# 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

IEEE INFOCOM 2022 - IEEE Conference on Computer Communications | 978-1-6654-5822-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/INFOCOM48880.2022.9796970

***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 $\mathcal{O}\left(\ln K/K\right)$. Simulation results validate the performance of our proposed algorithm.**

## I. Introduction

Timely status information is crucial for many real-time control system, 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 randomly by external environment, AoI minimum transmission and preemption strategies are proposed in [2]–[5]. When the generation of update packets can be controlled at the transmitter, the joint design of sampling, power control and retransmission strategies are studied in [6]–[8]. In continuous time scenarios, when the status update generation is controlled by an external random process, the expected AoI performance under different service disciplines are analyzed in [9], [10]. When the update packets can be generated at will, low complexity algorithms to obtain the optimum sampling and transmission strategies 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

This work was supported by Tsinghua University-China Mobile Research Institute Joint Innovation Center. (*Corresponding author: Jintao Wang*)

The authors have provided public access to their code or data at https://github.com/loveisbasa/Infocom2022

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 optimum 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 $\mathcal{O}\left(\ln K/K\right)$, 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.

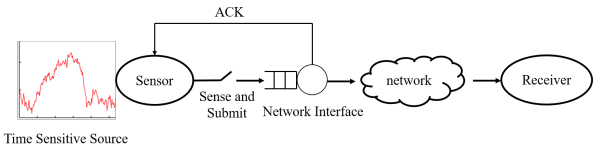


Fig. 1. System model.

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 $\mathsf{P}_D$.

*Assumption 1:* The probability measure $\mathsf{P}_D$ is absolutely continuous. Moreover, its expectation and second order moment is bounded, i.e.,

$$0 < \overline{D}_{\text{lb}} \leq \overline{D} \triangleq \mathbb{E}_{\mathsf{P}_D}[D] \leq \overline{D}_{\text{ub}} < \infty, \tag{1a}$$

$$0 < M_{\text{lb}} \leq \mathbb{E}_{\mathsf{P}_D}[D^2] \leq M_{\text{ub}} < \infty, \tag{1b}$$

where $\overline{D}_{\text{lb}}, \overline{D}_{\text{ub}}$ 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 $M_{\text{lb}}, M_{\text{ub}}$, 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 [1], namely the time elapsed since the latest information stored at the receiver is generated. Let $i(t) := \arg\max_{k\in\mathbb{N}}\{k | S_k + D_k \leq t\}$ be the index of the latest sample received by the destination before time $t$. The AoI at time $t$, denoted by $A(t)$ is:

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

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

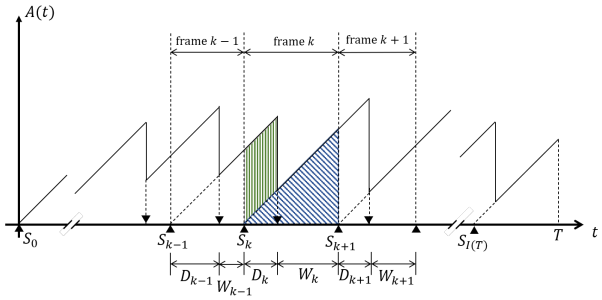


Fig. 2. Illustration of AoI evolution.

### 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.,

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

We assume an extra cost of $C \geq 0$ is incurred each time when the sensor samples and submits an update packet. Assume that $\mathsf{P}_D$ 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 $\overline{D}_{\text{lb}}, \overline{D}_{\text{ub}}, M_{\text{lb}}, M_{\text{ub}}$. 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 $\mathcal{F}_{k-1} := \sigma\left(\{(D_\kappa, W_\kappa)\}_{\kappa=1}^{k-1}\right)$, where $\sigma(\cdot)$ represents the $\sigma$-field generated by random variables. The overall optimization problem is as follows:

*Problem 1:*

$$\pi^* \triangleq \underset{\pi\in\Pi_{\text{Causal}}}{\arg\min}\, h_\pi := \left(\overline{A}_\pi + C\overline{F}_\pi\right), \tag{3a}$$

$$\text{where } \overline{A}_\pi := \limsup_{T\to\infty} \overline{A}_{\pi,T} = \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_{t=1}^{T} A(t)\mathrm{d}t\right], \tag{3b}$$

$$\overline{F}_\pi := \limsup_{T\to\infty} \overline{F}_{\pi,T} = \limsup_{T\to\infty} \mathbb{E}\left[\frac{i(T)}{T}\right]. \tag{3c}$$

## 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 $\mathsf{P}_D$ 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 $\mathsf{P}_D$ 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:

$$L_k := D_k + W_k. \tag{4}$$

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

$$X_k = (D_{k-1} + W_{k-1})D_k + \frac{1}{2}L_k^2. \tag{5}$$

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:

$$\begin{aligned} h_\pi &= \lim_{K\to\infty} \frac{\mathbb{E}_\pi[\sum_{k=1}^K X_k]}{\mathbb{E}_\pi[\sum_{k=1}^K L_k]} + \lim_{K\to\infty} \frac{C\cdot K}{\mathbb{E}_\pi[\sum_{k=1}^K L_k]} \\ &= \lim_{K\to\infty} \frac{\mathbb{E}_\pi[\sum_{k=1}^K Y_k]}{\mathbb{E}_\pi[\sum_{k=1}^K L_k]}. \end{aligned} \tag{6}$$

*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 $\mathsf{P}_D$ 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. ■

