Approximately Optimal Policies for a Class of Markov Decision Problems with Applications to Energy Harvesting

Dor Shaviv and Ayfer Özgür
Department of Electrical Engineering
Stanford University
Email: {shaviv, aozgur}@stanford.edu

***Abstract*—We consider a general class of stochastic optimization problems, in which the state represents a certain level or amount which can be partly used and depleted, and subsequently filled by a random amount. This is motivated by energy harvesting applications, in which one manages the amount of energy in a battery, but is also related to inventory models and queuing models. We propose a simple policy that requires minimal knowledge of the distribution of the stochastic process involved, and show that it is a close approximation to the optimal solution with bounded guarantees. Specifically, under natural assumptions on the reward function, we provide constant multiplicative and additive gaps to optimality, which do not depend on the problem parameters. This allows us to obtain a simple formula for approximating the long-term expected average reward, which gives some insight on its qualitative behavior as a function of the maximal state and the distribution of the disturbance.**

I. Introduction

We present a general class of Markov decision problems (MDPs), motivated by energy harvesting applications, but which also contains other problems of practical interest such as inventory management and queue optimization. Roughly speaking, the state represents an amount of a certain quantity, for example amount of energy in a battery or amount of product in an inventory. Each time slot, we can choose to expend a portion of this amount to gain a certain reward, which is typically increasing with the amount expended. At the next time slot, the amount available is the amount left after depletion of the amount expended, plus some new amount which is determined by an exogenous stochastic i.i.d. process. However, if the new amount exceeds some maximal quantity (such as battery size or storage capacity, for example), the excess amount is wasted.

We are interested in the long-term expected average reward of this model, under an *online* policy, i.e. when the future values of the disturbance are not known ahead of time. As an MDP, this problem can be solved using dynamic programming. However, this approach has several shortcomings. First, usually the problem cannot be solved explicitly and must be solved numerically. While there exist methods to numerically find a solution which is arbitrarily close to optimal, such as value iteration and policy iteration, these methods require quantization of the state and action spaces to finite sets. Specifically, the computational complexity of each iteration of these algorithms grows as the cube of the number of quantized states and/or actions. Second, the solution depends heavily on the exact statistical distribution of the exogenous stochastic process, which may be hard to obtain in practice. If this distribution is estimated from measurements, it may require recomputing the dynamic programming solution periodically to track changes in the process. Finally, the numerical solution does not provide much insight on the structure of the optimal policy and the qualitative behavior of the resultant average reward, namely how it varies with the parameters of the problem. This kind of insight can be critical for design considerations, such as choosing the size of the battery or storage capacity.

In this work, we provide a simple suboptimal policy that is provably close to optimal across all parameter regimes and any disturbance distribution. In particular, we find a policy that achieves the optimal long-term expected average reward of the problem simultaneously within a constant multiplicative factor of 2, also called a *2-approximation algorithm*, and a constant additive gap for all parameter values and disturbance distributions. Moreover, this simple policy has minimal dependence on the distribution of the disturbance, namely it depends only on its mean. This enables one to apply it to any given problem with arbitrary parameter values, without even knowing the exact distribution of the exogenous process, while she/he would be assured to achieve a performance that is very close to the one achieved by an optimal policy specifically optimized for the given problem, in particular the exact distribution of the disturbance.

Our policy is based on previous work [1], in which a similar result was derived for a specific example in this general class of MDPs (namely ex. a in Section II). The policy can be described as follows: at each time-slot, the policy uses a constant fraction of the available amount (i.e. the state), where the fraction is chosen as the ratio of the mean of the disturbance and the maximal state. We show that it is naturally motivated by the case where the disturbance is a binary random variable, in which case the optimal policy can be explicitly characterized. We then establish the near-optimality of this policy for any i.i.d. disturbance. In particular, we show that this policy achieves the optimal long-term average reward of the system simultaneously within a constant multiplicative

978-3-901882-90-6/17 ©2017 IFIP

factor and a constant additive gap for all parameter values. This implies that this policy can be applied under any i.i.d. disturbance, without even knowing its statistical distribution. The main ingredient of our proof is to show that for the proposed policy, a binary disturbance is the worst disturbance. Therefore, the performance of the scheme under a binary disturbance provides a lower bound on its performance under any i.i.d. disturbance. In this sense, our policy can be thought of as building on the insights from the worst-case scenario, hence performs well in the worst-case sense.

