entropy

MDPI

Article

Optimal Information Update for Energy Harvesting Sensor with Reliable Backup Energy

Lixin Wang [1], Fuzhou Peng [2], Xiang Chen [2] and Shidong Zhou [1,*]

[1] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China; wanglx19@mails.tsinghua.edu.cn
[2] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China; pengfzh@mail2.sysu.edu.cn (F.P.); chenxiang@mail.sysu.edu.cn (X.C.)
* Correspondence: zhousd@tsinghua.edu.cn

Abstract: The timely delivery of status information collected from sensors is critical in many real-time applications, e.g., monitoring and control. In this paper, we consider a scenario where a wireless sensor sends updates to the destination over an erasure channel with the supply of harvested energy and reliable backup energy. We adopt the metric age of information (AoI) to measure the timeliness of the received updates at the destination. We aim to find the optimal information updating policy that minimizes the time-average weighted sum of the AoI and the reliable backup energy cost. First, when all the environmental statistics are assumed to be known, the optimal information updating policy exists and is proved to have a threshold structure. Based on this special structure, an algorithm for efficiently computing the optimal policy is proposed. Then, for the unknown environment, a learning-based algorithm is employed to find a near-optimal policy. The simulation results verify the correctness of the theoretical derivation and the effectiveness of the proposed method.

Keywords: age of information; information update; energy harvesting; reliable backup energy

Citation: Wang, L.; Peng, F.; Chen, X.; Zhou, S. Optimal Information Update for Energy Harvesting Sensor with Reliable Backup Energy. *Entropy* **2022**, *24*, 961. https://doi.org/10.3390/e24070961

Academic Editors: Udo Von Toussaint and Chintha Tellambura

Received: 22 February 2022
Accepted: 7 July 2022
Published: 11 July 2022

1. Introduction

Timely information updates from wireless sensors to destinations are essential for real-time monitoring and control systems. To describe the timeliness of information updates from the receivers' perspective, a new metric called age of information (AoI) is proposed [1–3]. Unlike general performance metrics, such as delay and throughput, AoI refers to the time elapsed since the generation of the latest received information. A lower AoI generally reflects more timely information at the destination. Therefore, the AoI-minimal status updating policies in sensor networks have been widely studied [4–7].

The destinations always desire information updates in as timely a manner as possible, which is typically constrained by sensors' energy. Generally, energy sources include the grid and sensors' own non-rechargeable batteries. We call these sources *reliable energy* since they enable sensors to reliably operate until the power grid is cut off or sensors' batteries are exhausted [8]. Specifically, if sensors consume energy from the grid, they need to pay the electricity bill; if sensors only use the power of their own batteries, the price of sensing and transmitting updates will be the cost of frequent battery replacement. There is clearly a price to pay for using reliable energy to update. Energy harvesting (EH) is a promising technology that can help reduce the consumption of reliable energy for information update [9,10]. It can continuously extract energy from solar power, ambient RF, and thermal energy and store the harvested energy in sensors' rechargeable batteries. The stored energy is renewable and can be used for free. Hence, in this case, the reliable energy can serve as *backup* energy. The design of the coexistence of reliable backup energy and harvested energy has been researched and promoted in academia and industry [8,11–14]. The mixed energy supply mode can enhance the reliability of the system.

However, the irregular arrivals of harvested energy and the limited capacity of rechargeable batteries still motivate us to schedule the energy usage properly to reduce

the cost of using reliable backup energy while maintaining the timeliness of information updates (i.e., the average AoI). Intuitively, the average AoI and the cost of using reliable energy cannot be minimized simultaneously. On the one hand, a lower average AoI means that the sensor senses and transmits updates more frequently, which will increase the consumption of reliable backup energy since the harvested energy is limited. On the other hand, to reduce the cost of reliable backup energy, the sensor will only exploit the harvested energy. Due to the uncertainty of the energy harvesting behavior, the average AoI of the system will inevitably increase. Therefore, in this paper, we focus on achieving the best trade-off between the average AoI and the cost of reliable backup energy.

We consider a sensor-based information update system, where an energy harvesting sensor with reliable backup energy sends timely updates to the destination through an erasure channel. Based on our settings, we will minimize the long-term average weighted sum of the AoI and the paid reliable energy cost to find the optimal information updating policy by Markov decision process (MDP) theory [15]. First, we assume that the sensor knows the relevant statistics in advance, such as the success probability of each transmission and the probability of energy arrival, so that the sensor can make the optimal decision at any time. Then we consider a more realistic scenario where the sensor has no knowledge of the environment. In such an unknown environment, learning-based approaches should be adopted to obtain the updating policy.

1.1. Related Work

There have been a series of related works studying AoI minimization in EH communication systems [16–34]. In these systems, each update consumes harvested energy and is constrained by the energy causality.

Refs. [16–23] focus on how to optimize AoI under general energy causality constraints, where different battery model settings are considered. Constrained by the average power available in the infinite-sized battery, ref. [16] shows that a lazy policy which leaves a certain idle period between updates outperforms the greedy policy under random service times. With the same assumption of an infinite-sized battery, ref. [17] focuses on both offline and online policies under energy replenishment constraints with zero service time. While considering fixed service times, the offline results in [17] are extended to a two-hop scenario in [18], and online policy is provided in [19]. In the case of the delay being controlled by transmission energy, ref. [20] also investigated the optimal offline policy. For the error-free and delay-free channel, the optimal updating policies were investigated for different battery settings [21,22]. Ref. [21] derived the asymptotically optimal policies for the infinite-sized, finite-sized, and unit-sized battery by renewal theory. It turned out to be a threshold policy for the unit-sized battery case. More general battery models were considered in [22]. The optimal policy was also proved to be multi-threshold and the energy-dependent thresholds were characterized explicitly. When the battery is finite sized and there is no feedback from the destination, it was shown that the optimal updating policy is of a threshold structure and the threshold is non-increasing with the battery level [23].

Refs. [24–30] studied how to properly utilize the harvested energy to transmit updates over imperfect channels. For the noisy channel, ref. [24] considered an infinite-sized battery model and derived the different optimal policies for updating with and without feedback. Ref. [25] further derived a closed-form expression for the threshold of the unit-sized battery model and extended the threshold-based policies to multiple sources case. To combat the noisy channel, some channel coding schemes for EH communication were investigated in [26,27]. In [28], the HARQ protocol was applied for a single EH sensor to send updates to the destination. The optimal policies were obtained by employing reinforcement learning in both known and unknown environments, but no clear intuition on the policy structure was provided. Considering energy harvesting wireless sensor networks (EH-WSNs), ref. [29] suggested to estimate the channel state of a Rayleigh fading channel before transmitting to improve the AoI, update interval and packet loss performance, despite the associated time and energy costs. Ref. [30] aimed to minimize the average AoI of an EH-aided secondary

user(SU) in a cognitive ratio network. The SU has to make sensing and updating decisions subject to random energy arrivals and the available spectrum. The sequential decision problem is formulated as a partially observable Markov decision process (POMDP).

Refs. [31–34] paid attention to other AoI-related metrics in EH communication and even the distributional properties of AoI, not just the average AoI. Different freshness metrics were considered, such as nonlinear AoI [31], urgency-aware AoI (U-AoI) [32], and peak AoI [33] in EH sensor network. To better understand the distributional properties of AoI, ref. [34] further derived closed-form expressions of the moment generating function (MGF) of AoI in an EH-powered queuing system using the stochastic hybrid systems (SHS) framework.

The above works focus on optimizing information freshness under the EH supply. Different from them, energy sources in this paper include both harvested energy and reliable backup energy, and our goal is to achieve the best trade-off between age and reliable energy consumption, instead of merely optimizing AoI. Among the above works, refs. [23,25] are the most related to our paper. The following Table 1 summarizes the detailed differences. It is worth noting that by letting the reliable energy consumption be small enough, our results can be compared with some prior results in [23,25].

Table 1. Comparative summary of the most related works in contrast to our paper.

Feature \ Ref	[23]	[25]	Our
Energy supply	EH	EH	EH + reliable energy
Battery capacity	Finite-sized	Unit-sized	Finite-sized
Wireless channel	Error-free	Error-prone	Error-prone
Optimization objective	AoI	AoI	AoI-reliable energy trade-off

The age–energy trade-off has been widely studied in [35–39]. The age–energy trade-off in the erasure channel was studied in [35], and the fading channel case was investigated in [36]. Ref. [37] adopted a truncated automatic repeat request (TARQ) scheme and characterized the age–energy trade-off for the IoT monitoring system. Optimum energy efficiency and AoI trade-off was considered in a multicast system in [38]. In [39] , the authors investigated the optimal age–energy trade-off, where status sensing and data transmission can be carried out separately. By the MDP analysis similar to [6,15], the optimal policy exists and is proved to have two thresholds. The energy sources are all reliable in these works, which means that the energy cost of the update is easy to track. However, the uncertainty of the energy arrival and mixed energy supplies bring more challenges to the MDP analysis in this paper. To the best of our knowledge, this paper is the first to consider the timeliness of the system under mixed energy supplies. The preliminary results of this paper are presented in [40].

1.2. Main Contributions

The main contributions of this paper are as follows:

- We consider an information update system where the harvested energy and reliable energy coexist. The goal is to find the optimal policy that achieves the best trade-off between age and reliable energy consumption. Compared to the existing works [23,25], our problem is more practical and general, which will provide some insights for future green and durable update system designs.
- For the case that all the statistics such as channel erasure probability and EH probability are known a priori, we formulate an unconstrained infinite space Markov decision process (MDP) problem, and prove the existence of the optimal policy. By revealing the monotonicity and *proportional differential property* of the value function, we find

that the optimal policy is of the threshold-type. Based on this special structure, we propose an efficient algorithm to compute the optimal policy.
- In an unknown environment, we propose an average cost Q-learning algorithm to obtain the updating policy.
- Simulation results show that the optimal policy outperforms other baseline policies when the environmental statistics are known. At the same time, the performance of the policy learned in the unknown environment is very close to the theoretical optimal policy. We also compare the age-reliable energy trade-off curves of the optimal updating policies under different energy supply conditions, which reflects the rationality of mixed energy supplies. The optimal policy can also be particularized to a special case, where the sensor can only utilize the harvested energy and the battery is unit-sized, and its performance coincides with the existing results in [23,25].

1.3. Organization

The rest of this paper is organized as follows. In Section 2, we introduce the model of the information update system and formulate the problem. In Section 3, we analyze the optimal policy when all the statistics are known. In Section 4, we aim to minimize the average cost of updating in an unknown environment. In Section 5, we present the simulation results. Finally, in Section 6, we conclude the paper.

2. System Model and Problem Formulation

2.1. System Model

In this paper, we consider a point-to-point information update system, where a wireless sensor and a destination are connected by an erasure channel, as shown in Figure 1. The channel is assumed to be noisy and time invariant, and each update is corrupted with probability p during transmission (Note $p \in (0,1)$). Both the free harvested energy stored in the rechargeable battery and the reliable backup energy that needs to be paid can be used for real-time environmental status updates.

Without loss of generality, time is slotted with equal length and indexed by $t = 0, 1, 2 \ldots$. At the beginning of each time slot, the sensor decides whether to generate and transmit an update to the destination or stay idle. The decision action at slot t, denoted by $a[t]$, takes value from action set $\mathcal{A} = \{0, 1\}$, where $a[t] = 1$ means that the sensor decides to generate and transmit an update to the destination while $a[t] = 0$ means the sensor is idle. The destination will feed back an instantaneous ACK to the sensor through an error-free channel when it has successfully received an update and a NACK otherwise. We assume the above processes can be completed in one time slot. The destination keeps track of the environment status through the received updates. We apply the metric age of information to measure the freshness of the status information available at the destination.

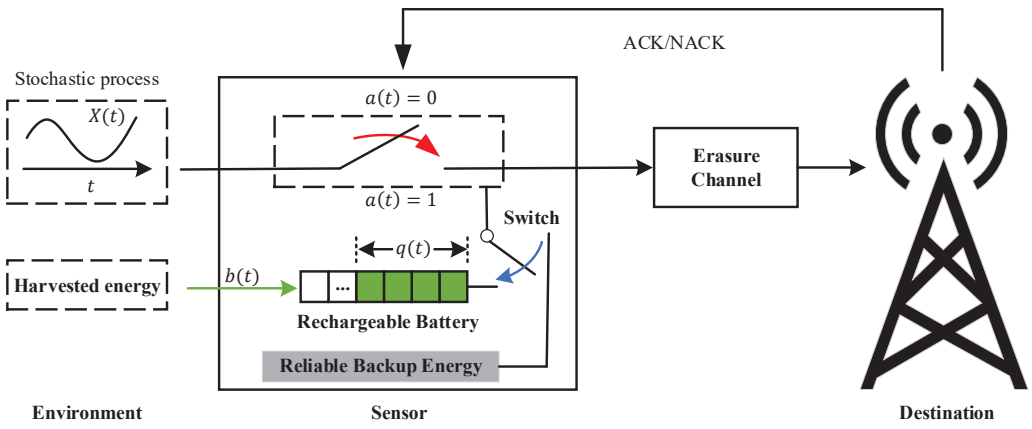

Figure 1. System model.