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

$$R_k := \frac{1}{2}L_k^2 + C. \tag{7}$$

Therefore the total cost in frame $k$ can be written as

$$Y_k = R_k + L_{k-1}D_k \tag{8}$$

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

$$\begin{aligned} h_\pi &= \lim_{K\to\infty} \frac{\mathbb{E}[\sum_{k=1}^K (\frac{1}{2}L_k^2 + L_{k-1}D_k + C)]}{\mathbb{E}[\sum_{k=1}^K L_k]} \\ &= \lim_{K\to\infty} \frac{\mathbb{E}[\sum_{k=1}^K R_k] + \mathbb{E}[\sum_{k=1}^K L_{k-1}D_k]}{\mathbb{E}[\sum_{k=1}^K L_k]} \\ &\overset{(a)}{=} \lim_{K\to\infty} \frac{\mathbb{E}[\sum_{k=1}^K R_k] + \mathbb{E}[\sum_{k=1}^K L_{k-1}]\overline{D}}{\mathbb{E}[\sum_{k=1}^K L_k]}, \end{aligned} \tag{9}$$

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):*

$$\pi^* = \underset{\pi\in\Pi_{\text{SD}}}{\arg\min} \left(\frac{\mathbb{E}[\frac{1}{2}(D+\pi(D))^2 + C]}{\mathbb{E}[D+\pi(D)]} + \overline{D}\right). \tag{10}$$

### B. Optimal Offline Algorithm based on $\mathsf{P}_D$

We first analyze the optimum stationary policy when the delay distribution $\mathsf{P}_D$ 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.,

$$\frac{\mathbb{E}_\pi[\frac{1}{2}(D+\pi(D))^2 + C]}{\mathbb{E}_\pi[D+\pi(D)]} + \overline{D} \geq h_{\pi^*}, \forall \pi \in \Pi_{\text{SD}}. \tag{11}$$

Denote $\gamma^* = h_{\pi^*} - \overline{D}$. Multiplying $\mathbb{E}_\pi[D+\pi(D)]$ on both sides of inequality (11), we have:

$$\frac{1}{2}\mathbb{E}[(D+\pi(D))^2+C] - \gamma^*\mathbb{E}[D+\pi(D)] \geq 0, \forall \pi \in \Pi_{\text{SD}}. \tag{12}$$

For simplicity, denote function

$$g(\pi,\gamma) := \frac{1}{2}\mathbb{E}[(D+\pi(D))^2+C] - \gamma^*\mathbb{E}[D+\pi(D)].$$

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 $h_{\pi^*}$ can be obtained by solving the following constrained optimization problem:

*Problem 3 (Functional Optimization Problem):*

$$\min_{\pi\in\Pi_{\text{SD}}} g(\pi,\gamma^*) := \mathbb{E}\left[\frac{1}{2}(D+\pi(D))^2 + C - \gamma^*(D+\pi(D))\right]. \tag{13}$$

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 $\gamma^* = \frac{1}{2}\frac{\mathbb{E}\left[(D+\pi^*(D))^2\right]}{\mathbb{E}[D+\pi^*(D)]}$ and can be lower and upper bounded by:

$$\gamma_{\text{lb}} := \frac{1}{2}\overline{D}_{\text{lb}} \leq \gamma \leq \frac{\frac{1}{2}M_{\text{ub}}+C}{\overline{D}_{\text{lb}}} =: \gamma_{\text{ub}}. \tag{14}$$

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

$$\mathcal{T}(\gamma) := \frac{\mathbb{E}[\frac{1}{2}(D+\tilde{\pi}^*_\gamma(D))^2+C]}{\mathbb{E}[D+\tilde{\pi}^*_\gamma(D)]}. \tag{15}$$

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

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 $\mathsf{P}_D$ 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 $\mathsf{P}_D$. 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:
$$W_k = (\gamma_k - D_k)^+. \tag{16a}$$
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:
$$\gamma_{k+1} = \left[\gamma_k + \eta_k\left(R_k - \gamma_k L_k\right)\right]_{\gamma_{\text{lb}}}^{\gamma_{\text{ub}}}, \tag{16b}$$
where $[\gamma]_a^b = \min\{b,\max\{\gamma,a\}\}$ and $\{\eta_k\}$ is a set of diminishing step sizes that is selected to be:
$$\eta_k = \begin{cases} \frac{1}{2\overline{D}_{\text{lb}}}, & k=1; \\ \frac{1}{(k+2)\overline{D}_{\text{lb}}}, & k\geq 2. \end{cases} \tag{16c}$$

### D. Theoretic 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 $Y_k = R_k + L_{k-1}D_k$ 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 $\overline{h}_K \triangleq \frac{\mathbb{E}\left[\sum_{k=1}^K Y_k\right]}{\mathbb{E}\left[\sum_{k=1}^K L_k\right]}$ 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 $\overline{h}_K - h_{\pi^*}$ and $h_{\pi_K} - h_{\pi^*}$, and the main results are as follow:

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

$$\overline{h}_K - h_{\pi^*} \leq \frac{(L_{\text{ub}}^2+C)^2}{\overline{DD}_{\text{lb}}^2} \times \frac{1+\ln K}{K}. \tag{17a}$$

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 $h_{\pi^*}$ can be upper bounded by:

$$\mathbb{E}\left[h_{\pi_K} - h_{\pi^*}\right] \leq \frac{(L_{\text{ub}}^2+C)^2}{\overline{DD}_{\text{lb}}^2} \times \frac{1}{K}. \tag{17b}$$

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

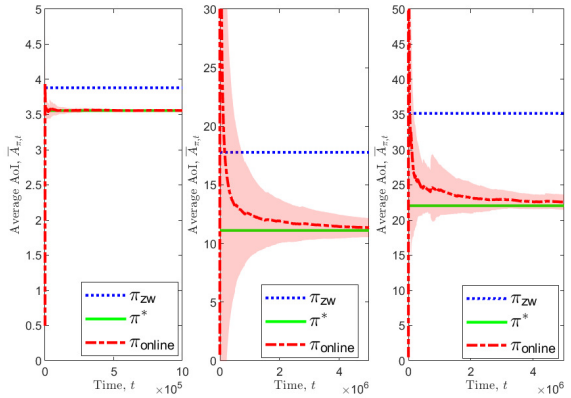


Fig. 3. The average AoI $\overline{A}_{\pi,t}$ evolution with time $t$ by using various algorithm. From left to right $(\mu, \sigma) = (1, 1), (1, 1.5), (2, 1.5)$.