This result also leads to a simple approximation of the optimal long-term expected average reward of this class of models. In particular, we show that within a constant gap, the average expected reward is given by

$$J^* \approx r\left(\mathbb{E}[\min(w_t, \bar{x})]\right), \tag{1}$$

where $r(\,\cdot\,)$ is the reward function, w_t is the disturbance, and $\bar{x}$ is the maximal state.

A. Related Work

MDPs of this type have been studied mostly in the context of power control for energy harvesting communication [1]–[23]. While the problem can be solved numerically using dynamic programming [5]–[11], there has been significant effort in the recent literature to develop simple heuristic online policies [14]–[23]. However, these policies come either with no guarantees or only asymptotic guarantees on optimality.

II. Model

We consider an MDP with state $x_t \in [0, \bar{x}]$ for some fixed $\bar{x} > 0$. The action is a non-negative real number u_t, which must be less than or equal to the state x_t. Therefore the action space is $\mathcal{U}(x_t) = [0, x_t]$. There is a disturbance process $w_t \in \mathcal{W}$, which is assumed to be non-negative, and distributed i.i.d. independently of the state and action according to some distribution P_w. The state dynamics are

$$x_{t+1} = \min\left(x_t - u_t + w_t,\, \bar{x}\right). \tag{2}$$

Note that without loss of generality, we can assume $\mathcal{W} \subseteq [0, \bar{x}]$, since if $w_t \geq \bar{x}$ the next state will be $x_{t+1} = \bar{x}$ regardless of the value of w_t. This explains the term $\min(w_t, \bar{x})$ in (1); the process w_t can be equivalently replaced by the process $\min(w_t, \bar{x})$. The reward at time t is $r(u_t)$, where $r(\,\cdot\,)$ is a non-negative, non-decreasing, and concave utility function. We assume $r(0) = 0$; otherwise we can take $\tilde{r}(x) = r(x) - r(0)$ to be the reward function. A policy π is a set of mappings $\mu_t : [0, \bar{B}] \to \mathbb{R}^+$, $t = 1, 2, \ldots$, such that $\mu_t(x_t) \in [0, x_t]$. The goal is to maximize the long-term average expected reward:

$$J^* = \sup_{\pi} \liminf_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \mathbb{E}[r(\mu_t(x_t))]. \tag{3}$$

Many problems fall under this class of MDPs. We bring here a few examples:

a) Throughput maximization for energy harvesting nodes: Consider a point-to-point communication channel with additive white Gaussian noise (AWGN), in which the transmitter harvests random energy from the environment. By allocation power u_t at time t, the instantaneous rate is given by the AWGN capacity formula $r(u) = \frac{1}{2}\log(1 + \frac{u_t}{\sigma^2})$, where σ^2 is the noise variance. The state x_t is the amount of energy in the transmitter's battery, which is non-negative and must be less than $\bar{x}$, the battery size. The disturbance w_t is the amount of energy harvested at time t.

b) Source distortion minimization for energy harvesting nodes: Suppose a sensor node observes an i.i.d. Gaussian source with mean zero and variance σ^2, and transmits a compressed version of the source over an AWGN channel with noise variance 1. Assume the node harvests random energy w_t each time slot, and has a battery of size $\bar{x}$ as before. By allocating power u_t at time t, the node can transmit at rate $R = \frac{1}{2}\log(1 + u_t)$. The rate-distortion function of a Gaussian source is $R(D) = \frac{1}{2}\log(\sigma^2/D)$, hence the *instantaneous distortion* incurred by transmitting at rate R is

$$\begin{aligned} D &= \sigma^2 2^{-2R} \\ &= \sigma^2 2^{-2\cdot\frac{1}{2}\log(1+u_t)} \\ &= \frac{\sigma^2}{1+u_t}. \end{aligned}$$