2.1.1. Age of Information

Age of Information (AoI) is defined as the elapsed time since the generation of the latest successfully received update [1–3]. Denote $\Delta[t]$ as the AoI of destination in time slot t. Then, we have

$$\Delta[t] = t - U[t]. \tag{1}$$

where $U[t]$ denotes the time slot when the most recently received update was generated before time slot t. In particular, the AoI will decrease to one if a new update is successfully received. Otherwise, it will increase by one. The evolution of AoI can be expressed as follows:

$$\Delta[t+1] = \begin{cases} 1, & \text{successful transmission,} \\ \Delta[t]+1, & \text{otherwise.} \end{cases} \tag{2}$$

A sample path of AoI is depicted in Figure 2.

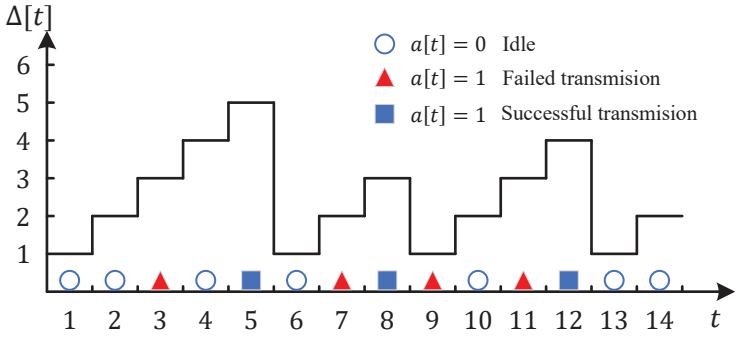

Figure 2. A sample path of AoI with initial age 1.

2.1.2. Description of Energy Supply

We assume that only the sensor's measurement and transmission process will consume energy and ignore other energy consumption. The energy unit is normalized, so the generation and transmission for each update will consume one energy unit. As previously described, the energy sources of the sensor include energy harvested from nature and reliable backup energy.

For the harvested energy, the sensor can store it in a rechargeable battery for later use. The maximum capacity of the rechargeable battery is B units ($B > 1$). Considering the scarcity of energy in nature, the total energy harvested in one time slot may sometimes not reach an energy unit. So we consider using the Bernoulli process with the parameter λ to approximately capture the arrival process of harvested energy, which was also adopted in [41–43]. Let $b[t]$ be the accumulated harvested energy in time slot t. That is, we have $\Pr\{b[t] = 1\} = \lambda$ and $\Pr\{b[t] = 0\} = 1 - \lambda$ in each time slot t (note $\lambda \in (0, 1)$). Here, we assume that the energy arrival at each slot is independently and identically distributed. Time-correlated energy arrival processes, such as Markov process, will be considered in future work.

For reliable backup energy, we assume that it contains much more energy units than the rechargeable battery, so the energy it contains can be viewed as infinite. However, it needs to be used for a fee compared to the free renewable energy stored in the rechargeable battery. Therefore, when the stored renewable energy is not zero, the sensor will prioritize using it for status updates; otherwise, it will automatically switch to the reliable backup energy until the sensor has harvested energy. Defining the power of the rechargeable battery at the beginning of time slot t as the battery state $q[t]$, then the evolution of battery state from time slot t to $t+1$ can be summarized as follows:

$$q[t+1] = \min\{q[t] + b[t] - a[t]u(q[t]), B\}, \tag{3}$$

where $u(\cdot)$ is unit step function, which is defined as

$$u(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Suppose that under reliable energy supply, the cost of generating and transmitting an update is a non-negative value C_r. Defining $E[t]$ as the paid reliable energy cost at the time slot t, then we have

$$E[t] = C_r a[t](1 - u(q[t])). \tag{5}$$

2.2. Problem Formulation

Let Π denote the set of non-anticipative policies in which scheduling decision $a[t]$ are made based on the action history $\{a[k]\}_{k=0}^{t-1}$, the evolution of AoI $\{\Delta[k]\}_{k=0}^{t}$, the evolution of battery state $\{q[k]\}_{k=0}^{t}$ as well as the system parameters (e.g., p and λ). In order to keep the information freshness at the destination, the sensor needs to send updates. However, due to the randomness of energy arrivals, the battery energy may sometimes be insufficient to support updates, and the sensor has to take energy from reliable backup energy. To balance the information freshness and the paid reliable backup energy cost, we aim to find the optimal information updating policy $\pi \in \Pi$ that achieves the minimum of the time-average weighted sum of the AoI and the paid reliable backup energy cost. The problem is formulated as follows:

$$\begin{aligned} &\min_{\pi \in \Pi} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_\pi \left\{ \sum_{t=0}^{T-1} [\Delta[t] + \omega E[t]] \right\}, \\ &\text{s.t.} \qquad (2), (3), (5), \end{aligned} \tag{6}$$

where ω is the pre-defined non-negative weighting factor. If $\omega = 0$, the optimal policy is to update in each time slot, i.e., zero-wait policy [4]. Since the effect of energy can be ignored, if the rechargeable battery is not empty, the sensor uses the renewable energy; otherwise, the sensor will use the reliable energy directly. When $\omega > 0$, the optimal policy is non-trivial. So we will focus on the optimal policy for $\omega > 0$ in the rest of the paper. The smaller ω is, the more we attach importance to the system AoI; otherwise, the more emphasis is placed on the cost of reliable energy.

Remark 1. *The optimal trade-off between age and reliable energy consumption can also be formulated as a constrained problem, where the reliable energy consumption serves as a constraint (not exceeding E_m) but not a penalty, and the goal is to minimize the long-term average age. By the Lagrangian method, it can be converted into an unconstrained weighted sum problem, where the Lagrangian multiplier is exactly the weight factor ω. So the solution proposed in this paper can be used. If there exists an ω such that the average reliable energy consumption in the minimum weighted sum is E_m, the optimal policy of the weighted sum problem also minimizes the long-term average age with the E_m constraint. Otherwise, a randomized optimal policy for the constrained problem needs to be considered; see details in [44].*

3. Optimal Policy Analysis In A Known Environment

In this section, we aim to solve the problem (6) in a known environment and obtain the optimal policy. It is difficult to solve the original problem directly due to the random erasures and the temporal dependency in both AoI and battery state evolution. However, since the statistics such as channel erasure probability and EH probability are known, we can reformulate the original problem as a time-average cost MDP with infinite state space and analyze the structure of the optimal policy.

3.1. Markov Decision Process Formulation

According to the system description mentioned in the previous section, the MDP is formulated as follows:

- **State space** . The sensor's state $\mathbf{x}[t]$ in slot t is a couple of the current destination AoI and the battery state, i.e., $(\Delta[t], q[t])$. Define $\mathcal{B} = \{0, 1, \ldots, B\}$. The state space $\mathcal{S} = \mathbb{Z}^+ \times \mathcal{B}$ is thus infinite countable.
- **Action space**. The sensor's action $a[t]$ in time slot t takes value from the action set $\mathcal{A} = \{0, 1\}$.
- **Transition probabilities**. Denote $\Pr(\mathbf{x}[t+1]|\mathbf{x}[t], a[t])$ as the transition probability that current state $\mathbf{x}[t]$ transits to next state $\mathbf{x}[t+1]$ after taking action $a[t]$. Suppose the current state $\mathbf{x}[t] = (\Delta, q)$ and action $a[t] = a$, then the transition probability is divided into two following cases conditioned on different values of action.
  Case 1 . $a = 0$,

$$\begin{cases} \Pr\{(\Delta+1, q+1)|(\Delta, q), 0\} = \lambda, & \text{if } q < B, \\ \Pr\{(\Delta+1, B)|(\Delta, B), 0\} = 1, & \text{if } q = B, \\ \Pr\{(\Delta+1, q)|(\Delta, q), 0\} = 1-\lambda, & \text{if } q < B. \end{cases} \tag{7}$$

  Case 2. $a = 1$,

$$\begin{cases} \Pr\{(\Delta+1, q)|(\Delta, q), 1\} = p\lambda, & \text{if } q > 0, \\ \Pr\{(1, q)|(\Delta, q), 1\} = (1-p)\lambda, & \text{if } q > 0, \\ \Pr\{\Delta+1, q-1)|(\Delta, q), 1\} = p(1-\lambda), & \text{if } q > 0, \\ \Pr\{(1, q-1)|(\Delta, q), 1\} = (1-p)(1-\lambda), & \text{if } q > 0, \\ \Pr\{(\Delta+1, 1)|(\Delta, 0), 1\} = p\lambda, & \text{if } q = 0, \\ \Pr\{(1, 1)|(\Delta, 0), 1\} = (1-p)\lambda, & \text{if } q = 0, \\ \Pr\{(\Delta+1, 0)|(\Delta, 0), 0\} = p(1-\lambda), & \text{if } q = 0, \\ \Pr\{(1, 0)|(\Delta, 0), 0\} = (1-p)(1-\lambda), & \text{if } q = 0. \end{cases} \tag{8}$$

  In both cases, the evolution of AoI still follows Equation (2) and the evolution of battery state follows Equation (3).
- **One-step cost**. For the current state $\mathbf{x} = (\Delta, q)$, the one-step cost $C(\mathbf{x}, a)$ of taking action a is expressed by

$$C(\mathbf{x}, a) = \Delta + \omega C_r a(1 - u(q)). \tag{9}$$

After the above modeling, the original problem (6) is transformed into obtaining the optimal policy for the MDP to minimize the average cost in an infinite horizon:

$$\min_{\pi \in \Pi} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_\pi \left\{ \sum_{t=0}^{T-1} C(\mathbf{x}[t], a[t]) \right\}. \tag{10}$$

Denote Π_{SD} as the set of stationary deterministic policies. Given observation $(\Delta[t], q[t]) = (\Delta, q)$, the policy $\pi \in \Pi_{SD}$ selects action $a[t] = \pi(\Delta, q)$, where $\pi(\cdot) : (\Delta, q) \to \{0, 1\}$ is a deterministic function from state space $\mathcal{S}$ to action space $\mathcal{A}$. In the next section, we prove that there is an optimal stationary deterministic policy for the above unconstrained MDP with infinite countable state and action space.

3.2. The Existence of the Optimal Stationary Deterministic Policy

According to [15], we need to first address a discounted cost MDP, then relate it to the original average cost problem. Given an initial state $\mathbf{x}[0] = \hat{\mathbf{x}}$, the total expected discounted cost under a policy π is given by

$$V_\gamma^\pi(\hat{\mathbf{x}}) = \limsup_{T\to\infty} \mathbb{E}_\pi \left\{ \sum_{t=0}^{T-1} \gamma^t C(\mathbf{x}[t], a[t]) \middle| \mathbf{x}[0] = \hat{\mathbf{x}} \right\}, \tag{11}$$

where the discounted factor is $\gamma \in (0,1)$. Therefore, the problem of minimizing the expected discounted cost can be formulated as

$$V_\gamma(\hat{\mathbf{x}}) \triangleq \min_{\pi\in\Pi} V_\gamma^\pi(\hat{\mathbf{x}}), \tag{12}$$

where value function $V_\gamma(\hat{\mathbf{x}})$ denotes the minimum expected discounted cost. The policy is γ-optimal if it minimizes the above discounted cost. The optimality equation of $V_\gamma(\hat{\mathbf{x}})$ is introduced in Proposition 1.

Proposition 1.

(a) The optimal expected discounted cost $V_\gamma(\hat{\mathbf{x}})$ satisfies the Bellman equation as follows:

$$V_\gamma(\hat{\mathbf{x}}) = \min_{a\in\mathcal{A}} Q_\gamma(\hat{\mathbf{x}}, a), \tag{13}$$

where the state–action value function $Q_\gamma(\hat{\mathbf{x}}, a)$ is defined as

$$Q_\gamma(\hat{\mathbf{x}}, a) = C(\hat{\mathbf{x}}, a) + \gamma \sum_{\mathbf{x}'\in\mathcal{S}} \Pr(\mathbf{x}'|\hat{\mathbf{x}}, a) V_\gamma(\mathbf{x}'). \tag{14}$$

(b) The policy π determined by the right hand side of (13) is γ-optimal, and $\pi \in \Pi_{SD}$.

(c) $V_\gamma(\hat{\mathbf{x}})$ can be solved by value iteration algorithm. Specifically, let $V_{\gamma,n}(\hat{\mathbf{x}})$ be the cost-to-go function and $V_{\gamma,0}(\hat{\mathbf{x}}) = 0$ for all state $\hat{\mathbf{x}} \in \mathcal{S}$. For all $n \geq 1$, we have:

$$V_{\gamma,n}(\hat{\mathbf{x}}) = \min_{a\in\mathcal{A}} Q_{\gamma,n}(\hat{\mathbf{x}}, a), \tag{15}$$

where $Q_{\gamma,n}(\hat{\mathbf{x}}, a)$ is obtained as follows:

$$Q_{\gamma,n}(\hat{\mathbf{x}}, a) = C(\hat{\mathbf{x}}, a) + \gamma \sum_{\mathbf{x}'\in\mathcal{S}} \Pr(\mathbf{x}'|\hat{\mathbf{x}}, a) V_{\gamma,n-1}(\mathbf{x}'). \tag{16}$$

Then the equation $\lim_{n\to\infty} V_{\gamma,n}(\hat{\mathbf{x}}) = V_\gamma(\hat{\mathbf{x}})$ holds for every state $\hat{\mathbf{x}}$ and γ.

Proof. See Appendix A. □

Now, we can show the monotonic properties of $V_\gamma(\hat{\mathbf{x}})$ in the following lemma by using (c) in Proposition 1.