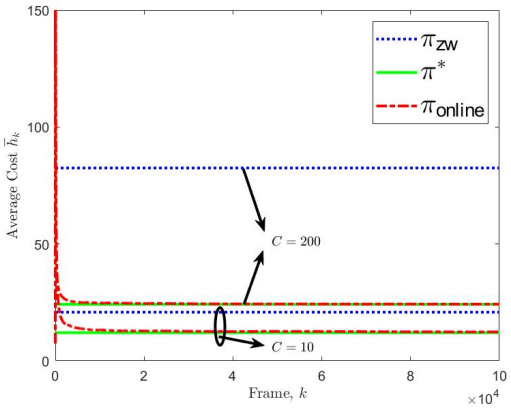


Fig. 4. The average cost evolution with the number of frames $k$. The transmission delay $\mathsf{P}_D$ follows the log-normal distribution parameterized by $(\mu, \sigma) = (1, 1.5)$.

$\pi^*$ with known $\mathsf{P}_D$ proposed in Section III-B; (3). The online algorithm in Section III-C. Simulations are carried out when the transmission delay $D$ follows a log-normal distribution parameterized by $\mu$ and $\sigma$, i.e., the density function

$$f_D(d) := \frac{\mathsf{P}_D(\mathsf{d}d)}{\mathsf{d}d} = \frac{1}{d\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln d - \mu)^2}{2\sigma^2}\right).$$

First we consider $C = 0$, i.e., sampling has no extra cost. Therefore, the total cost minimization problem degrades to the average AoI minimization problem considered in [13]. The expectation $\overline{A}_{\pi,t}$ is computed by taking the average of $\frac{1}{t}\int_0^t A(t)\mathsf{d}t$ over 50 simulations and the confidence region is marked in red in the figure. As is illustrated in Fig. 3, the time average AoI obtained by our algorithm converges to the expected average AoI obtained by the optimum algorithm [13] for all the parameters. Notice that a small $\sigma$ indicates the variance of the delay is small, comparing the first and 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 $X_k$ is the cumulative AoI in frame $K$ and $C$ is the sampling cost. We plot the average cost $\overline{h}_K = \frac{\mathbb{E}\left[\sum_{k=1}^K (X_k + C)\right]}{\mathbb{E}\left[\sum_{k=1}^K L_k\right]}$ up to frame $K$ by using different algorithms in Fig. 4. The transmission delay $D$ follows the log-normal distribution with parameter $\mu = 1$ and $\sigma = 1.5$. The proposed online learning algorithm adaptively learns the optimum policy that balances the average AoI performance and sampling cost. The zero-wait policy is far from optimum when the sampling cost is large.

## 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 1

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

$$\frac{\mathbb{E}[\frac{1}{2}(D + \pi(D))^2 + C]}{\mathbb{E}[D + \pi(D)]} \overset{(a)}{\geq} \frac{1}{2}\frac{\mathbb{E}[D + \pi(D)]^2}{\mathbb{E}[D + \pi(D)]}$$
$$= \frac{1}{2}\mathbb{E}[D + \pi(D)] \geq \frac{1}{2}D_{\text{lb}} =: \gamma_{\text{lb}}. \tag{18}$$

where inequality $(a)$ is because $C \geq 0$ and the Cauchy-Schwartz 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 $\pi_{\text{zw}}$ that selects $W \equiv 0, \forall D$. Since policy $\pi_{\text{zw}}$ may not be the optimum policy, we have:

$$\gamma^* \leq \frac{\mathbb{E}\left[\frac{1}{2}(D + \pi_{\text{zw}}(D))^2 + C\right]}{\mathbb{E}[D + \pi_{\text{zw}}(D)]}$$
$$= \frac{\frac{1}{2}\mathbb{E}[D^2] + C}{\overline{D}} \leq \frac{\frac{1}{2}M_{\text{ub}} + C}{\overline{D}_{\text{lb}}} =: \gamma_{\text{ub}}. \tag{19}$$

■

## Appendix B
## Proof of Corollary 2

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

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

and according to Theorem 2, the waiting time obtained by policy $\tilde{\pi}^*_\gamma$ is:

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

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

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

Since distribution $\mathsf{P}_D$ is absolutely continuous, $\mathcal{T}$ is continuous. Next, we will show that for any $\gamma \in [\gamma_{\text{lb}}, \gamma_{\text{ub}}]$, $\mathcal{T}(\gamma)$ is bounded via the following inequality:

$$\begin{aligned}\mathcal{T}(\gamma) &\leq \frac{\mathbb{E}[\frac{1}{2}(D+\gamma)^2 + C]}{\overline{D}} \\ &\leq \frac{1}{\overline{D}}\left(\frac{1}{2}M_{\text{ub}} + \overline{D}\gamma_{\text{ub}} + \gamma_{\text{ub}}^2 + C\right) =: \mathcal{T}_{\text{ub}}. \end{aligned}\tag{21}$$

Apparently, $\mathcal{T}_{\text{ub}} \geq \gamma_{\text{ub}}$ and the stationary point $\gamma \in [\gamma_{\text{lb}}, \mathcal{T}_{\text{ub}}]$.