We are interested in finding the policy that minimizes the long-term average distortion, so to put it in the reward maximization framework we define the reward function as $r(u) = -\frac{\sigma^2}{1+u}$.

c) Supplier inventory management: Consider a supplier of a certain product, that produces or harvests a random amount w_t every period. The supplier stores the product in a storage facility that can hold at most $\bar{x}$ amount of the product. Let x_t be the amount of product in storage at period t. Each period, the supplier can decide to sell an amount u_t of the product available in storage. Thus $0 \leq u_t \leq x_t$. The amount that will be available in the inventory at the next period is given by $x_{t+1} = \min(x_t - u_t + w_t,\, \bar{x})$, since if the storage facility is full, the excess product is discarded. For selling amount u, the supplier earns revenue $r(u)$. It is assumed that the supplier provides a discount of the price per unit when selling larger quantities; hence the increase in revenue generated by selling one more unit of product will get smaller as u grows larger. This suggests that the reward function $r(u)$ should be concave, in addition to the natural assumptions of non-negative and non-decreasing. The problem of maximizing the average revenue generated per period falls under the model proposed here.

d) Queue optimization: Suppose a service provider or a retailer attends a queue of customers. The amount of customers in the queue is represented by the state x_t. At each time period, the retailer can choose to service u_t customers, and at the same time a random amount of w_t new customers arrives into the queue. The queue is assumed to have a maximum capacity of $\bar{x}$; customers arriving when the queue is full will leave. The state evolution function is given as before. The retailer receives a reward $r(u)$ for servicing u customers, however there is a loss incurred by servicing too many customers. Hence, as in the previous example, it is natural to assume that $r(u)$ is non-decreasing and concave.

In general, the optimal policy can be found via dynamic programming, by solving the Bellman equation.

Proposition 1 (Bellman Equation [24, Theorem 6.1])**.** *If there exists a scalar $\lambda \in \mathbb{R}_+$ and a bounded function $h : [0, \bar{x}] \to \mathbb{R}_+$ that satisfy*

$$\lambda + h(x) = \max_{0 \le u \le x} \Big[r(u) + \mathbb{E}\big[h\big(\min(x - u + w,\, \bar{x}) \big) \big] \Big] \quad (4)$$

for all $0 \le x \le \bar{x}$, then the maximal average expected reward is $J^* = \lambda$. Additionally, if $u^*(x)$ achieves the maximum in (4) *then the optimal policy is stationary (i.e. it does not depend on t) and is given by $\mu_t^*(x_t) = u^*(x_t)$.*

The functional equation (4) is hard to solve explicitly, and requires an exact model for the statistical distribution of the disturbance w, which may be hard to obtain in practical scenarios. The equation can be solved numerically using value iteration [25], but this can be computationally demanding, especially when the state and actions need to be quantized, and the numerical solution cannot provide insight as to the structure of the optimal policy and the qualitative behavior of the optimal throughput, namely how it varies with the parameters of the problem.

In the sequel, we propose an explicit policy and show that it is within a constant gap to optimality for all disturbance distributions. This policy depends on the disturbance distribution only through its mean $\mathbb{E}[w]$. It also leads to a simple and insightful approximation of the optimal average reward. We first discuss a special case in which the optimal solution can be explicitly found. This inspires the approximately optimal policy for general disturbance distributions.

III. Binary Disturbance

We consider a special case, in which the disturbance is binary: $w_t \in \{0, \bar{x}\}$ and $\Pr(w_t = \bar{x}) = p$. That is, the state evolves according to the following transitions:

$$x_{t+1} = \begin{cases} \bar{x} & \text{w.p. } p, \\ x_t - u_t & \text{w.p. } 1 - p. \end{cases} \quad (5)$$

This special case can be solved explicitly, as detailed in the following theorem and proved in Appendix A.