Lemma 1. *Given fixed channel erasure probability p and EH probability λ, then*

*(a) value function $V_\gamma(\Delta, q)$ is **non-decreasing** in Δ, i.e., for any $1 \leq \Delta_1 \leq \Delta_2$ and any battery state $q \in \mathcal{B}$, we have*

$$V_\gamma(\Delta_1, q) \leq V_\gamma(\Delta_2, q), \tag{17}$$

and

$$V_\gamma(\Delta_2, q) - V_\gamma(\Delta_1, q) \geq \Delta_2 - \Delta_1. \tag{18}$$

(b) *value function* $V_\gamma(\Delta, q)$ *is* ***non-increasing*** *in* q*, i.e., for AoI* $\Delta \geq 1$ *and any battery state* $q \in \{0, 1, \ldots, B-1\}$*, we have*

$$V_\gamma(\Delta, q) \geq V_\gamma(\Delta, q+1), \tag{19}$$

Proof. See Appendix B. □

Based on the Proposition 1 and Lemma 1, we will verify the existence of the optimal stationary deterministic policy for the average cost problem (10) in the following theorem.

Theorem 1. *There exists an optimal policy* $\pi^\star \in \Pi_{SD}$ *for the average cost MDP in (10). Moreover, for every state* $\boldsymbol{x}$*, there exists a value function* $V(\cdot) : \mathcal{S} \to \mathbb{R}$ *and a unique constant* $g^\star \in \mathbb{R}$ *such that:*

$$g^\star + V(\boldsymbol{x}) = \min_{a \in \mathcal{A}} \left\{ C(\boldsymbol{x}, a) + \sum_{\boldsymbol{x}' \in \mathcal{S}} \Pr(\boldsymbol{x}' | \boldsymbol{x}, a) V(\boldsymbol{x}') \right\}, \tag{20}$$

where $g^\star$ *is the optimal average cost of problem (10) and satisfies* $g^\star = \lim_{\gamma \to 1} (1-\gamma) V_\gamma(\boldsymbol{x})$ *for every state* $\boldsymbol{x}$*, and the value function* $V(\boldsymbol{x})$ *satisfies*

$$V(\boldsymbol{x}) = \lim_{\gamma \to 1} \gamma V_\gamma(\boldsymbol{x}) = \lim_{\gamma \to 1} V_\gamma(\boldsymbol{x}) - g^\star = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_\pi \left\{ \sum_{t=0}^{T-1} [C(\boldsymbol{x}[t], a[t]) - g^\star] \right\}. \tag{21}$$

Proof. See Appendix C. □

Based on Theorem 1, we have the following corollary:

Corollary 1. *The state–action value function* $Q(\boldsymbol{x}, a)$ *for the average cost is given as follows:*

$$Q(\boldsymbol{x}, a) = C(\boldsymbol{x}, a) + \sum_{\boldsymbol{x}' \in \mathcal{S}} \Pr(\boldsymbol{x}' | \boldsymbol{x}, a) V(\boldsymbol{x}'), \tag{22}$$

which is similar to $Q_\gamma(\boldsymbol{x}, a)$ *in (14) by letting* $\gamma \to 1$*. Then the optimal policy* $\pi^\star \in \Pi_{SD}$ *for the average cost MDP in (10) can be expressed as follows:*

$$\pi^\star(\boldsymbol{x}) = \arg\min_{a \in \mathcal{A}} Q(\boldsymbol{x}, a), \forall \boldsymbol{x} \in \mathcal{S}. \tag{23}$$

3.3. Structure Analysis of Optimal Policy

Before analyzing the structure of the optimal policy $\pi^\star$, we first prove some monotonic properties of the value function $V(\mathbf{x})$ on different dimensions, which is summarized in the following lemma.

Lemma 2. *Given fixed channel erasure probability* p *and EH probability* λ*, then*

(a) *value function* $V(\Delta, q)$ *is* ***non-decreasing*** *in* Δ*, i.e., for any* $1 \leq \Delta_1 \leq \Delta_2$ *and any battery state* $q \in \mathcal{B}$*, we have*

$$V(\Delta_1, q) \leq V(\Delta_2, q), \tag{24}$$

and

$$V(\Delta_2, q) - V(\Delta_1, q) \geq \Delta_2 - \Delta_1. \tag{25}$$

(b) *value function* $V(\Delta, q)$ *is* ***non-increasing*** *in* q*, i.e., for AoI* $\Delta \geq 1$ *and any battery state* $q \in \{0, 1, \ldots, B-1\}$*, we have*

$$V(\Delta, q) \geq V(\Delta, q+1), \tag{26}$$

Proof. According to the (21), $V(x) = \lim_{\gamma \to 1} V_\gamma(\mathbf{x}) - g$. Therefore, the monotonic properties of $V_\gamma(\mathbf{x})$ in Lemma 1 are also valid for $V(x)$, which completes the proof. □

Based on Lemma 2, we will derive the **proportional differential property** of the value function in Lemma 3.

Lemma 3. *Given fixed channel erasure probability p and EH probability λ, then value function $V(\Delta, q)$ has the **proportional differential property**, i.e., the inequality*

$$\frac{V(\Delta+1, q+1) - V(\Delta, q+1)}{V(\Delta+1, q) - V(\Delta, q)} \geq p \tag{27}$$

holds for AoI $\Delta \geq 1$ and any battery state $q \in \{0, 1, ..., B-1\}$.

Proof. See Appendix D. □

With Corollary 1, Lemmas 2 and 3, we directly provide our main result in the following theorem.

Theorem 2. *Assuming that the channel erasure probability p and EH probability λ are both fixed , there exists a threshold $\Delta_q \in \mathbb{Z}^+$ for given battery state q, such that when $\Delta < \Delta_q$, the optimal action $\pi^\star(\Delta, q) = 0$, i.e., the sensor keeps idle; when $\Delta \geq \Delta_q$, the optimal action $\pi^\star(\Delta, q) = 1$, i.e., the sensor chooses to generate and transmit a new update.*

Proof. See Appendix E. □

Theorem 2 reveals the threshold structure of the optimal policy: if the optimal action in a certain state is to generate and transmit an update, then in the state with the same battery state and larger AoI, the optimal action must be the same. Note that the threshold Δ_q is actually determined by the channel erasure probability p, EH probability λ and pre-defined weighting factor ω. The closed-form expression of the threshold is difficult to be derived due to the complex transition probabilities. In the next section, we will show how to compute the optimal policy numerically.

3.4. Modified Relative Value Iteration Algorithm Design

In this section, we will propose a computationally efficient algorithm to find the optimal stationary deterministic policy based on the threshold structure.

Since the state space $\mathcal{S}$ is infinite, we will use a truncated space $\mathcal{S}^N$ for approximation in practice, where $\mathcal{S}^N = \{(\Delta, q) | \Delta \leq N, \Delta \in \mathbb{Z}^+, q \in \mathcal{B}\}$. It can be proved that when N is large enough, the optimal policy of the approximated MDP will be identical to that of the original problem [6].

However, the value iteration algorithm in Proposition 1 for the discounted cost problem cannot be applied to the average cost problem by letting $\gamma = 1$. It does not converge because the value function $V(\cdot)$ in (20) is not unique. One can check if $V(\cdot)$ satisfies (20), a new function $V'(\cdot) = V(\cdot) + c$ also satisfies (20), where $c \in \mathbb{R}$. Therefore, we introduce a *relative value iterative* (RVI) algorithm to obtain the optimal policy of the approximate average cost MDP [45]. We choose a reference state $\hat{x} \in \mathcal{S}^N$ and set $V_0(\mathbf{x}) = 0$ for all states $\mathbf{x} \in \mathcal{S}^N$ Then for all $n \geq 0$, we have

$$V_{n+1}(\mathbf{x}) = \min_{a \in \mathcal{A}} Q_{n+1}(\mathbf{x}, a), \tag{28}$$

and $Q_{n+1}(\mathbf{x}, a)$ is obtained as follows:

$$Q_{n+1}(\mathbf{x}, a) = C(\mathbf{x}, a) + \sum_{\mathbf{x}' \in \mathcal{S}^N} \Pr(\mathbf{x}' | \mathbf{x}, a) h_n(\mathbf{x}'), \tag{29}$$

where the differential value function is $h_n(\mathbf{x}) = V_n(\mathbf{x}) - V_n(\hat{\mathbf{x}})$. The equation $\lim_{n\to\infty} Q_n(\mathbf{x}, a) = Q(\mathbf{x}, a)$ holds for every state $\mathbf{x} \in \mathcal{S}^N$ and action $a \in \mathcal{A}$. Finally, we compute the optimal policy by

$$\pi^\star(\mathbf{x}) = \arg\min_{a\in\mathcal{A}} Q(\mathbf{x}, a). \tag{30}$$

Note that the optimal policy is still of a threshold structure. The corresponding proof is similar to that of Theorem 2.

Moreover, based on the RVI algorithm, we can exploit this threshold structure to reduce the computational complexity. When the optimal policy of a state $\mathbf{x}' = (\Delta', q')$ is 1, the optimal policy of state $\mathbf{x}' \in \{(\Delta, q) | \Delta > \Delta', \Delta \leq N, q = q'\}$ will also be 1 without the need to calculate (30). Therefore, we propose a modified RVI algorithm, and the details are given in Algorithm 1.

Algorithm 1 Modified relative value iteration algorithm.

Input:
Iteration number K,
Iteration threshold ϵ,
Maximum of AoI N,
Maximum of battery state B,
Reference state $\hat{\mathbf{x}}$.

Output:
Optimal policy $\pi^\star(\mathbf{x})$ for all state $\mathbf{x}$.

Initialization: $h_0(\mathbf{x}) = 0$, for all $\mathbf{x} \in \mathcal{S}^N$
for episodes $n = 0, 1, 2, \ldots, K$ **do**
  for state $\mathbf{x} \in \mathcal{S}^N$ **do**
    for action $a \in \mathcal{A}$ **do**
      $Q_n(\mathbf{x}, a) \leftarrow C(\mathbf{x}, a) + \sum_{\mathbf{x}'\in\mathcal{S}^N} \Pr(\mathbf{x}'|\mathbf{x}, a) h_n(\mathbf{x}')$ // Update the state-action value function.
    end for
    $V_{n+1}(\mathbf{x}) \leftarrow \min_{a\in\mathcal{A}} Q_n(\mathbf{x}, a)$ // Update the value function.
    $h_{n+1}(\mathbf{x}) \leftarrow V_{n+1}(\mathbf{x}) - V_{n+1}(\hat{\mathbf{x}})$ // Update the differential value function.
  end for
  if $\|h_{n+1}(\mathbf{x}) - h_n(\mathbf{x})\| \leq \epsilon, \forall \mathbf{x} \in \mathcal{S}^N$ **then**
    for $\mathbf{x} = (\Delta, q) \in \mathcal{S}^N$ **do**
      if $\pi^\star(\Delta - 1, q) = 1$ **then**
        $\pi^\star(\mathbf{x}) \leftarrow 1$, // Leverage the threshold structure of the optimal policy.
      else
        $\pi^\star(\mathbf{x}) \leftarrow \arg\min_{a\in\mathcal{A}} Q_n(\mathbf{x}, a)$
      end if
    end for
    break
  end if
end for

4. Minimize Average Cost in an Unknown Environment

In the previous sections, we assumed that the channel erasure probability p and EH probability λ are known in advance. Thus, the *model-based* RVI method can be employed to obtain the optimal updating policy. However, statistics such as p and λ may be unknown and even time variant in many practical scenarios, which makes it impossible to apply modified RVI algorithm because the transition probabilities are not explicit and Equation (29) cannot be applied to estimate the state-action value function $Q(\mathbf{x}, a)$. In the field of reinforcement learning, alternatively, *model-free* methods can solve MDP problems with unknown

transition probabilities. An example of a model-free algorithm is *Q-learning* [46]. Q-learning finds an optimal policy in the sense of maximizing the expected value of the total reward over any and all successive steps. However, it is only designed for discounted MDP. For the average cost problem in (10), we employ an average cost Q-learning algorithm. The basic idea of this algorithm comes from the *SMART* algorithm in [47], which is a model-free reinforcement learning algorithm proposed for semi-Markov decision problems (SMDP) under the average-reward criterion. We modify it to fit the average cost MDP problem.

The state–action value function $Q(\mathbf{x}, a)$ is essential for solving the optimal policy. When the model is unknown, as long as $Q(\mathbf{x}, a)$ can be estimated accurately, the optimal policy can also be obtained immediately by (30). So the key question is how to estimate the $Q(\mathbf{x}, a)$ function, or equivalently, the value of all state–action pairs. Similar to Q-learning, the average cost Q-learning algorithm uses the minimum value of the next state–action pairs to update the value of the current state–action pair. Moreover, it needs to estimate the shift value g by averaging all immediate cost.