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

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

The proof is divided into two cases:

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

$$\begin{aligned}&\mathcal{T}(\gamma) - \gamma^* = \frac{\mathbb{E}\left[\frac{1}{2}(\max\{D,\gamma\})^2 + C\right]}{\mathbb{E}\left[\max\{D,\gamma\}\right]} - \gamma^* \\ =&\frac{\mathbb{E}\left[\frac{1}{2}(\max\{D,\gamma\})^2 - \gamma\max\{D,\gamma\} + C\right]}{\mathbb{E}\left[\max\{D,\gamma\}\right]} \\ &+ (\gamma - \gamma^*) \\ \overset{(a)}{<}&\frac{\mathbb{E}\left[\frac{1}{2}(\max\{D,\gamma^*\})^2 - \gamma\max\{D,\gamma^*\} + C\right]}{\mathbb{E}\left[\max\{D,\gamma\}\right]} \\ &+ (\gamma - \gamma^*) \\ =&\frac{\mathbb{E}\left[\frac{1}{2}(\max\{D,\gamma^*\})^2 - \gamma^*\max\{D,\gamma^*\} + C\right]}{\mathbb{E}\left[\max\{D,\gamma\}\right]} \\ &+ (\gamma^* - \gamma)\frac{\mathbb{E}\left[\max\{D,\gamma^*\}\right]}{\mathbb{E}\left[\max\{D,\gamma\}\right]} + (\gamma - \gamma^*) \\ \overset{(b)}{=}&(\gamma - \gamma^*)\left(1 - \frac{\mathbb{E}\left[\max\{D,\gamma^*\}\right]}{\mathbb{E}\left[\max\{D,\gamma\}\right]}\right) \leq (\gamma - \gamma^*),\end{aligned}\tag{23}$$

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 $\mathcal{T}(\gamma) - \gamma^*$ is positive:

$$\begin{aligned}&\mathcal{T}(\gamma) - \gamma^* \\ =&\frac{\mathbb{E}\left[\frac{1}{2}(\max\{D,\gamma\})^2 + C\right]}{\mathbb{E}\left[\max\{D,\gamma\}\right]} - \gamma^* \\ =&\frac{\mathbb{E}\left[\frac{1}{2}(\max\{D,\gamma\})^2 - \gamma^*\cdot\max\{D,\gamma\} + C\right]}{\mathbb{E}\left[\max\{D,\gamma\}\right]} \\ \overset{(c)}{\geq}&\frac{\mathbb{E}\left[\frac{1}{2}(\max\{D,\gamma^*\})^2 - \gamma^*\cdot\max\{D,\gamma^*\} + C\right]}{\mathbb{E}\left[\max\{D,\gamma\}\right]} \\ \overset{(d)}{=}&0,\end{aligned}\tag{24}$$

where inequality (c) is because policy $\tilde{\pi}_{\gamma^*}(D) = (\gamma^* - D)^+$ is the optimum policy to minimize $g(\pi, \gamma^*)$ and equality (d) is obtained because of (22). Since the stationary point $\gamma^*$ exists, combining (23) and (24) leads to the conclusion that $\mathcal{T}(\cdot)$ is a sub-contraction mapping:

$$|\mathcal{T}(\gamma) - \gamma^*| < (\gamma - \gamma^*), \forall \gamma > \gamma^*.$$

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

$$\lim_{\tau\to\infty} \mathcal{T}^{(\tau)}(\gamma) = \gamma^*. \tag{25}$$

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

$$\begin{aligned}&\mathcal{T}(\gamma) - \gamma^* = \frac{\mathbb{E}\left[\frac{1}{2}(\max\{D,\gamma\})^2 + C\right]}{\mathbb{E}\left[\max\{D,\gamma\}\right]} - \gamma^* \\ \overset{(e)}{>}&\frac{\mathbb{E}\left[\frac{1}{2}(\max\{D,\gamma\})^2 + C\right]}{\mathbb{E}\left[\max\{D,\gamma^*\}\right]} - \gamma^* \\ =&\frac{\mathbb{E}\left[\frac{1}{2}(\max\{D,\gamma\})^2 - \gamma^*\cdot\max\{D,\gamma\} + C\right]}{\mathbb{E}\left[\max\{D,\gamma^*\}\right]} \\ &+ \gamma^*\left(\frac{\mathbb{E}\left[\max\{D,\gamma\}\right]}{\mathbb{E}\left[\max\{D,\gamma^*\}\right]} - 1\right) \\ \overset{(f)}{\geq}&\frac{\mathbb{E}\left[\frac{1}{2}(\max\{D,\gamma^*\})^2 - \gamma^*\cdot\max\{D,\gamma^*\} + C\right]}{\mathbb{E}\left[\max\{D,\gamma^*\}\right]} \\ &+ \gamma^*\left(\frac{\mathbb{E}\left[\max\{D,\gamma\}\right]}{\mathbb{E}\left[\max\{D,\gamma^*\}\right]} - 1\right) \\ \overset{(g)}{=}&\frac{\gamma^*}{\mathbb{E}\left[\max\{D,\gamma^*\}\right]}\left(\mathbb{E}\left[\max\{D,\gamma\}\right] - \mathbb{E}\left[\max\{D,\gamma^*\}\right]\right) \\ \overset{(h)}{\geq}&(\gamma - \gamma^*),\end{aligned}\tag{26}$$