Theorem 1. *Let j_t be the last time in which the state was $\bar{x}$, i.e.*

$$j_t = \{\sup\ \tau \le t : \ x_\tau = \bar{x}\}.$$

If the reward function $r(u)$ is differentiable and strictly concave, then the optimal policy is given by

$$\mu_t^*(x_1, \ldots, x_t) = (r')^{-1} \left(\frac{\nu}{p(1-p)^{t - j_t}} \right),$$

where $(r')^{-1}(u)$ is the inverse function of the derivative of $r(u)$, i.e. $(r')^{-1}(r'(u)) = u$. The parameter ν is the solution to the equation

$$\bar{x} = \sum_{i=1}^{\tilde{N}(\nu)} (r')^{-1} \left(\frac{\nu}{p(1-p)^{i-1}} \right),$$

where

$$\tilde{N}(\nu) = \left\lfloor 1 + \frac{\log \nu - \log(p r'(0))}{\log(1-p)} \right\rfloor .$$

Remark. While the optimal policy is stated as a function of all the past states $x_1, \ldots, x_t$, it is shown in Appendix A that this is equivalent to a stationary policy, i.e. μ_t^* can be written as a time-invariant function of only the current state x_t.

For the purpose of extending this policy to general i.i.d. disturbances in the next section, it is useful to simplify it to the following form by preserving its exponentially decaying structure:

$$\mu_t(x_1, \ldots, x_t) = \bar{x} p (1-p)^{t - j_t}, \quad (6)$$

where j_t is the time of the last positive disturbance, as defined above. With this simplified policy, the action decreases exactly exponentially with the time elapsed since the last time the state was $\bar{x}$. Note that the factor $\bar{x}p$ was chosen so that the sum $\sum_{k=j_t}^{\infty} \bar{x}p(1-p)^{k-j_t}$ is $\bar{x}$, which ensures this is an admissible policy. Another way to view this policy is that we always use p fraction of the state, i.e.

$$\mu(x_t) = p x_t, \quad (7)$$

where x_t is the state given by $x_t = (1-p)^{t-j_t}\bar{x}$. Hence, it is a stationary policy.

This simplified policy can be intuitively motivated as follows: for the binary disturbance w_t, the inter-arrival time is a geometric random variable with parameter p. Because the geometric random variable is memoryless and has mean $1/p$, at each time step the expected time to the next energy arrival is $1/p$. Since $r(u)$ is a concave function, uniform allocation of the actions maximizes the reward, i.e. if the current state is x_t and we knew that the next time the state would be $\bar{x}$ is in exactly m time slots, allocating x_t/m to each of the next m time slots would maximize the total reward. For the online case of interest here, we can instead use the expected time: since at each time slot, the expected time to the next positive disturbance is $1/p$, we always allocate a fraction p of the state. Fig. 1 illustrates this policy.

This policy is clearly suboptimal, however in what follows we will show that it is within constant multiplicative and additive gaps to optimality, for all values of $\bar{x}$ and p. Before we state these results, we present the following simple upper bound on the maximal expected average reward, which is true for any disturbance distribution (not just the binary case discussed in this section).

Proposition 2. *The optimal average reward is upper bounded by*

$$J^* \le r(\mathbb{E}[w_t]).$$

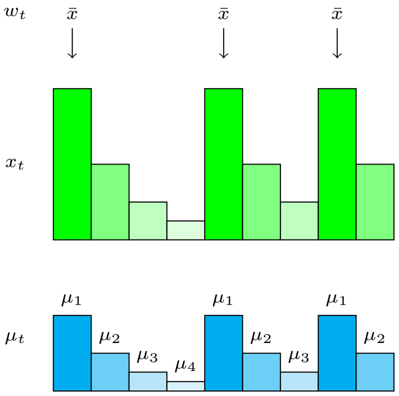

Fig. 1. The approximately-optimal policy for binary disturbance.

The proof follows from a simple application of (2) and Jensen's inequality, and hence will be omitted.

Denote by J_π the expected average reward obtained by our simplified policy. Next, we lower bound J_π in terms of this upper bound.