Specifically, the average cost Q-learning algorithm learns $Q(\mathbf{x}, a)$ by episodes. Each episode contains several iterations, and each iteration corresponds to one time slot. Then in the nth time slot of an episode, the algorithm first observes the current state $\mathbf{x}[n] = (\Delta[n], q[n])$, selects an action $a[n]$ according to the ϵ-greedy policy:

$$a[n] = \begin{cases} \arg\min\limits_{a \in \mathcal{A}} Q(\mathbf{x}[n], a), & \text{with probability } 1 - \epsilon, \\ \text{random action}, & \text{otherwise.} \end{cases} \tag{31}$$

By (9), the immediate cost $C[n] = \Delta[n] + \omega C_r a[n](1 - u(q[n]))$ occurs, and the system will transit to the next state $\mathbf{x}[n+1]$. The value of $Q(\mathbf{x}[n], a[n])$ is updated as follows:

$$Q(\mathbf{x}[n], a[n]) = (1 - \alpha[n]) Q(\mathbf{x}[n], a[n]) + \alpha[n](C[n] - g + \min_{a \in \mathcal{A}} Q(\mathbf{x}[n+1], a)), \tag{32}$$

where $\alpha[n]$ is the learning rate. The shift value g is updated as follows:

$$g = (1 - \beta[n]) g + \beta[n] C[n] \tag{33}$$

where $\beta[n] = \frac{1}{n}$. The details are given in Algorithm 2. We leverage the parameter ϵ to balance exploration and exploitation. As the number of epochs increases, the learned $Q(\mathbf{x}, a)$ value will approach its true value, so we can gradually decrease ϵ to 0 to reduce invalid exploration. At the same time, the shift value g will also be close to the optimal average cost $g^\star$ in (20). Note that in [47], the shift value g is updated only in a non-exploratory time slot. Here we update it by simply averaging all cost, similar to [48]. The performance comparison of the average cost Q-learing algorithm and modified RVI algorithm is shown in the next section.

Algorithm 2 Average cost Q-learning algorithm.

Input:
Maximum number of episodes K,
Maximum iteration number of an episode N_e,
Maximum of AoI N,
Maximum of battery state B,
Initial value of $Q^{N\times B\times 2} \leftarrow \mathbf{0}$,
Initial value of $\epsilon \leftarrow 0$,
Initial value of the shift value $g \leftarrow 0$.
Output:
Learned policy $\pi(\mathbf{x})$ for all state $\mathbf{x}$,
Average cost $g^\star$ by following the policy π.

for episodes $k = 0, 1, 2, \ldots, K$ **do**
  $g \leftarrow 0$ // Initialize the shift value at the beginning of every episode.
  for $n = 1, 2, \ldots, N_e$ **do**
    Observe the current state $\mathbf{x}[n]$
    Select an action $a[n]$ according to ϵ-greedy policy in (31)
    Calculate immediate cost $C[n] \leftarrow \Delta[n] + \omega C_r a[n](1 - u(q[n]))$
    Observe the next state $\mathbf{x}[n+1]$
    $\alpha[n] \leftarrow \frac{1}{\sqrt{n}}$
    $Q(\mathbf{x}[n], a[n]) \leftarrow (1-\alpha[n])Q(\mathbf{x}[n], a[n]) + \alpha[n](C[n] - g + \min\limits_{a\in\mathcal{A}} Q(\mathbf{x}[n+1], a))$ // Update the state-action value function.
    $\beta[n] \leftarrow \frac{1}{n}$
    $g \leftarrow (1-\beta[n])g + \beta[n]C[n]$// Update the shift value.
  end for
  Decrease ϵ
end for
for $\mathbf{x} = (\Delta, q) \in \mathcal{S}^N$ **do**
  $\pi(\mathbf{x}) \leftarrow \arg\min\limits_{a\in\mathcal{A}} Q(\mathbf{x}, a)$ // Calculate the learned policy π.
end for

5. Numerical Results

In this section, we first show the threshold structure of the optimal policy by the simulation results. Then we compare the performance of the optimal policy with the following representative policies under different system parameters:

- Zero-wait policy [4]. The sensor generates and transmits an update in every time slot.
- Periodic policy. The sensor periodically generates and sends updates to the destination.
- Randomized policy. The sensor chooses to send an update or stay idle in each time slot with the same probability.
- Energy first policy. The sensor only uses the harvested energy, that is, as long as the battery state is not zero, it will choose to sense and send updates, otherwise it will remain idle.

Moreover, we will show the average cost Q-learning algorithm performs very close to the modified RVI with known statistics. We will also compare age and reliable energy cost trade-off curves of the optimal updating policies under EH supply, reliable energy supply and mixed energy supplies. Finally, we compare the performance of the optimal policy under the only EH supply and unit-sized battery setting with the prior results in [23,25].

5.1. Simulation Setup

In our simulations, we set the maximum of AoI $N = 500$, and the maximum of battery state $B = 20$. So the finite state space $\mathcal{S}^N = \{(\Delta, q) | \Delta \leq 500, \Delta \in \mathbb{Z}^+, q \in \mathcal{B}\}$. The cost of reliable energy C_r for one update is equal to 2. For the modified RVI algorithm, we set the iteration number $K = 1000$, iteration threshold $\epsilon = 10^{-5}$ and reference state $\hat{\mathbf{x}} = (1, B)$.

The optimal policy and other baseline policies are run for $T = 10{,}000$ time slots to compute the average cost. For the average cost Q-learning algorithm, we set the total number of episodes $K = 1000$, and the maximum iteration number in an episode $N_e = 100{,}000$.

5.2. Results

Figure 3 shows the optimal policy under different system parameters. All the subfigures in Figure 3 exhibit the threshold structure described in Theorem 2. Intuitively, when ω is very small, the optimal action for every state should be 1, and when ω is very large, the optimal action for battery state $q = 0$ should be 0. It can be observed from Figure 3a,b that when ω is small (i.e., $\omega = 0.1$), the optimal policy is to update for every state, which is exactly the zero-wait policy. Figure 3 also shows that when ω is relatively large (e.g., $\omega = 10$), and the AoI is small, even if the battery state is not zero, the optimal action in the corresponding state is to keep idle. When the AoI is large or the battery state is large, the optimal action is to measure and send updates. Moreover, in all the subfigures, the threshold Δ_q keeps monotonically non-increasing with the battery state q. However, this conclusion has not been rigorously proven.

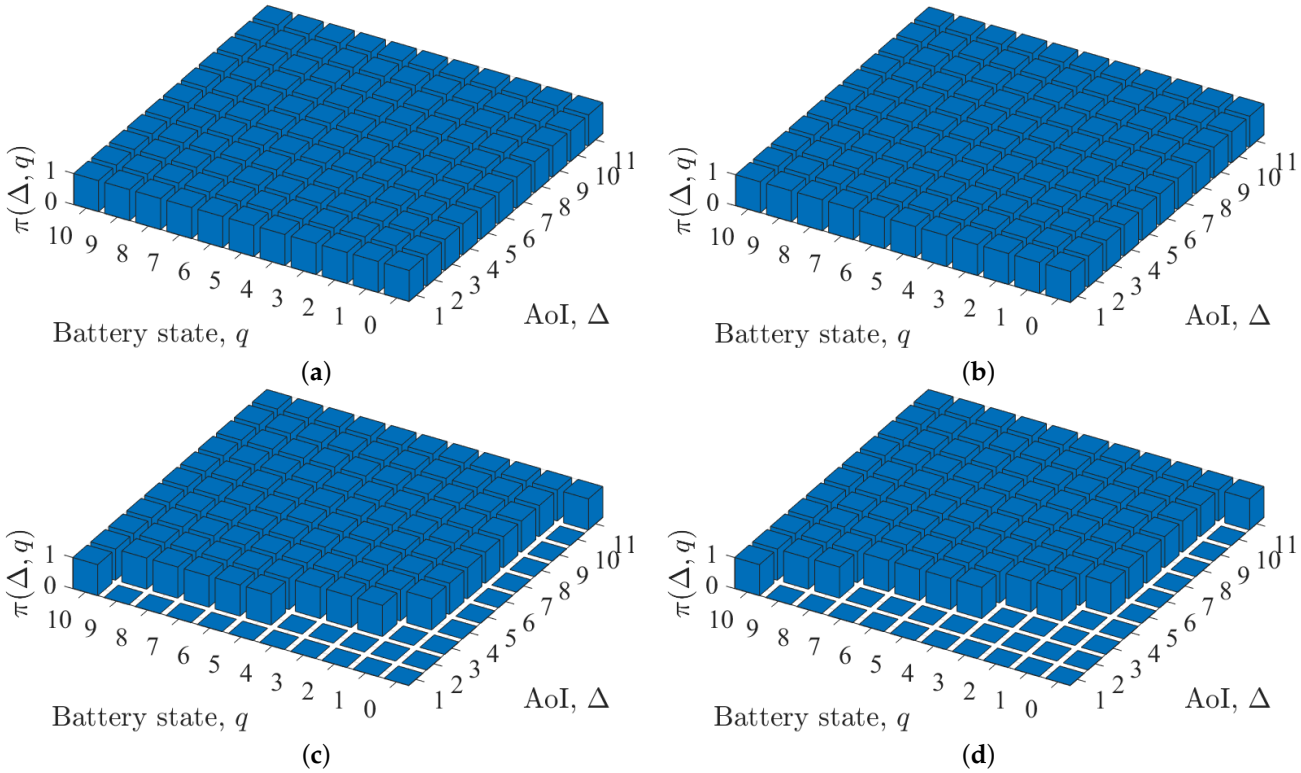

Figure 3. Optimal policy conditioned on different parameters: (**a**) $\omega = 0.1$, $p = 0.2$, $\lambda = 0.5$, (**b**) $\omega = 0.1$, $p = 0.4$, $\lambda = 0.5$, (**c**) $\omega = 10$, $p = 0.2$, $\lambda = 0.5$ and (**d**) $\omega = 10$, $p = 0.4$, $\lambda = 0.5$.

Figure 4 shows the time average cost with respect to ω under different policies. Here, we set the period of the periodic policy to 5 and 10 for comparison without loss of generality. It can be found that under different weighting factor ω, the optimal policy proposed in this paper can obtain the minimum long-term average cost compared with the other policies, which indicates the best trade-off between the average AoI and the cost of reliable energy. When ω tends to 0, the zero-wait policy tends to be optimal. Since there is no need to consider the update cost brought by paid reliable backup energy, the optimal policy should maximize the utilization of the updating opportunities.

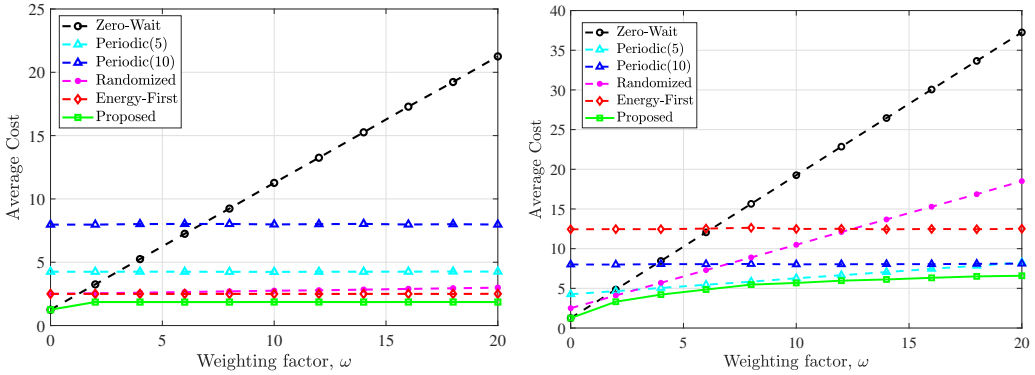

Figure 4. Performance comparison of the proposed optimal policy, zero-wait policy, periodic policy (period = 5), periodic policy (period = 10), randomized policy and energy first policy versus the weighting factor ω with simulation conditions: (**a**) $p = 0.2$, $\lambda = 0.5$ and (**b**) $p = 0.2$, $\lambda = 0.1$.

It can also be observed from Figure 4 that the growth of the optimal policy curve slows down as ω increases. This is because the optimal policy in the case of large ω does not tend to use the reliable energy when battery state $q = 0$, but prefers to wait for harvested energy, as shown in Figure 3. Since the EH probability is constant, the average AoI does not change much, resulting in no significant increase in the total average cost. Comparing Figure 4a,b, it is found that the larger the λ, the smaller the average cost variation with ω. This is because there is not much opportunity for the sensor to use reliable energy in the case of sufficient harvested energy.

Figure 5 reveals the impact of EH probabilities λ. In Figure 5a,b, we set $p = 0.2$, $\omega = 10$ and $p = 0.2, \omega = 1$, respectively.

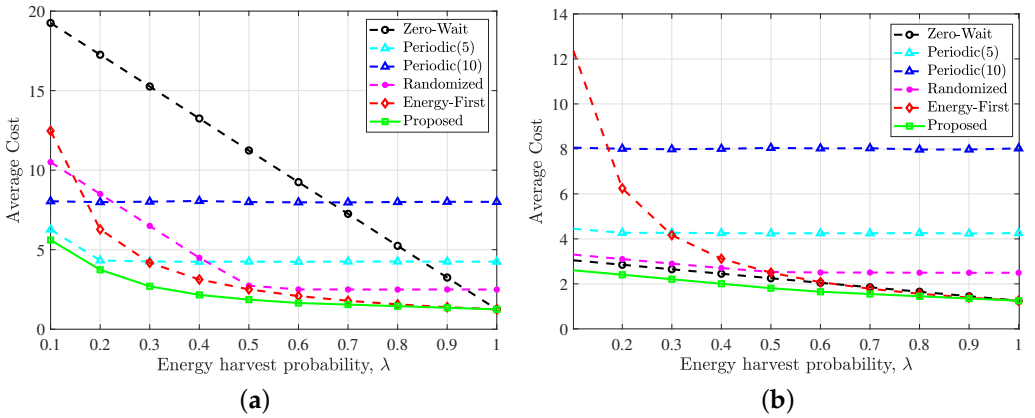

Figure 5. Performance comparison of the proposed optimal policy, zero-wait policy, periodic policy (period = 5), periodic policy (period = 10), randomized policy and energy first policy versus the EH probability λ with simulation conditions: (**a**) $p = 0.2$, $\omega = 10$ and (**b**) $p = 0.2$, $\omega = 1$.