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

$$\begin{aligned}&\mathbb{E}\left[\max\{D,\gamma\}\right] - \mathbb{E}\left[\max\{D,\gamma^*\}\right] \\ \geq&\mathbb{E}\left[(\gamma - \gamma^*)\right] = \gamma - \gamma^*.\end{aligned}$$

For the case that $\gamma < \gamma^*$, denote $\iota$ be the stopping time that:

$$\iota(\gamma) = \arg\min_{\tau\in\mathbb{N}^+}\{\mathcal{T}^{(\tau)}(\gamma) \geq \gamma^*\}.$$

If $\iota(\gamma) < \infty$, then $\mathcal{T}^{(\iota(\tau))} > \gamma^*$, and according to (25) we have

$$\lim_{\tau\to\infty} \mathcal{T}^{(\tau)}(\mathcal{T}^{(\iota(\gamma))}(\gamma)) = \gamma^*.$$

Otherwise, if $\iota(\gamma) = \infty$, due to (26), the mapping is subcontract. Since the stationary point exists, $\lim_{\tau\to\infty} \mathcal{T}^{(\tau)}(\gamma) = \gamma^*$.

■

## APPENDIX C
## PROOF OF THEOREM 3

*Proof:* First, recall that the ratio $\gamma_k$ used in any frame $k$ is upper bounded by $\gamma_{\text{ub}}$, since the transmission delay is bounded $D_k \le B$, the frame length $L_k$ and the reward $R_k$ can be upper bounded by:

$$L_k \le D_k + (\gamma - D_k)^+ \le B + \gamma_{\text{ub}} =: L_{\text{ub}}, \tag{27a}$$

$$R_k = \frac{1}{2}L_k^2 + C \le L_{\text{ub}}^2 + C. \tag{27b}$$

Denote $\overline{L}^* := \mathbb{E}[D + \pi^*(D)]$ and $\overline{R}^* := \mathbb{E}[\frac{1}{2}(D + \pi^*(D))^2 + C]$ 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 $\mathbb{E}[L_k|\mathcal{F}_{k-1}]$ and the expected reward $\mathbb{E}[R_k|\mathcal{F}_{k-1}]$ of frame $k$ satisfies:

$$\mathbb{E}\left[R_k - \gamma_k L_k|\mathcal{F}_{k-1}\right] \le (\gamma^* - \gamma_k)\overline{L}^*, \tag{28a}$$

$$\mathbb{E}\left[R_k - \gamma^* L_k|\mathcal{F}_{k-1}\right] \le -(\gamma^* - \gamma_k)\left(\mathbb{E}[L_k|\mathcal{F}_{k-1}] - \overline{L}^*\right). \tag{28b}$$

*Lemma 2:* Recall that $Y_k = R_k + L_{k-1}D_k$ is the sum of cumulative AoI and extra cost within frame $k$ according to (8). Denote

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

Then, $\Delta_K$ can be upper bounded by:

$$\Delta_K \le \mathbb{E}\left[\sum_{k=1}^{K}(\gamma^* - \gamma_k)^2\right]. \tag{29}$$

Notice that the average cost deviation $\overline{h}_K - h_{\pi^*} = \frac{1}{\mathbb{E}[\sum_{k=1}^{K} L_k]}\Delta_K$, it is suffice 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:

$$\mathbb{E}[(\gamma_k - \gamma^*)^2] \le \frac{1}{k}\frac{(L_{\text{ub}}^2 + C)^2}{\overline{D}_{\text{lb}}^2}. \tag{30}$$

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

*Proof of* (17a): Plugging (3) into inequality (29) from Lemma 2, we can first upper bound $\Delta_K$ by:

$$\Delta_K \le \frac{(L_{\text{ub}}^2 + C)^2}{\overline{D}_{\text{lb}}^2}\left(\sum_{k=1}^{K}\frac{1}{k}\right)$$

$$\overset{(a)}{\le} \frac{(L_{\text{ub}}^2 + C)^2}{\overline{D}_{\text{lb}}^2}\left(1 + \int_{k=1}^{K}\frac{1}{k}\mathrm{d}k\right)$$

$$= \frac{(L_{\text{ub}}^2 + C)^2}{\overline{D}_{\text{lb}}^2}(1 + \ln K), \tag{31}$$

where inequality $(a)$ is because $\frac{1}{k} \le \int_{k-1}^{k}\frac{1}{x}\mathrm{d}x$.

Notice that:

$$\overline{h}_K - h_{\pi^*} = \frac{1}{\mathbb{E}\left[\sum_{k=1}^{K} L_k\right]}\Delta_K \overset{(b)}{\le} \frac{(L_{\text{ub}}^2 + C)^2}{\overline{D}\overline{D}_{\text{lb}}^2}\frac{(1 + \ln K)}{K}, \tag{32}$$

where inequality $(b)$ is because

$$\mathbb{E}\left[\sum_{k=1}^{K} L_k\right] \ge \mathbb{E}\left[\sum_{k=1}^{K} D_k\right] \ge K\overline{D}_{\text{lb}}.$$