Proposition 3. *The expected average reward of the suggested policy J_π under the binary disturbance in* (5) *is bounded below by*

$$J_\pi \geq \frac{1}{2} r(\mathbb{E}[w_t]). \tag{8}$$

This is proved in Section V-A. This multiplicative gap means that using our simple policy incurs at most 50% loss in performance relative to the optimal policy. This is sometimes referred to as a *2-approximation algorithm*.

In what follows, under an extra assumption on the reward function, we provide a lower bound in the form of an *additive* gap to optimality, which is especially useful when the maximal state $\bar{x}$ is very large.

Assumption 1 (Sub-logarithmic differences)**.** There exists a positive constant η such that for every $x \leq y$:

$$r(y) - r(x) \leq \eta \log \frac{y}{x}.$$

Note that this holds for the throughput maximization reward function discussed in Section II with $\eta = \frac{1}{2}$:

$$\begin{aligned} r(y) &= \frac{1}{2}\log(1+y) \\ &\leq \frac{1}{2}\log\left(\frac{y}{x} + y\right) \\ &= \frac{1}{2}\log\frac{y}{x} + \frac{1}{2}\log(1+x) \\ &= \frac{1}{2}\log\frac{y}{x} + r(x). \end{aligned}$$

In general, if the reward function $r(u)$ is differentiable and the derivative satisfies $r'(u) \leq \frac{c}{u}$ for some $c \geq 0$, we have:

$$\begin{aligned} r(y) - r(x) &= \int_x^y r'(u)du \\ &\leq \int_x^y \frac{c}{u} du \\ &= \frac{c}{\log e} \log \frac{y}{x}. \end{aligned}$$

The source distortion minimization reward function from Section II, for example, satisfies

$$r'(u) = \frac{\sigma^2}{(1+u)^2} \leq \frac{\sigma^2}{u}.$$

Proposition 4. *If Assumption 1 holds and the disturbance is binary as in* (5)*, then the expected average reward obtained by the suggested policy J_π is bounded by*

$$J_\pi \geq r(\mathbb{E}[w_t]) - \eta \log e \tag{9}$$

See Section V-B for the proof. This proposition, along with the upper bound in Proposition 2, suggests that the approximate policy is within $\eta \log e$ of optimality, regardless of $\bar{x}$ and the distribution of w_t, while the parameter η depends only on the reward function $r(u)$.

IV. Approximately Optimal Policy for General Disturbance

We now assume that w_t is an i.i.d. process with an arbitrary distribution P_w. As discussed in Section II, finding the optimal solution for this general case is a hard problem. In this section, we present a natural extension of the approximately optimal policy (7) in the binary disturbance case and show that it is approximately optimal for any disturbance distribution. The policy reduces to (7) when the disturbance is binary.

The Fixed Fraction Policy: Let $q \triangleq \mathbb{E}[w_t]/\bar{x}$. Note that $\mathbb{E}[w_t] \in [0, \bar{x}]$ so $q \in [0, 1]$. We will use q here instead of the parameter p in the binary case. Notice that in that case, we also have $\mathbb{E}[w_t] = p\bar{x}$, hence this is a natural definition. The Fixed Fraction Policy is defined as follows:

$$\mu(x_t) = q x_t. \tag{10}$$

Inspired by (7), at each time slot, this policy allocates a fraction q of the currently available amount (i.e. the state). Clearly this is an admissible policy, since $q \leq 1$.

The main result of this paper is that the Fixed Fraction Policy achieves the upper bound in Proposition 2 within a constant multiplicative factor and a constant additive gap for any i.i.d. disturbance process. We prove this result by showing that under this policy, the binary disturbance process yields the *worst* performance compared to all other i.i.d. disturbances with the same mean. This implies that the lower bounds obtained for the expected average reward achieved under the binary disturbance of (5) apply also to any disturbance with the same mean, giving the following theorem.

Theorem 2. *Let w_t be an i.i.d. non-negative process with bounded support $[0, \bar{x}]$, and let π be the Fixed Fraction Policy* (10)*. Then, the long-term expected average reward achieved by π is bounded by*

$$J_\pi \geq \frac{1}{2} r(\mathbb{E}[w_t]). \tag{11}$$

Furthermore, if Assumption 1 holds for the reward function $r(u)$, then the expected average reward is also bounded by

$$J_\pi \geq r(\mathbb{E}[w_t]) - \eta \log e. \tag{12}$$

The proof of this theorem is given in Section V-C. The following approximation for the optimal expected average reward is an immediate corollary of the above theorem and proposition 2.