It can also be found from both Figure 5a,b that the proposed optimal update policy outperforms all other policies under different EH probability. The interesting point is that when the EH probability tends to 1, i.e., energy arrives in each time slot, the performance of the zero-wait policy and the energy first policy is equal to the optimal policy, while there is still a performance gap between the optimal policy and the other two polices. This is intuitive because when the free harvested energy is sufficient, the optimal policy must be to generate and transmit updates in every time slot. However, the periodic policy and the randomized policy still keep idle in many time slots, which will lead to a higher average AoI and thus increase the average cost. Results show that the performance of zero-wait

policy approaches the optimal policy for large λ, which is consistent with our findings in Figure 4.

In Figure 6, we compare the five policies under different channel erasure probability p.

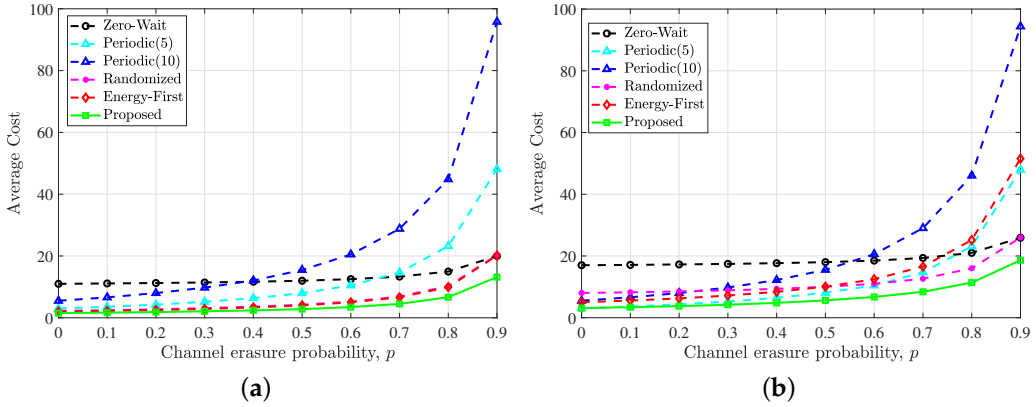

Figure 6. Performance comparison of the proposed optimal policy, zero-wait policy, periodic policy (period = 5), periodic policy (period = 10), randomized policy and energy first policy versus the erasure probability p with simulation conditions: (**a**) $\lambda = 0.5$, $\omega = 10$ and (**b**) $\lambda = 0.2$, $\omega = 10$.

It can be found that when the erasure probability increases from 0 to 0.9, the proposed optimal update policy always performs better than the other baseline policies. As p tends to 1, the average cost under all policies theoretically tends to infinity because all updates will be erased by the noisy channel and cannot be received by the destination. The simulation results confirmed this conjecture. Comparing Figure 6a,b, we can observe that when λ is large, the energy-first strategy will be close to the optimal strategy, which is also illustrated in Figure 5.

Figure 7 shows the performance of the average cost Q-learning algorithm. In every episode, the shift value g of the last inner iteration is recorded as the average cost. It can be found from Figure 7a that the average cost achieved by Algorithm 2 converges to that obtained by the modified RVI algorithm under different EH probability λ and channel erasure probability p. The age–energy trade-off is shown in Figure 7b. By fixing λ and p and changing ω from 0 to 1000, we run the modified RVI algorithm and average cost Q-learning algorithm to obtain the corresponding trade-off curve. It can be found that the curve obtained by the average cost Q-learning algorithm is very close to the optimal trade-off curve under the same condition, which further verifies the near-optimal performance of the average cost Q-learning algorithm in an unknown environment.

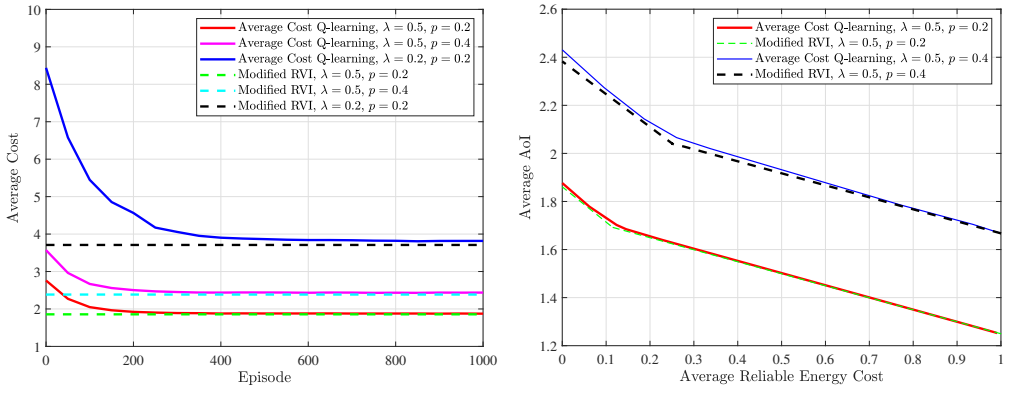

Figure 7. (**a**) Performance of the average cost Q-learning with respect to the modified RVI algorithm under different system parameters ($\omega = 10$); (**b**) age–energy trade-off curves computed by the average cost Q-learning and modified RVI algorithm.

Figure 8 shows the optimal age and reliable energy cost trade-off curves for different energy supplies. By fixing EH probability λ and channel erasure probability p and changing ω from 0 to 10,000, we run the modified RVI algorithm to get the optimal trade-off curve for mixed energy supplies. By letting EH probability $\lambda = 0$ and following the same steps, we can obtain the optimal trade-off curve for reliable energy supply. By letting weighting factor ω go to infinity, we can theoretically obtain the optimal trade-off "curve" corresponding to the EH supply. The "curve" contains only one point because the reliable energy consumption can only be 0 for the EH supply case. It should be noted that ω cannot be infinite in a simulation. Instead, we can set ω to a relatively large number (e.g., 10,000). To facilitate comparison, the channel erasure probability is set as $p = 0.2$, and the EH probability λ is set as 0.1, 0.3 and 0.7. It can be observed that the curves for the mixed energy supplies are always at the lower left of the curve for relying solely on reliable energy, which indicates that under the same average AoI, the reliable energy required by the system under the mixed energy supplies is smaller, and under the same reliable energy consumption, the AoI of the system under the mixed energy supplies is lower. The mixed energy design also achieves lower AoI than that with only EH, at the cost of paying for reliable energy. The optimal updating policy proposed in this paper makes full use of the harvested energy.

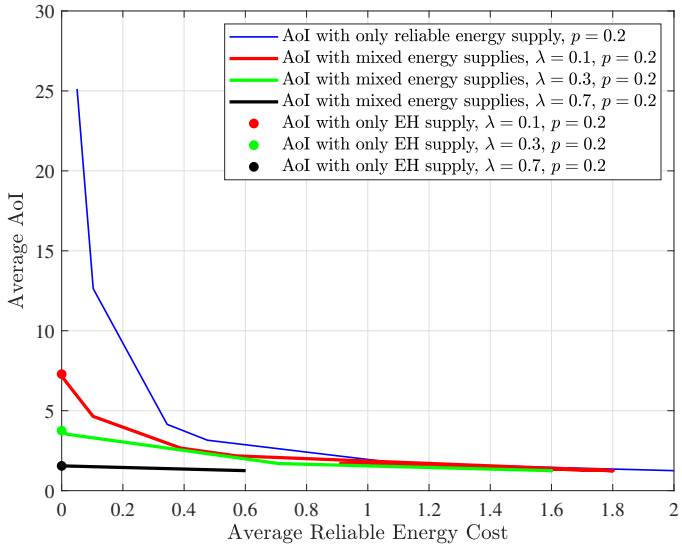

Figure 8. Age-reliable energy trade-off for different energy supplies: mixed energy supply, reliable energy supply and EH supply. The channel erasure probability $p = 0.2$, and the EH probability λ is set as 0.1, 0.3 and 0.7, respectively.

Figure 9 compares the performance of the optimal policy with the prior results in [23,25] for a special case where the sensor only uses the harvested energy and the battery capacity $B = 1$. Both [23,25] considered a continuous-time model, i.e., the energy arrival process is a Poisson process with an arrival rate of Λ energy units per *time unit* (TU), and proved that the optimal policies are threshold structure, in which a new update is generated and transmitted only if the time until the next energy arrival since the latest successful transmission exceeds a certain threshold. Specifically, [23] (Theorem 4, Equation (13)) provided the average AoI and threshold in closed-form under the optimal update policy for any energy arrival rate Λ in the error-free channel case. It is interesting that the optimal average AoI and the corresponding threshold are equal. Ref. [25] (Theorem 4, Equation (14)) extended the results of [23] to an error-prone channel case, while the energy arrival rate Λ is assumed to be 1. So we first show the results of [23,25] vs. different channel erasure probability p in Figure 9, where the energy arrival rate $\Lambda = 1$. It should be emphasized that the unit of the average AoI and threshold is TU. According to Theorems 1 and 2 in this paper, the optimal update policy exists and admits a threshold

structure for any EH probability λ, channel erasure probability p, weighting factor ω and battery capacity B. This conclusion is based on the discrete-time model, i.e., the energy arrives as a Bernoulli process with parameter λ, which is different from the continuous-time model in [23,25], and the reliable backup energy is also considered. However, by the choice of some parameters (large ω, small λ), our results can be a good approximation of the results in [23,25]. First, by choosing a large ω, the reliable energy will almost never be used, and equivalently, only the EH supply exists. Secondly, by choosing a small λ, the Poisson process can be approximated as a Bernoulli process. This is because for a Poisson process with parameter Λ, we can discretize a TU into n small time slots of equal length, then according to probability theory, when n is large enough, the energy arrival process within a time slot can be approximated as a Bernoulli process with parameter $\lambda = \Lambda/n$, which is relatively small. In our simulation, we set the battery capacity $B = 1$, and take $\lambda = 0.1$ (i.e., $n = 10$) and $\omega = 10{,}000$. By changing the channel erasure probability p, we can run the modified RVI algorithm to compute the minimum average AoI and the optimal threshold. It needs to be mentioned that the unit of them is a time slot. For comparison, we need to divide their values by n to obtain the average AoI and threshold in TU. The final results are shown by the dashed lines in Figure 9. It can be observed that the results of this paper are extremely close to the explicit results in [23,25], which verifies the correctness of the analysis and also reflects the generality of our system model.

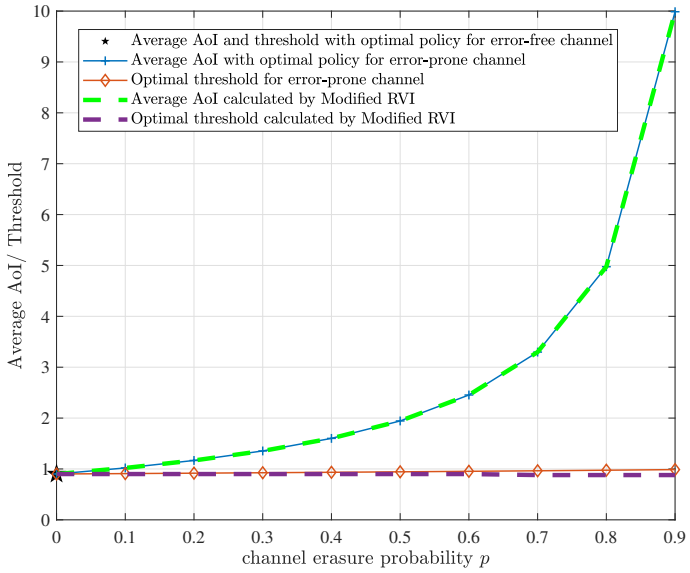

Figure 9. AoI and threshold with the proposed optimal policy for a special case where the sensor only uses the harvested energy and the battery capacity $B = 1$, and those with a unit-sized battery in [23,25] (error-free channel case and error-prone channel case, respectively), vs. the channel erasure probability p.

6. Conclusions

In this paper, we studied the optimal updating policy for an information update system, where a wireless sensor sends updates over an erasure channel using both harvested energy and reliable backup energy. Theoretical analysis indicates the threshold structure of the optimal policy and simulation results verify its performance. For the practical case where the statistics, such as the EH probability and channel erasure probability, are unknown in advance, a learning-based algorithm is proposed to compute the updating policy. Simulation results show its performance is close to that of the optimal policy. With the optimal policy, the design of mixed energy supplies can make full use of harvested energy and achieve the best age–energy trade-off. In future work, we will focus on the timeliness of the multi-sensor system under mixed energy supplies.