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

$$h_{\pi_K} = \frac{\mathbb{E}[\frac{1}{2}((\gamma_K - D)^+ + D)^2 + C]}{\mathbb{E}[(\gamma_K - D)^+ + D]} + \overline{D},$$

which is smaller or equal to the minimum cost, i.e., $h_{\pi_K} \ge h_{\pi^*}$. Therefore for each $\gamma_K$, we can upper bound the gap between policy $\pi_K$ and $\pi^*$ by:

$$\begin{aligned}
&h_{\pi_K} - h_{\pi^*} = h_{\pi_K} - h_{\pi^*}\\
=&\frac{\mathbb{E}\left[\frac{1}{2}((\gamma_K - D)^+ + D)^2 + C\right]}{\mathbb{E}[(\gamma_K - D)^+ + D]} - \gamma^*\\
=&\frac{\mathbb{E}\left[\frac{1}{2}((\gamma_K - D)^+ + D)^2 + C - \gamma_K((\gamma_K - D)^+ + D)\right]}{\mathbb{E}\left[(\gamma_K - D)^+ + D\right]}\\
&+ (\gamma_K - \gamma^*)\\
\le&\frac{\mathbb{E}[\frac{1}{2}((\gamma^* - D)^+ + D)^2 + C - \gamma_K((\gamma^* - D)^+ + D)]}{\mathbb{E}[(\gamma_K - D)^+ + D]}\\
&+ (\gamma_K - \gamma^*)\\
\le&\frac{(\gamma^* - \gamma_K)\mathbb{E}[(\gamma_K - D)^+ + D]}{\mathbb{E}[(\gamma_K - D)^+ + D]} + (\gamma_K - \gamma^*)\\
=&\frac{(\gamma^* - \gamma_K)}{\mathbb{E}[(\gamma_K - D)^+ + D]}\mathbb{E}[(\gamma_K - D)^+ - (\gamma^* - D)^+]\\
\le&\frac{1}{\overline{D}}(\gamma_K - \gamma^*)^2.
\end{aligned} \tag{33}$$

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

$$\mathbb{E}[h_{\pi_K} - h_{\pi^*}] \le \frac{(L_{\text{ub}}^2 + C)^2}{\overline{D}\overline{D}_{\text{lb}}^2}\frac{1}{K}.$$

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:

$$\mathbb{E}\left[R_k - \gamma_k L_k|\mathcal{F}_{k-1}\right] \overset{(a)}{\le} (\overline{R}^* - \gamma_k\overline{L}^*)$$

$$=(\overline{R}^* - \gamma^*\overline{L}^*) + (\gamma^* - \gamma_k)\overline{L}^* \overset{(b)}{=} (\gamma^* - \gamma_k)\overline{L}^*, \quad (34)$$

where equality $(a)$ is because policy $\pi_k$ used in frame $k$ satisfies $g(\pi_k, \gamma_k) \leq g(\pi^*, \gamma_k)$. Equality $(b)$ is obtained because on the stationary point $\gamma^*$ we have $\overline{R}^* = \gamma^*\overline{L}^*$. This verifies the first inequality in Lemma 1.

Then, adding $(\gamma_k - \gamma^*)\mathbb{E}[L_k|\mathcal{F}_{k-1}]$ to both sides of (34), we have:

$$\mathbb{E}\left[R_k - \gamma^* L_k|\mathcal{F}_{k-1}\right] \leq (\gamma_k - \gamma^*)\mathbb{E}\left[L_k - \overline{L}^*|\mathcal{F}_{k-1}\right]. \quad (35)$$

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 $\mathbb{E}[L_{k-1}D_k|\mathcal{F}_{k-1}]$ on both sides of (28b). By replacing $R_k + L_{k-1}D_k$ with $Y_k$, we then have the following inequality:

$$\begin{aligned}&\mathbb{E}[Y_k - \gamma^* L_k|\mathcal{F}_{k-1}]\\ &\leq -(\gamma^* - \gamma_k)\left(\mathbb{E}[L_k|\mathcal{F}_{k-1}] - \overline{L}^*\right) + \mathbb{E}[L_{k-1}|\mathcal{F}_{k-1}]\overline{D}\\ &\overset{(d)}{\leq}(\gamma^* - \gamma_k)^2 + \mathbb{E}[L_{k-1}|\mathcal{F}_{k-1}]\overline{D}, \end{aligned} \quad (36)$$

where inequality $(d)$ is because

$$\begin{aligned}\mathbb{E}[L_k - \overline{L}^*|\mathcal{F}_{k-1}] &= \mathbb{E}\left[(\gamma_k - D)^+ - (\gamma^* - D)^+\right]\\ &\leq |\gamma_k - \gamma^*|. \end{aligned} \quad (37)$$

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

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}\left(Y_k - \left(\gamma^* + \overline{D}\right)L_k\right)\right]\\ \leq&\mathbb{E}\left[\sum_{k=1}^{K}(\gamma^* - \gamma_k)^2\right] - \mathbb{E}[L_K]\overline{D}. \end{aligned} \quad (38)$$

And this completes the proof of Lemma 2. ■

## Appendix F
## Proof of Lemma 3

*Proof:* For simplicity, denote

$$z_{k+1} := \gamma_k + \eta_k(R_k - \gamma_k L_k). \quad (39)$$

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

$$(\gamma_{k+1} - \gamma^*)^2 = ([z_{k+1}]_{\gamma_{\text{lb}}}^{\gamma_{\text{ub}}} - [\gamma^*]_{\gamma_{\text{lb}}}^{\gamma_{\text{ub}}}) \leq (z_{k+1} - \gamma^*)^2. \quad (40)$$

