k - \gamma^*|.
$$

(37)

Summing up inequality (36) from frame $k = 1$ to $K$ and taking the expectation with respect to $F_{K-1}$, we have:

$$
\frac{1}{2} E \left[ \sum_{k=1}^{K} (Y_k - (\gamma^* + D_k) L_k) \right] \leq \frac{1}{2} E \left[ \sum_{k=1}^{K} (\gamma_k - \gamma^*)^2 \right] - E[L_K] D_k.
$$

(38)

And this completes the proof of Lemma 2.

APPENDIX F
PROOF OF LEMMA 3

Proof: For simplicity, denote $z_{k+1} := \gamma_k + \eta_k (R_k - \gamma_k L_k)$.

(39)

Since $\gamma_{k+1} = [z_{k+1}]_+^{\eta_k}$ and the optimum ratio $\gamma^* \in [\gamma_h, \gamma_a]$, we can bound the derivation $(\gamma_{k+1} - \gamma^*)^2$ using $(z_{k+1} - \gamma^*)^2$ following the following inequality:

$$
(\gamma_{k+1} - \gamma^*)^2 = ([z_{k+1}]_+^{\eta_k} - [\gamma^*]_+^{\eta_k})^2 \leq (z_{k+1} - \gamma^*)^2.
$$

(40)

Next, recall the update rule given in (16b), we have:

$$
\frac{1}{2} (z_{k+1} - \gamma^*)^2 = \frac{1}{2} (\gamma - \gamma^* + \eta_k (R_k - \gamma_k L_k))^2 = \frac{1}{2} (\gamma - \gamma^*)^2 + \frac{1}{2} \eta_k^2 (R_k - \gamma_k L_k)^2
$$

+ $\eta_k (\gamma_k - \gamma^*) (R_k - \gamma_k L_k) 
\leq \frac{1}{2} (\gamma - \gamma^*)^2 + \frac{1}{2} \eta_k^2 (L_{ab}^2 + C)^2
$$

+ $\eta_k (\gamma_k - \gamma^*) (R_k - \gamma_k L_k),

(41)

where the last inequality is obtained because $R_k = \frac{1}{2} L_{ab}^2 + C$ and $\gamma_k L_k \leq L_{ab}^2$. Then, conditioned on filtration $F_{k-1}$ and take the expectation on both sides of (41), we have:

$$
\frac{1}{2} E[(z_{k+1} - \gamma^*)^2 | F_{k-1}] 
\leq \frac{1}{2} (\gamma - \gamma^*)^2 + \frac{1}{2} \eta_k^2 (L_{ab}^2 + C)^2
$$

+ $\eta_k (\gamma_k - \gamma^*) E[R_k - \gamma_k L_k | F_{k-1}].

(42)

We then proceed to bound the last term in inequality (42). The analysis is divided into two cases:

- If the current $\gamma_k - \gamma^* \geq 0$, by plugging (28a) into the above equation, we have:

$$
(\gamma_k - \gamma^*) E[R_k - \gamma_k L_k | F_{k-1}] \leq (\gamma_k - \gamma^*)^2 T^* \leq - (\gamma_k - \gamma^*)^2 D_k.
$$

(43)

where the last inequality is obtained because $T^* \geq D_k$.

- If the current $\gamma_k - \gamma^* \leq 0$, we can upper the last term in inequality (42) as follows:

$$
(\gamma_k - \gamma^*) E[R_k - \gamma_k L_k | F_{k-1}] = (\gamma_k - \gamma^*) E[R_k - \gamma_k L_k | F_{k-1}]
$$

- $- (\gamma_k - \gamma^*)^2 E[L_k | F_{k-1}]

(a) \leq (\gamma_k - \gamma^*) (R_k - \gamma^*) - (\gamma_k - \gamma^*)^2 E[L_k | F_{k-1}]

\leq - (\gamma_k - \gamma^*)^2 E[L_k | F_{k-1}]

(b) \leq - (\gamma_k - \gamma^*)^2 D_k,

(44)

where inequality (a) is because $E[R_k - \gamma_k L_k | F_{k-1}] \geq T^* - \gamma^* T^*$ and $E[L_k | F_{k-1}] \geq D_k$. Inequality (b) is because $E[L_k | F_{k-1}] \geq D_k$.

Plugging (43) and (44) into (42) yields:

$$
\frac{1}{2} E[(z_{k+1} - \gamma^*)^2 | F_{k-1}]
$$

= $\left(1 - \eta_k D_k\right) (\gamma - \gamma^*)^2 + \frac{1}{2} \eta_k^2 (L_{ab}^2 + C)^2
$$

\leq \left(1 - \eta_k D_k\right) (\gamma - \gamma^*)^2 + \frac{1}{2} \eta_k^2 (L_{ab}^2 + C)^2.

(45)

By taking the expectation with respect to filtration $F_{k-1}$ on both sides of inequality (45) and then plugging it into (40), we can upper bound $E[(z_{k+1} - \gamma^*)^2]$ through:

$$
\frac{1}{2} E[(z_{k+1} - \gamma^*)^2] \leq \frac{1}{2} E[(z_{k+1} - \gamma^*)^2]
$$

\leq \left(1 - 2 \eta_k D_k\right) E[(\gamma - \gamma^*)^2] + \frac{1}{2} \eta_k^2 (L_{ab}^2 + C)^2.

(46)
where inequality (c) is obtained because of inequality (40).

Recall that the stepsizes are selected through \( \eta_1 = \frac{1}{2\alpha_0} \)
and \( \eta_k = \frac{1}{(k+1)\alpha_0} \), \( \forall k > 1 \), we can then show by induction that

\[
\frac{1}{2} \mathbb{E}[(\gamma_k - \gamma^*)^2] \leq \frac{1}{2k} \left( \frac{L^2_{ah} + C)^2}{D_{lb}} \right).
\]

Detailed proofs for (47) are as follow:

- When \( k = 2 \), by choosing \( \eta_1 = \frac{1}{2\alpha_0} \) we have

\[
\frac{1}{2} \mathbb{E}[(\gamma_2 - \gamma^*)^2] \leq \frac{1}{8} \left( \frac{L^2_{ah} + C)^2}{D_{lb}} \right).
\]

- When \( k > 2 \), assuming that \( \frac{1}{2} \mathbb{E}[(\gamma_k - \gamma^*)^2] \leq \frac{1}{2k} \left( \frac{L^2_{ah} + C)^2}{D_{lb}} \right) \), recall that the stepsize \( \eta_k \) is chosen to be \( \eta_k = \frac{1}{(k+1)\alpha_0} \), we have:

\[
\frac{1}{2} \mathbb{E}[(\gamma_{k+1} - \gamma^*)^2] \leq \left( \frac{1}{2 - \eta_k D_{lb}} \right) \left( \gamma_k - \gamma^* \right)^2 + \frac{1}{2} \eta_k^2 \left( \frac{L^2_{ah} + C)^2}{D_{lb}} \right)
\]

\[
\leq \frac{1}{2k} \left( \frac{L^2_{ah} + C)^2}{D_{lb}} \right)
\]

\[
\leq \frac{1}{2} \left( \frac{(k+1) \gamma_k - \gamma^*}{k+2} \right)^2 + \frac{1}{2 (k+2) \eta_k^2} \left( \frac{L^2_{ah} + C)^2}{D_{lb}} \right)
\]

\[
\leq \frac{1}{2} \frac{1}{(k+1) \gamma_k - \gamma^*}
\]

where the final inequality is obtained because \( (k+1)(k+3) \leq (k+2)^2 \).