Author Contributions: Conceptualization, L.W., F.P., X.C. and S.Z.; methodology, L.W. and F.P.; software, L.W. and F.P.; validation, L.W. and F.P.; formal analysis, L.W. and F.P.; investigation, L.W.; writing—original draft preparation, L.W.; writing—review and editing, F.P., X.C. and S.Z.; visualization, L.W.; supervision, S.Z.; project administration, S.Z.; funding acquisition, S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Key Research and Development Program of China under Grant 2019YFE0113200&2019YFE0196600, Tsinghua University-China Mobile Communications Group Co.,Ltd. Joint Institute, Huawei Company Cooperation Project under Contract No. TC20210519013.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AoI	Age of Information
EH	Energy Harvesting
MDP	Markov Decision Process
RVI	Relative Value Iteration
SMART	Semi-Markov Average Reward Technique
SMDP	Semi-Markov Decision problems

Appendix A. Proof of Proposition 1

According to [15], the proof of Proposition 1 is equivalent to proving that there is a stationary deterministic policy π such that the expected discounted cost $V_\gamma^\pi(\mathbf{x})$ is finite for all $\mathbf{x}$, γ. So we can select a policy π which chooses to keep idle in each time slot. Then by (11), for any state $\mathbf{x} = (\Delta, q) \in \mathcal{S}$ and $\gamma \in (0,1)$, we have

$$\begin{aligned} V_\gamma^\pi(\mathbf{x}) &= \mathbb{E}_\pi\left\{\sum_{t=0}^{\infty}\gamma^t C(\mathbf{x}[t], a[t])|\mathbf{x}[0]=\mathbf{x}\right\} \\ &= \sum_{t=0}^{\infty}\gamma^t C(\mathbf{x}[t], a[t]) \\ &= \sum_{t=0}^{\infty}\gamma^t(\Delta+t) \\ &= \frac{1}{1-\gamma}(\Delta+\frac{\gamma}{1-\gamma}) < \infty, \end{aligned} \tag{A1}$$

which completes the proof.

Appendix B. Proof of Lemma 1

The proof requires the use of value iteration algorithm(VIA) and mathematical induction. According to (c) in Proposition 1, The specific iteration process of VIA is as follows:

$$\begin{cases} V_{\gamma,0}(\mathbf{x}) = 0, \\ Q_{\gamma,k}(\mathbf{x},a) = C(\mathbf{x},a) + \gamma\sum\limits_{\mathbf{x}'\in\mathcal{S}}\Pr(\mathbf{x}'|\mathbf{x},a)V_{\gamma,k}(\mathbf{x}'), \\ V_{\gamma,k+1}(\mathbf{x}) = \min\limits_{a\in\mathcal{A}} Q_{\gamma,k}(\mathbf{x},a), \end{cases} \tag{A2}$$

where $k \in \mathbb{Z}^+$. For any state $\mathbf{x} \in \mathcal{S}$, $V_{\gamma,k}(\mathbf{x})$ will converge when k goes into infinity:

$$\lim_{k\to\infty} V_{\gamma,k}(\mathbf{x}) = V_\gamma(\mathbf{x}). \tag{A3}$$

Then we will use mathematical induction to prove the monotonicity of the value function in each component.

First let us tackle part (a) of Lemma 1.

For (17), we can verify that the inequality $V_{\gamma,1}(\Delta_1, q) \leq V_{\gamma,1}(\Delta_2, q)$ holds when $k = 1$. Then we assume that at the kth step of the induction method, the following formula holds:

$$V_{\gamma,k}(\Delta_1, q) \leq V_{\gamma,k}(\Delta_2, q), \forall \Delta_1 \leq \Delta_2. \tag{A4}$$

So the next formula that needs to be verified is

$$V_{\gamma,k+1}(\Delta_1, q) \leq V_{\gamma,k+1}(\Delta_2, q), \forall \Delta_1 \leq \Delta_2 \tag{A5}$$

Since $V_{\gamma,k+1}(\mathbf{x}) = \min_{a \in \mathcal{A}} Q_{\gamma,k}(\mathbf{x}, a)$, we need to bring out $Q_{\gamma,k}(\mathbf{x}, a)$ first. Due to the complexity of the transition probabilities and one-step cost function, we need to discuss the following three cases: $q = 0$, $0 < q < B$ and $q = B$. For the sake of brevity, we only give the calculation details of the case $0 < q < B$, and the other two cases can be verified by following the exact same steps.

According to transition probability (7) and (8), we have the state-action value function $Q_{\gamma,k}(\Delta, q, 0)$ and $Q_{\gamma,k}(\Delta, q, 1)$ as follows:

$$Q_{\gamma,k}(\Delta, q, 0) = \Delta + \gamma\lambda V_{\gamma,k}(\Delta + 1, q + 1) + \gamma(1 - \lambda)V_{\gamma,k}(\Delta + 1, q), \tag{A6}$$

and

$$\begin{aligned} Q_{\gamma,k}(\Delta, q, 1) = \Delta &+ \gamma p\lambda V_{\gamma,k}(\Delta + 1, q) + \gamma p(1 - \lambda)V_{\gamma,k}(\Delta + 1, q - 1) \\ &+ \gamma(1 - p)\lambda V_{\gamma,k}(1, q) + \gamma(1 - p)(1 - \lambda)V_{\gamma,k}(1, q - 1). \end{aligned} \tag{A7}$$

Because $V_{\gamma,k}(\Delta, q)$ is assumed to be non-decreasing function with respect to Δ for any fixed q, it is obvious that both $Q_{\gamma,k}(\Delta, q, 0)$ and $Q_{\gamma,k}(\Delta, q, 1)$ are non-decreasing with respect to Δ. Therefore, for any $\Delta_1 \leq \Delta_2$, we have

$$\begin{aligned} V_{\gamma,k+1}(\Delta_1, q) &= \min_{a \in \mathcal{A}} \{Q_{\gamma,k}(\Delta_1, q, a)\} \\ &= \min\{Q_{\gamma,k}(\Delta_1, q, 0), Q_{\gamma,k}(\Delta_1, q, 1)\} \\ &\leq \min\{Q_{\gamma,k}(\Delta_2, q, 0), Q_{\gamma,k}(\Delta_2, q, 1)\} \\ &= V_{\gamma,k+1}(\Delta_2, q). \end{aligned} \tag{A8}$$

As a result, with the induction we prove that $V_{\gamma,k}(\Delta, q)$ is a non-decreasing function with respect to Δ for any $q \in \{1, \ldots, B - 1\}$, i.e., the Equation (A4) holds. When k goes to infinity, combining (A3) and (A4), we prove that (17) holds in the case $0 < q < B$. In the other two cases, (17) still holds. So we have proved that (17) holds for any $q \in \mathcal{B}$.

For (18), it is easy to yield

$$\begin{aligned} Q_\gamma(\Delta_2, q, 0) - Q_\gamma(\Delta_1, q, 0) &= \Delta_2 - \Delta_1 \\ &+ \gamma\lambda[V_\gamma(\Delta_2 + 1, q + 1) - V_\gamma(\Delta_1 + 1, q + 1)] \\ &+ \gamma(1 - \lambda)[V_\gamma(\Delta_2 + 1, q) - V_\gamma(\Delta_1 + 1, q)] \\ &\overset{(a)}{\geq} \Delta_2 - \Delta_1, \end{aligned} \tag{A9}$$

and

$$\begin{aligned} Q_\gamma(\Delta_2,q,1) - Q_\gamma(\Delta_1,q,1) &= \Delta_2 - \Delta_1 \\ &+ \gamma p\lambda[V_\gamma(\Delta_2+1,q) - V_\gamma(\Delta_1+1,q)] \\ &+ \gamma p(1-\lambda)[V_\gamma(\Delta_2+1,q-1) - V_\gamma(\Delta_1+1,q-1)] \\ &+ \gamma(1-p)\lambda[V_\gamma(1,q) - V_\gamma(1,q)] \\ &+ \gamma(1-p)(1-\lambda)[V_\gamma(1,q-1) - V_\gamma(1,q-1)] \\ &\overset{(b)}{\geq} \Delta_2 - \Delta_1, \end{aligned} \tag{A10}$$

where (a) and (b) are due to (17). Since $V_\gamma(\mathbf{x}) = \min_{a\in\mathcal{A}} Q_\gamma(\mathbf{x},a)$, we prove that Equation (18) holds for all $q \in \{1,\ldots,B-1\}$. Through the same proof process, it can also be verified that (18) is also valid when $q = 0$ and $q = B$. Therefore, we have completed the proof of part (a).

Second, we will tackle the part (b) of Lemma 1.

For (19), according to the exact same mathematical induction we have applied to (17), we can also verify that Equation (19) holds. Due to limited space, the details are omitted here.

Hence, we have completed the whole proof.

Appendix C. Proof of Theorem 1

By Proposition 4 in [15], it suffices to show that the following four conditions hold:

(1): For every state $\mathbf{x}$ and discount factor γ, the discount value function $V_\gamma(\mathbf{x})$ is finite.
(2): There exists a non-negative value L such that $L \leq h_\gamma(\mathbf{x})$ for all $\mathbf{x}$ and γ, where $h_\gamma(\mathbf{x}) = V_\gamma(\mathbf{x}) - V_\gamma(\hat{\mathbf{x}})$, and $\hat{\mathbf{x}}$ is a reference state.
(3): There exists a non-negative value $M(\mathbf{x})$, such that $h_\gamma(\mathbf{x}) \leq M(\mathbf{x})$ for every $\mathbf{x}$ and γ.
(4): The inequality $\sum_{\mathbf{x}'\in\mathcal{S}} Pr(\mathbf{x}'|\mathbf{x},a)M(\mathbf{x}') < \infty$ holds for all $\mathbf{x}$ and a.

For condition (1), recall that we have verified that there exists a stationary deterministic policy π such that the expected discounted cost V_γ^π is finite in the proof of Proposition 1, and here we will extend this conclusion to any policy $\pi \in \Pi$. For any non-anticipative policy π and state $\mathbf{x} = (\Delta, q)$, we have

$$C(\mathbf{x}[t], a[t]) = \Delta + \omega C_r a(1-u(t)) \leq \Delta + \omega C_r. \tag{A11}$$

Since the AoI grows linearly at most, for any state $\mathbf{x} = (\Delta, q)$ and discounted factor γ, we have

$$\begin{aligned} V_\gamma(\mathbf{x}) &= \min_{\pi\in\Pi} V_\gamma^\pi(\mathbf{x}) = \min_{\pi\in\Pi} \mathbb{E}_\pi\left\{\sum_{t=0}^{\infty} \gamma^t C(\mathbf{x}[t], a[t]) | \mathbf{x}[0] = (\Delta, q)\right\} \\ &\leq \sum_{t=0}^{\infty} \gamma^t(\Delta + t + \omega C_r) \\ &= \frac{1}{1-\gamma}(\Delta + \omega C_r + \frac{\gamma}{1-\gamma}) < \infty, \end{aligned} \tag{A12}$$

which verifies condition (1).

Next let us focus on condition (2). By (17) and (19) in Lemma 1, $V_\gamma(\Delta, q)$ is non-decreasing with regard to age Δ and non-increasing with regard to battery state q. Hence, we can choose $L = 0$ and reference state $\hat{\mathbf{x}} = (1, B)$. Then we have $L = 0 \leq V_\gamma(\mathbf{x}) - V_\gamma(\hat{\mathbf{x}}) = h_\gamma(\mathbf{x})$, which verifies condition (2).

To prove that condition (3) holds, we need to introduce the following lemma:

Lemma A1. *Denote $\hat{\boldsymbol{x}} = (1, B)$ as the reference state, and $T = \inf\{t : t \geq 0, \boldsymbol{x}[t] = \hat{\boldsymbol{x}}\}$ as the first hitting time from the initial state $\boldsymbol{x}$ to $\hat{\boldsymbol{x}}$. Under the following lazy policy π':*

$$\pi'(\Delta, q) = \begin{cases} 1, & \text{if } q = B, \\ 0, & \text{otherwise}, \end{cases} \tag{A13}$$

the expected discounted cost from $\boldsymbol{x}$ to $\hat{\boldsymbol{x}}$ is finite for all initial state $\boldsymbol{x} \in \mathcal{S}$, i.e.,

$$C^{\pi'}(\boldsymbol{x}) = \mathbb{E}_{\pi'}\left\{\sum_{t=0}^{T-1} \gamma^t C(\boldsymbol{x}[t], a[t]) | \boldsymbol{x}[0] = \boldsymbol{x}\right\} < \infty. \tag{A14}$$

Note that if $\boldsymbol{x} = \hat{\boldsymbol{x}}$, $C^{\pi'}(\boldsymbol{x}) = 0$.

Proof. see Appendix F. □

Considering a mixed non-anticipative policy π^m consisting of π' and optimal policy $\pi^\star$ for (12) from the initial state $\mathbf{x}$ as follows,

$$\pi^m(x[t]) = \begin{cases} \pi'(x[t]), & \text{if } t < T, \\ \pi^\star(x[t]), & \text{otherwise}, \end{cases} \tag{A15}$$

we have

$$\begin{aligned} V_\gamma(\mathbf{x}) \leq V_\gamma^{\pi^m}(\mathbf{x}) &= \mathbb{E}_{\pi'}\left\{\sum_{t=0}^{T-1} \gamma^t C(\mathbf{x}[t], a[t]) | \mathbf{x}[0] = \mathbf{x}\right\} + \mathbb{E}_{\pi^\star}\left\{\sum_{t=T}^{\infty} \gamma^t C(\mathbf{x}[t], a[t]) | \mathbf{x}[T] = \hat{\mathbf{x}}\right\} \\ &\overset{(a)}{=} C^{\pi'}(\mathbf{x}) + \mathbb{E}_{\pi^\star}\left\{\gamma^T V_\gamma(\hat{\mathbf{x}})\right\} \\ &\overset{(b)}{\leq} C^{\pi'}(\mathbf{x}) + V_\gamma(\hat{\mathbf{x}}), \end{aligned} \tag{A16}$$

where (a) is due to (A14) and (12), (b) is due to $\gamma \in (0, 1)$. Recall the definition of $h_\gamma(\mathbf{x})$, by setting $M(\mathbf{x}) = C^{\pi'}(\mathbf{x})$, the condition (3) holds.