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

$$\begin{aligned}&\frac{1}{2}(z_{k+1} - \gamma^*)^2\\ =&\frac{1}{2}\left(\gamma_k - \gamma^* + \eta_k\left(R_k - \gamma_k L_k\right)\right)^2\\ =&\frac{1}{2}(\gamma_k - \gamma^*)^2 + \frac{1}{2}\eta_k^2\left(R_k - \gamma_k L_k\right)^2\\ &+ \eta_k(\gamma_k - \gamma^*)\left(R_k - \gamma_k L_k\right)\\ \leq&\frac{1}{2}(\gamma_k - \gamma^*)^2 + \frac{1}{2}\eta_k^2(L_{\text{ub}}^2 + C)^2\\ &+ \eta_k(\gamma_k - \gamma^*)\left(R_k - \gamma_k L_k\right), \end{aligned} \quad (41)$$

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

$$\begin{aligned}&\frac{1}{2}\mathbb{E}\left[(z_{k+1} - \gamma^*)^2|\mathcal{F}_{k-1}\right]\\ \leq&\frac{1}{2}(\gamma_k - \gamma^*)^2 + \frac{1}{2}\eta_k^2(L_{\text{ub}}^2 + C)^2\\ &+ \eta_k(\gamma_k - \gamma^*)\mathbb{E}\left[R_k - \gamma_k L_k|\mathcal{F}_{k-1}\right]. \end{aligned} \quad (42)$$

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:

$$\begin{aligned}&(\gamma_k - \gamma^*)\mathbb{E}[R_k - \gamma_k L_k|\mathcal{F}_{k-1}]\\ \leq&-(\gamma_k - \gamma^*)^2\overline{L}^* \leq -(\gamma_k - \gamma^*)^2\overline{D}, \end{aligned} \quad (43)$$

  where the last inequality is obtained because $\overline{L}^* \geq \overline{D}$.
- If the current $\gamma_k - \gamma^* \leq 0$, we can upper the last term in inequality (42) as follows:

$$\begin{aligned}&(\gamma_k - \gamma^*)\mathbb{E}[R_k - \gamma_k L_k|\mathcal{F}_{k-1}]\\ =&(\gamma_k - \gamma^*)\mathbb{E}[R_k - \gamma^* L_k|\mathcal{F}_{k-1}]\\ &-(\gamma_k - \gamma^*)^2\mathbb{E}[L_k|\mathcal{F}_{k-1}]\\ \overset{(a)}{\leq}&(\gamma_k - \gamma^*)(\overline{R}^* - \gamma^*\overline{L}^*) - (\gamma_k - \gamma^*)^2\mathbb{E}[L_k|\mathcal{F}_{k-1}]\\ =&-(\gamma_k - \gamma^*)^2\mathbb{E}[L_k|\mathcal{F}_{k-1}]\\ \overset{(b)}{\leq}&-(\gamma_k - \gamma^*)^2\overline{D}, \end{aligned} \quad (44)$$

  where inequality $(a)$ is because $\mathbb{E}[R_k - \gamma^* L_k|\mathcal{F}_{k-1}] \geq \overline{R}^* - \gamma^*\overline{L}^* = 0$ and inequality $(b)$ is because $\mathbb{E}[L_k|\mathcal{F}_{k-1}] \geq \overline{D}$.

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

$$\begin{aligned}&\frac{1}{2}\mathbb{E}\left[(z_{k+1} - \gamma^*)^2|\mathcal{F}_{k-1}\right]\\ =&\left(\frac{1}{2} - \eta_k\overline{D}\right)(\gamma_k - \gamma^*)^2 + \frac{1}{2}\eta_k^2(L_{\text{ub}}^2 + C)^2\\ \leq&\left(\frac{1}{2} - \eta_k\overline{D}_{\text{lb}}\right)(\gamma_k - \gamma^*)^2 + \frac{1}{2}\eta_k^2(L_{\text{ub}}^2 + C)^2. \end{aligned} \quad (45)$$

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

$$\begin{aligned}&\frac{1}{2}\mathbb{E}\left[(\gamma_{k+1} - \gamma^*)^2\right] \overset{(c)}{\leq} \frac{1}{2}\mathbb{E}[(z_{k+1} - \gamma^*)^2]\\ \leq&\frac{1}{2}\left(1 - 2\eta_k\overline{D}_{\text{lb}}\right)\mathbb{E}\left[(\gamma_k - \gamma^*)^2\right] + \frac{1}{2}\eta_k^2(L_{\text{ub}}^2 + C)^2, \end{aligned} \quad (46)$$

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

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

$$\frac{1}{2}\mathbb{E}[(\gamma_k - \gamma^*)^2] \leq \frac{1}{2k}\frac{(L_{\text{ub}}^2 + C)^2}{\overline{D}_{\text{lb}}^2}. \tag{47}$$

Detailed proofs for (47) are as follow:

- When $k = 2$, by choosing $\eta_1 = \frac{1}{2\overline{D}_{\text{lb}}}$ we have

$$\frac{1}{2}\mathbb{E}[(\gamma_2 - \gamma^*)^2] \leq \frac{1}{8}\frac{(L_{\text{ub}}^2 + C)^2}{\overline{D}_{\text{lb}}^2} \leq \frac{1}{4}\frac{(L_{\text{ub}}^2 + C)^2}{\overline{D}_{\text{lb}}^2}.$$

- When $k > 2$, assuming that $\frac{1}{2}\mathbb{E}[(\gamma_k - \gamma^*)^2] \leq \frac{1}{2k}\frac{(L_{\text{ub}}^2+C)^2}{\overline{D}_{\text{lb}}^2}$, recall that the stepsize $\eta_k$ is chosen to be $\eta_k = \frac{1}{(k+2)\overline{D}_{\text{lb}}}$, we have:

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}\left[(\gamma_{k+1} - \gamma^*)^2\right] \\
\leq &\left(\frac{1}{2} - \eta_k\overline{D}_{\text{lb}}\right)(\gamma_k - \gamma^*)^2 + \frac{1}{2}\eta_k^2(L_{\text{ub}}^2 + C)^2 \\
\leq &\left(1 - \frac{2}{k+2}\right)\frac{1}{2k}\frac{(L_{\text{ub}}^2 + C)^2}{\overline{D}_{\text{lb}}^2} + \frac{1}{2}\frac{1}{(k+2)^2}\frac{(L_{\text{ub}}^2 + C)^2}{\overline{D}_{\text{lb}}^2} \\
= &\frac{1}{2}\left(\frac{1}{k+2} + \frac{1}{(k+2)^2}\right)\frac{(L_{\text{ub}}^2 + C)^2}{\overline{D}_{\text{lb}}^2} \\
= &\frac{1}{2}\frac{k+3}{(k+2)^2}\frac{(L_{\text{ub}}^2 + C)^2}{\overline{D}_{\text{lb}}^2} \\
\leq &\frac{1}{2}\frac{1}{(k+1)}\frac{(L_{\text{ub}}^2 + C)^2}{\overline{D}_{\text{lb}}^2},
\end{aligned} \tag{48}$$

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

■

## REFERENCES


[1] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *2012 Proceedings IEEE INFOCOM*, 2012, pp. 2731–2735.

[2] Y. Wang and W. Chen, "Adaptive power and rate control for real-time status updating over fading channels," *IEEE Transactions on Wireless Communications*, vol. 20, no. 5, pp. 3095–3106, 2021.

[3] B. Wang, S. Feng, and J. Yang, "When to preempt? age of information minimization under link capacity constraint," *Journal of Communications and Networks*, vol. 21, no. 3, pp. 220–232, 2019.

[4] B. Zhou and W. Saad, "Joint status sampling and updating for minimizing age of information in the internet of things," *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7468–7482, 2019.

[5] H. Tang, J. Wang, L. Song, and J. Song, "Minimizing age of information with power constraints: Multi-user opportunistic scheduling in multi-state time-varying channels," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 5, pp. 854–868, 2020.

[6] E. T. Ceran, D. Gündüz, and A. György, "Average age of information with hybrid arq under a resource constraint," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, 2018, pp. 1–6.

[7] M. A. Abd-Elmagid, H. S. Dhillon, and N. Pappas, "A reinforcement learning framework for optimizing age of information in rf-powered communication systems," *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 4747–4760, 2020.

[8] A. Arafa, J. Yang, S. Ulukus, and H. V. Poor, "Online timely status updates with erasures for energy harvesting sensors," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2018, pp. 966–972.

[9] A. M. Bedewy, Y. Sun, S. Kompella, and N. B. Shroff, "Optimal sampling and scheduling for timely status updates in multi-source networks," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 4019–4034, 2021.

[10] A. Soysal and S. Ulukus, "Age of information in g/g/1/1 systems," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 2022–2027.

[11] E. Najm, R. Nasser, and E. Telatar, "Content based status updates," *IEEE Transactions on Information Theory*, vol. 66, no. 6, pp. 3846–3863, 2020.

[12] R. D. Yates, "Lazy is timely: Status updates by an energy harvesting source," in *2015 IEEE International Symposium on Information Theory (ISIT)*, 2015, pp. 3008–3012.

[13] Y. Sun, E. Uysal-Biyikoglu, R. D. Yates, C. E. Koksal, and N. B. Shroff, "Update or wait: How to keep your data fresh," *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 7492–7508, 2017.

[14] E. T. Ceran, D. Gündüz, and A. György, "Reinforcement learning to minimize age of information with an energy harvesting sensor with harq and sensing cost," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019, pp. 656–661.

[15] ——, "A reinforcement learning approach to age of information in multi-user networks with harq," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 5, pp. 1412–1426, 2021.

[16] C. Kam, S. Kompella, and A. Ephremides, "Learning to sample a signal through an unknown system for minimum aoi," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019, pp. 177–182.

[17] S. Leng and A. Yener, "Age of information minimization for wireless ad hoc networks: A deep reinforcement learning approach," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[18] K. Bhandari, S. Fatale, U. Narula, S. Moharir, and M. K. Hanawal, "Age-of-information bandits," in *2020 18th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, 2020, pp. 1–8.

[19] E. U. Atay, I. Kadota, and E. Modiano, "Aging bandits: Regret analysis and order-optimal learning algorithm for wireless networks with stochastic arrivals," 2020.

[20] S. Banerjee, R. Bhattacharjee, and A. Sinha, "Fundamental limits of age-of-information in stationary and non-stationary environments," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 1741–1746.

[21] V. Tripathi and E. Modiano, "An online learning approach to optimizing time-varying costs of aoi," 2021.

[22] B. Li, "Efficient learning-based scheduling for information freshness in wireless networks," in *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, 2021.

[23] C.-H. Tsai and C.-C. Wang, "Age-of-information revisited: Two-way delay and distribution-oblivious online algorithm," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 1782–1787.

[24] A. Arafa, R. D. Yates, and H. V. Poor, "Timely cloud computing: Preemption and waiting," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2019, pp. 528–535.

[25] H. Robbins and S. Monro, "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400 – 407, 1951.

[26] M. J. Neely, "Fast learning for renewal optimization in online task scheduling," 2021.

[27] A. A. Goldstein, "Constructive real analysis," Harper's Series in Modern Mathematics. New York-Evanston-London: Harper and Row, Publishers. XII, 178 p. (1967)., 1967.