Based on Lemma A1, $M(\mathbf{x}) = C^{\pi'}(\mathbf{x}) < \infty$ holds for any state $\mathbf{x}$. Since there will be finite possible states after transition from $\mathbf{x}$ under any action, the sum of finite $M(\cdot)$ is also finite. Hence, condition (4) holds.

Appendix D. Proof of Lemma 3

For (27), an equivalent transformation is made as follows:

$$V(\Delta + 1, q + 1) + pV(\Delta, q) \geq V(\Delta, q + 1) + pV(\Delta + 1, q). \tag{A17}$$

For every state $\mathbf{x}$, we have

$$V(\mathbf{x}) = \min_{a \in \mathcal{A}} Q(\mathbf{x}, a) = \min\{Q(\mathbf{x}, 0), Q(\mathbf{x}, 1)\}. \tag{A18}$$

So every value function in (A17) has two possible values. In order to prove Equation (A17), theoretically we need to discuss $2^4 = 16$ cases, which is obviously a bit too cumbersome. Here we use a little trick, that is, as long as we prove that for the $2^2 = 4$ possible combinations on the left hand side(LHS) of (A17), there exists a combination on the right hand side (RHS) of (A17) to make "$\geq$" hold, then we complete the proof. For convenience, we make a mapping by using four numbers to sequentially represent the action taken by the

minimum state–action value function in Equation (A17). For example, "1010" represents the following:

$$Q(\Delta+1,q+1,\mathbf{1})+pQ(\Delta,q,\mathbf{0})\geq Q(\Delta,q+1,\mathbf{1})+pQ(\Delta+1,q,\mathbf{0}), \tag{A19}$$

So according to the previous trick, we only need to verify "0000", "1010", "0101", and "1111" to prove Equation (A17). For brevity, we only show the verification process of "1010" in the following proof. The other three cases can also be proved by exactly the same steps.

Now we start to apply VIA and mathematical induction. Assuming that $V_0(\mathbf{x})=0$ for any states $\mathbf{x}$, it is easy to yield

$$V_1(\Delta+1,q+1)+pV_1(\Delta,q)\geq V_1(\Delta,q+1)+pV_1(\Delta+1,q), \tag{A20}$$

for any $q\in\{0,1,...,B-1\}$ and $\Delta\in\mathbb{Z}^+$. By induction, assuming that for any $q\in\{0,1,\ldots,B-1\}$ and $\Delta\in\mathbb{Z}^+$, we have:

$$V_k(\Delta+1,q+1)+pV_k(\Delta,q)\geq V_k(\Delta,q+1)+pV_k(\Delta+1,q). \tag{A21}$$

What we need to do is to verify that Equation (A21) still holds in the next value iteration. Based on our previous analysis, we will focus on the "1010" case. For $\Delta\in\mathbb{Z}^+$ and $q\in\{0,1,\ldots,B-1\}$, we have

$$\begin{aligned}
&Q_k(\Delta+1,q+1,1)+pQ_k(\Delta,q,0)\\
&-[Q_k(\Delta,q+1,1)+pQ_k(\Delta+1,q,0)]\\
=&\Delta+1+p\lambda V_k(\Delta+2,q+1)+p(1-\lambda)V_k(\Delta+2,q)\\
&+(1-p)\lambda V_k(1,q+1)+(1-p)(1-\lambda)V_k(1,q)\\
&+p[\Delta+\omega C_r+\lambda V_k(\Delta+1,q+1)+(1-\lambda)V_k(\Delta+1,q)]\\
&-\Delta-p\lambda V_k(\Delta+1,q+1)-p(1-\lambda)V_k(\Delta+1,q)\\
&-(1-p)\lambda V_k(1,q+1)-(1-p)(1-\lambda)V_k(1,q)\\
&-p[\Delta+1+\omega C_r+\lambda V_k(\Delta+2,q+1)-(1-\lambda)V_k(\Delta+2,q)]\\
=&1-p\geq 0.
\end{aligned} \tag{A22}$$

Therefore, by the similar step, we can verify the other three cases and confirm that the following formula

$$V_{k+1}(\Delta+1,q+1)+pV_{k+1}(\Delta,q)\geq V_{k+1}(\Delta,q+1)+pV_{k+1}(\Delta+1,q) \tag{A23}$$

holds for any $\Delta\in\mathbb{Z}^+$ and $q\in\{0,1,\ldots,B-1\}$. Therefore, by induction, we prove that for any k, the Equation (A21) holds. Take the limits of k on both sides, then we are able to prove that (A17) holds, which indicates that (27) holds. Therefore, we have completed the proof.

Appendix E. Proof of Theorem 2

By Corollary 1, the optimal policy is of a threshold structure if $Q(\mathbf{x},a)$ has a *sub-modular* structure, that is,

$$Q(\Delta,q,0)-Q(\Delta,q,1)\leq Q(\Delta+1,q,0)-Q(\Delta+1,q,1). \tag{A24}$$

We will divide the whole proof into the following three cases:

Case 1. When $q=0$, for any $\Delta\in\mathbb{Z}^+$ we have:

$$\begin{aligned}
&Q(\Delta,q,0)-Q(\Delta,q,1)\\
=&\Delta+\lambda V(\Delta+1,q+1)+(1-\lambda)V(\Delta+1,q)\\
&-\Delta-\omega C_r-p\lambda V(\Delta+1,q+1)+p(1-\lambda)V(\Delta+1,q)\\
&-(1-p)\lambda V(1,q+1)-(1-p)(1-\lambda)V(1,q)\\
=&(1-p)\lambda(V(\Delta+1,q+1)-V(1,q+1))\\
&+(1-p)(1-\lambda)(V(\Delta+1,q)-V(1,q))-\omega C_r. \qquad \text{(A25)}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&Q(\Delta+1,q,0)-Q(\Delta+1,q,1)-[Q(\Delta,q,0)-Q(\Delta,q,1)]\\
=&(1-p)\lambda(V(\Delta+2,q+1)-V(\Delta+1,q+1))\\
&+(1-p)(1-\lambda)(V(\Delta+2,q)-V(\Delta,q))\\
\overset{(a)}{\geq}&0, \qquad \text{(A26)}
\end{aligned}$$

where the last inequality (a) is due to (24) in Lemma 2.

Case 2. When $q \in \{1,\ldots,B-1\}$, for any $\Delta \in \mathbb{Z}^+$ we have

$$\begin{aligned}
&Q(\Delta+1,q,0)-Q(\Delta+1,q,1)-[Q(\Delta,q,0)-Q(\Delta,q,1)]\\
=&Q(\Delta+1,q,0)-Q(\Delta,q,0)-[Q(\Delta+1,q,1)-Q(\Delta,q,1)]\\
=&\lambda[V(\Delta+2,q+1)-V(\Delta+1,q+1)]\\
&-p\lambda[V(\Delta+2,q)-V(\Delta+1,q)]\\
&+(1-\lambda)[V(\Delta+2,q)-V(\Delta+1,q)]\\
&-p(1-\lambda)[V(\Delta+2,q-1)-V(\Delta+1,q-1)]\\
\overset{(a)}{\geq}&0, \qquad \text{(A27)}
\end{aligned}$$

where the last inequality (a) is due to (27) in Lemma 3.

Case 3. When $q = B$, for any $\Delta \in \mathbb{Z}^+$ we have

$$\begin{aligned}
&Q(\Delta+1,q,0)-Q(\Delta+1,q,1)-[Q(\Delta,q,0)-Q(\Delta,q,1)]\\
=&Q(\Delta+1,q,0)-Q(\Delta,q,0)-[Q(\Delta+1,q,1)-Q(\Delta,q,1)]\\
=&(1-\lambda)[V(\Delta+2,q)-V(\Delta+1,q)]\\
&-p(1-\lambda)[V(\Delta+2,q-1)-V(\Delta+1,q-1)]\\
\overset{(a)}{\geq}&0, \qquad \text{(A28)}
\end{aligned}$$

where the last inequality (a) is also due to (27) in Lemma 3.

Therefore, we have completed the whole proof.

Appendix F. Proof of Lemma A1

Before dealing with the expected discounted cost $C^{\pi'}(\mathbf{x})$, we need to find the probability distribution of the first hitting time T, which is determined by the transition probabilities of system states. Under the lazy policy π', we can formulate a two-dimensional Markov chain to describe the dynamic changes of system states. The state transition probabilities of the formulated Markov chain is extremely complicated, and we can simplify it by combining some states, as depicted in Figure A1.

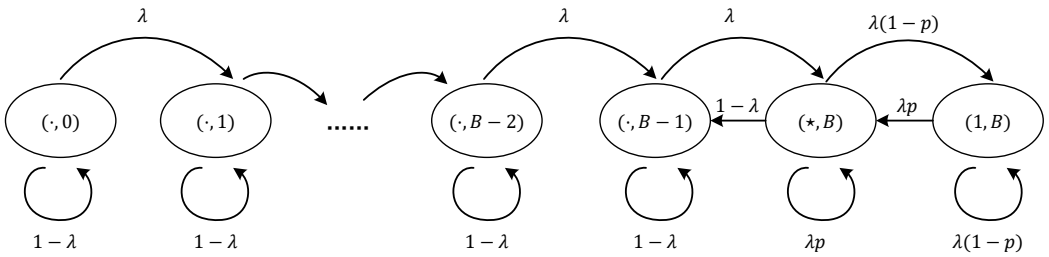

Figure A1. A simplified Markov chain of system states under the lazy policy. Note that $(1, B)$ is the reference state. $(\cdot, 1)$ means the state set $\{(\Delta, q)|\Delta \in \mathbb{Z}^+, q = 1\}$, $(\star, B)$ means the state set $\{(\Delta, q)|\Delta \in \mathbb{Z}^+, \Delta > 1, q = B\}$ and so on for the rest.

According to the simplified Markov chain, the initial state $\mathbf{x}$ can be divided into three cases: $(\star, B)$, $(\cdot, B-1)$, and $(\cdot, q)$ where $q < B-1$. Note that for the special case $\mathbf{x} = \hat{\mathbf{x}}$, $C^{\pi'}(x)$ is set to be 0. First, we focus on the case $\mathbf{x} = (\cdot, q)$, where $q < B-1$. Suppose it takes $T = k$ time slots for state $\mathbf{x}$ to transit to state $\hat{\mathbf{x}}$ for the first time. Then state $\mathbf{x}' = (\cdot, B-1)$ must be passed during these k time slots. Therefore, we can divide the entire transition process into two parts: state $\mathbf{x}$ first visits state $\mathbf{x}'$ after k_1 time slots, and then starts from state $\mathbf{x}'$ and enters state $\hat{\mathbf{x}}$ for the first time after $k_2 = k - k_1$ time slots. Denote $f^{(n)}_{\mathbf{x}_1,\mathbf{x}_2}$ as the first hitting probability from state $\mathbf{x}_1$ to state $\mathbf{x}_2$ after n time slots, then we have

$$f^{(k)}_{\mathbf{x},\hat{\mathbf{x}}} = \sum_{k_1=0}^{k} f^{(k_1)}_{\mathbf{x},\mathbf{x}'} f^{(k_2)}_{\mathbf{x}',\hat{\mathbf{x}}}. \tag{A29}$$

When the initial state first transits to state $\mathbf{x}'$, the total energy arrivals must be exactly $B-q-1$. Hence, the first hitting probability $f^{(k_1)}_{\mathbf{x},\mathbf{x}'}$ from state $\mathbf{x}$ to state $\mathbf{x}'$ can be expressed as follows:

$$\begin{aligned} f^{(k_1)}_{\mathbf{x},\mathbf{x}'} &= \binom{k_1-1}{B-q-2} \lambda^{B-q-2}(1-\lambda)^{k_1-1-(B-q-2)}\lambda \\ &= \binom{k_1-1}{B-q-2} (\frac{\lambda}{1-\lambda})^{B-q-1}(1-\lambda)^{k_1} \\ &\overset{(a)}{\leq} (k_1-1)^{B-q-2}(\frac{\lambda}{1-\lambda})^{B-q-1}(1-\lambda)^{k_1}, \end{aligned} \tag{A30}$$

where $k_1 \geq B-q-1$. The inequality (a) in (A30) is due to combination $\binom{N}{r} \leq N^r, \forall N \geq r$. For any $k_1 < B-q-1$, we have $f^{(k_1)}_{\mathbf{x}\mathbf{x}'} = 0$.

After entering state $\mathbf{x}'$, the system state will always change between states $\mathbf{x}' = (\cdot, B-1)$ and $(\star, B)$ before entering state $\hat{\mathbf{x}}$ for the first time. By mathematical induction, $f^{(k_2)}_{\mathbf{x}',\hat{\mathbf{x}}}$ is given as follows:

$$\begin{aligned} f^{(k_2)}_{\mathbf{x}',\hat{\mathbf{x}}} &= \begin{bmatrix} 1-\lambda & \lambda \end{bmatrix} \begin{bmatrix} 1-\lambda & \lambda \\ 1-\lambda & \lambda p \end{bmatrix}^{k_2-2} \begin{bmatrix} 0 \\ \lambda(1-p) \end{bmatrix} \\ &= (1-p)\lambda^2 \frac{\beta_1^{k_2-1} - \beta_2^{k_2-1}}{\beta_1 - \beta_2} \\ &= (1-p)\lambda^2 \sum_{i=0}^{k_2-2} \beta_1^i \beta_2^{k_2-2-i} \\ &\overset{(a)}{<} (1-p)\lambda^2 (k_2-1)\beta_1^{k_2-2}, \end{aligned} \tag{A31}$$

where $k_2 \geq 2$, β_1 and β_2 are the eigenvalues of the matrix $\begin{bmatrix} 1-\lambda & \lambda \\ 1-\lambda & \lambda p \end{bmatrix}$ and satisfy $-1 < \beta_2 < 0 < 1-\lambda < \beta_1 < 1$. The last inequality (a) of (A31) is due to $\beta_2 < 0 < \beta_1$ and $|\beta_2| < |\beta_1|$. For any $k_2 < 2$, we have $f^{(k_2)}_{\mathbf{x}',\hat{\mathbf{x}}} = 0$.

Therefore, we will verify the discounted cost from the initial state $\mathbf{x}$ to reference state $\hat{\mathbf{x}}$ is finite as follows:

$$\begin{aligned} C^{\pi'}(\mathbf{x}) &= \mathbb{E}_{\pi'}\left\{\sum_{t=0}^{T-1}\gamma^t C(\mathbf{x}[t], a[t])|\mathbf{x}[0]=\mathbf{x}\right\} \\ &\overset{(a)}{\leq} \sum_{k=0}^{\infty} f^{(k)}_{\mathbf{x},\hat{\mathbf{x}}}[\sum_{t=0}^{k}(\Delta + t + \omega C_r)] \\ &\overset{(b)}{=} \sum_{k=B-q+1}^{\infty}\sum_{k_1=B-q-1}^{k} f^{(k_1)}_{\mathbf{x},\mathbf{x}'} f^{(k_2)}_{\mathbf{x}',\hat{\mathbf{x}}}[\sum_{t=0}^{k}(\Delta + t + \omega C_r)] \\ &\overset{(c)}{\leq} (1-p)\lambda^2 \frac{(\frac{1-\lambda}{\beta_1})^{B-q-1}}{1-\frac{1-\lambda}{\beta_1}} \sum_{k=2}^{\infty}\beta_1^{k-2}k^{B-q-1}[\sum_{t=0}^{k}(\Delta + t + \omega C_r)] \\ &\overset{(d)}{<} \infty. \end{aligned} \tag{A32}$$

where inequality (a) is due to (A11), equality (b) is due to (A29), inequality (c) is due to (A30) and (A31), and inequality (d) is due to $0 < \beta_1 < 1$.

For the other two case where the initial state is $(\cdot, B-1)$ or $(\star, B)$, the discounted cost to the reference state for the first time can also be verified to be finite by similar steps. Therefore, we have completed the proof of Lemma A1.

References

1. Kaul, S.; Yates, R.; Gruteser, M. Real-time status: How often should one update? In Proceedings of the IEEE INFOCOM, Orlando, FL, USA, 25–30 March 2012; pp. 2731–2735.
2. Sun, Y.; Kadota, I.; Talak, R.; Modiano, E. Age of information: A new metric for information freshness. *Synth. Lect. Commun. Netw.* **2019**, *12*, 1–224. [CrossRef]
3. Yates, R.D.; Sun, Y.; Brown, D.R.; Kaul, S.K.; Modiano, E.; Ulukus, S. Age of information: An introduction and survey. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 1183–1210. [CrossRef]
4. Sun, Y.; Uysal-Biyikoglu, E.; Yates, R.D.; Koksal, C.E.; Shroff, N.B. Update or wait: How to keep your data fresh. *IEEE Trans. Inf. Theory* **2017**, *63*, 7492–7508. [CrossRef]
5. Kadota, I.; Sinha, A.; Uysal-Biyikoglu, E.; Singh, R.; Modiano, E. Scheduling policies for minimizing age of information in broadcast wireless networks. *IEEE/ACM Trans. Netw.* **2018**, *26*, 2637–2650. [CrossRef]
6. Hsu, Y.P.; Modiano, E.; Duan, L. Scheduling algorithms for minimizing age of information in wireless broadcast networks with random arrivals. *IEEE Trans. Mob. Comput.* **2019**, *19*, 2903–2915. [CrossRef]
7. Tang, H.; Wang, J.; Song, L.; Song, J. Minimizing age of information with power constraints: Multi-user opportunistic scheduling in multi-state time-varying channels. *IEEE J. Sel. Areas Commun.* **2020**, *38*, 854–868. [CrossRef]
8. Jackson, N.; Adkins, J.; Dutta, P. Capacity over capacitance for reliable energy harvesting sensors. In Proceedings of the 18th International Conference on Information Processing in Sensor Networks, Montreal, QC, Canada, 16–18 April 2019; pp. 193–204.
9. Ma, D.; Lan, G.; Hassan, M.; Hu, W.; Das, S.K. Sensing, computing, and communications for energy harvesting IoTs: A survey. *IEEE Commun. Surv. Tutorials* **2019**, *22*, 1222–1250. [CrossRef]
10. Sudevalayam, S.; Kulkarni, P. Energy harvesting sensor nodes: Survey and implications. *IEEE Commun. Surv. Tutorials* **2010**, *13*, 443–461. [CrossRef]
11. TEXAS Instruments. BQ25505 Ultra Low-Power Boost Charger with Battery Management and Autonomous Power Multiplexer for Primary Battery in Energy Harvester Applications. *BQ25505 Datasheet* **2019**, *3*. Available online: https://www.ti.com/lit/ds/symlink/bq25505.pdf (accessed on 10 March 2019).
12. Wu, X.; Tan, L.; Tang, S. Optimal Energy Supplementary and Data Transmission Schedule for Energy Harvesting Transmitter With Reliable Energy Backup. *IEEE Access* **2020**, *8*, 161838–161846. [CrossRef]
13. Wu, J.; Chen, W. Delay-Optimal Scheduling for Energy Harvesting Aided mmWave Communications with Random Blocking. In Proceedings of the ICC 2020—2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 7–11 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.

14. Draskovic, S.; Thiele, L. Optimal Power Management for Energy Harvesting Systems with A Backup Power Source. In Proceedings of the 2021 10th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 7–10 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–9.
15. Sennott, L.I. Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs. *Oper. Res.* **1989**, *37*, 626–633. [CrossRef]
16. Yates, R.D. Lazy is timely: Status updates by an energy harvesting source. In Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT), Hong Kong, China, 14–19 June 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 3008–3012.
17. Bacinoglu, B.T.; Ceran, E.T.; Uysal-Biyikoglu, E. Age of information under energy replenishment constraints. In Proceedings of the 2015 Information Theory and Applications Workshop (ITA), San Diego, CA, USA, 1–6 February 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 25–31.
18. Arafa, A.; Ulukus, S. Age-minimal transmission in energy harvesting two-hop networks. In Proceedings of the GLOBECOM 2017—2017 IEEE Global Communications Conference, Singapore, 4–8 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
19. Arafa, A.; Ulukus, S. Timely updates in energy harvesting two-hop networks: Offline and online policies. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 4017–4030. [CrossRef]
20. Arafa, A.; Ulukus, S. Age minimization in energy harvesting communications: Energy-controlled delays. In Proceedings of the 2017 51st Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 29 October–1 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1801–1805.
21. Wu, X.; Yang, J.; Wu, J. Optimal status update for age of information minimization with an energy harvesting source. *IEEE Trans. Green Commun. Netw.* **2017**, *2*, 193–204. [CrossRef]
22. Arafa, A.; Yang, J.; Ulukus, S.; Poor, H.V. Age-minimal transmission for energy harvesting sensors with finite batteries: Online policies. *IEEE Trans. Inf. Theory* **2019**, *66*, 534–556. [CrossRef]
23. Bacinoglu, B.T.; Sun, Y.; Uysal, E.; Mutlu, V. Optimal status updating with a finite-battery energy harvesting source. *J. Commun. Netw.* **2019**, *21*, 280–294. [CrossRef]
24. Feng, S.; Yang, J. Age of information minimization for an energy harvesting source with updating erasures: Without and with feedback. *IEEE Trans. Commun.* **2021**, *69*, 5091–5105. [CrossRef]
25. Arafa, A.; Yang, J.; Ulukus, S.; Poor, H.V. Timely Status Updating Over Erasure Channels Using an Energy Harvesting Sensor: Single and Multiple Sources. *IEEE Trans. Green Commun. Netw.* **2021**, *6*, 6–19 . [CrossRef]
26. Baknina, A.; Ulukus, S. Coded status updates in an energy harvesting erasure channel. In Proceedings of the 2018 52nd Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, 21–23 March 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.
27. Baknina, A.; Ozel, O.; Yang, J.; Ulukus, S.; Yener, A. Sending information through status updates. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 2271–2275.
28. Ceran, E.T.; Gündüz, D.; György, A. Reinforcement learning to minimize age of information with an energy harvesting sensor with HARQ and sensing cost. In Proceedings of the IEEE INFOCOM 2019—IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Paris, France, 29 April–2 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 656–661.
29. Hentati, A.; Frigon, J.F.; Ajib, W. Energy harvesting wireless sensor networks with channel estimation: Delay and packet loss performance analysis. *IEEE Trans. Veh. Technol.* **2019**, *69*, 1956–1969. [CrossRef]
30. Leng, S.; Yener, A. Age of Information Minimization for an Energy Harvesting Cognitive Radio. *IEEE Trans. Cogn. Commun. Netw.* **2019**, *5*, 427–439. [CrossRef]
31. Zheng, X.; Zhou, S.; Jiang, Z.; Niu, Z. Closed-form analysis of non-linear age of information in status updates with an energy harvesting transmitter. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 4129–4142. [CrossRef]
32. Lu, Y.; Xiong, K.; Fan, P.; Zhong, Z.; Letaief, K.B. Online transmission policy in wireless powered networks with urgency-aware age of information. In Proceedings of the 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), Tangier, Morocco, 24–28 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1096–1101.
33. Saurav, K.; Vaze, R. Online energy minimization under a peak age of information constraint. In Proceedings of the 2021 19th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt), Virtual, 18–21 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8.
34. Abd-Elmagid, M.A.; Dhillon, H.S. Closed-form characterization of the MGF of AoI in energy harvesting status update systems. *IEEE Trans. Inf. Theory* **2022**, *68*, 3896–3919. [CrossRef]
35. Gong, J.; Chen, X.; Ma, X. Energy-age tradeoff in status update communication systems with retransmission. In Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 9–13 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.
36. Huang, H.; Qiao, D.; Gursoy, M.C. Age-energy tradeoff in fading channels with packet-based transmissions. In Proceedings of the IEEE INFOCOM 2020—IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada, 6–9 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 323–328.
37. Gu, Y.; Chen, H.; Zhou, Y.; Li, Y.; Vucetic, B. Timely status update in Internet of Things monitoring systems: An age-energy tradeoff. *IEEE Internet Things J.* **2019**, *6*, 5324–5335. [CrossRef]

38. Nath, S.; Wu, J.; Yang, J. Optimum energy efficiency and age-of-information tradeoff in multicast scheduling. In Proceedings of the 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 20–24 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.
39. Gong, J.; Zhu, J.; Chen, X.; Ma, X. Sleep, Sense or Transmit: Energy-Age Tradeoff for Status Update with Two-Thresholds Optimal Policy. *IEEE Trans. Wirel. Commun.* **2021**, *21*, 1751–1765. [CrossRef]
40. Wang, L.; Peng, F.; Chen, X.; Zhou, S. Optimal Update for Energy Harvesting Sensor with Reliable Backup Energy. *arXiv* **2022**, arXiv:2201.01686.
41. Valentini, R.; Levorato, M. Optimal aging-aware channel access control for wireless networks with energy harvesting. In Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 2754–2758.
42. Dong, Y.; Fan, P.; Letaief, K.B. Energy harvesting powered sensing in IoT: Timeliness versus distortion. *IEEE Internet Things J.* **2020**, *7*, 10897–10911. [CrossRef]
43. Gindullina, E.; Badia, L.; Gündüz, D. Age-of-information with information source diversity in an energy harvesting system. *IEEE Trans. Green Commun. Netw.* **2021**, *5*, 1529–1540. [CrossRef]
44. Sennott, L.I. Constrained average cost Markov decision chains. *Probab. Eng. Informational Sci.* **1993**, *7*, 69–83. [CrossRef]
45. Puterman, M.L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
46. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
47. Das, T.K.; Gosavi, A.; Mahadevan, S.; Marchalleck, N. Solving semi-Markov decision problems using average reward reinforcement learning. *Manag. Sci.* **1999**, *45*, 560–574. [CrossRef]
48. Ceran, E.T.; Gündüz, D.; György, A. Average age of information with hybrid ARQ under a resource constraint. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 1900–1913. [CrossRef]