Corollary 1. *The optimal long-term expected average reward J^* under any i.i.d. disturbance process w_t is bounded by*

$$\frac{1}{2} \leq \frac{J^*}{r(\mathbb{E}[w_t])} \leq 1,$$

and if in addition Assumption 1 holds, then

$$r(\mathbb{E}[w_t]) - \eta \log e \leq J^* \leq r(\mathbb{E}[w_t]).$$

This corollary gives a simple approximation of how the optimal average reward depends on the disturbance w_t and the maximal state $\bar{x}$.

V. Lower Bounds on the Average Reward

A. Multiplicative Gap for Binary Disturbance: Proof of Proposition 3

Before establishing the approximate optimality of the suggested policy, we provide a few definitions and results from renewal theory.

Definition 1. A stochastic process $\{X_t\}_{t=1}^{\infty}$ is called a *non-delayed regenerative process* if there exists a random time $\tau > 0$ such that the process $\{X_{\tau+t}\}_{t=1}^{\infty}$ has the same distribution as $\{X_t\}_{t=1}^{\infty}$ and is independent of the past (τ, X^τ).

Observe that a regenerative process is composed of i.i.d. "cycles" or *epochs*, which have i.i.d. durations $\tau_1, \tau_2, \ldots$. At the beginning of each epoch, the process "regenerates" and all memory of the past is essentially erased. The following lemma establishes an important time-average property of regenerative processes.

Lemma 1 (LLN for Regenerative Processes)**.** *Let $\{X_t\}_{t=1}^{\infty}$, $X_t \in \mathcal{X}$, be a non-delayed regenerative process with associated epoch duration τ, and let $f : \mathcal{X} \to \mathbb{R}$. If $\mathbb{E}\tau < \infty$ and $\mathbb{E}[\sum_{t=1}^{\tau} |f(X_t)|] < \infty$ then:*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} f(X_t) = \frac{1}{\mathbb{E}\tau} \mathbb{E}\left[\sum_{t=1}^{\tau} f(X_t)\right] \quad \textit{a.s.}$$

This is an immediate consequence of Theorem 3.1 in [26, Ch. VI] or of the renewal reward theorem [27, Prop. 7.3].

Going back to our MDP with binary disturbance, denote by L the random time between two consecutive positive disturbances, i.e. the time between consecutive occurrences of the state $x_t = \bar{x}$. This is called an *epoch*. Evidently, $L \sim \text{Geometric}(p)$. That is,

$$\Pr(L = k) = p(1-p)^{k-1} \qquad , k = 1, 2, \ldots$$

Since the optimal long-term average reward does not depend on the initial state x_1, we assume without loss of generality that $x_1 = \bar{x}$.

Equipped with Lemma 1, we consider the policy (6) (or equivalently (7)). Observe that $u_t = \mu(x_t) = px_t$ is a non-delayed regenerative process with epoch duration L. We apply Lemma 1 with $f(x) = r(x)$. Note that $\mathbb{E}L = 1/p < \infty$ and $\mathbb{E}[\sum_{t=1}^{L} |r(u_t)|] \leq \mathbb{E}[L \cdot r(\bar{x})] < \infty$, so the conditions of the lemma are satisfied. We obtain

$$\lim_{N\to\infty} \frac{1}{N} r(u_t) = \frac{1}{\mathbb{E}[L]} \mathbb{E}\left[\sum_{t=1}^{L} r(u_t)\right] \qquad \text{a.s.} \tag{13}$$

We proceed to lower bound the expected average reward obtained by our suggested policy:

$$\begin{aligned}
J_\pi &= \liminf_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \mathbb{E}[r(u_t)] \\
&\overset{\text{(i)}}{\geq} \mathbb{E}\left[\liminf_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} r(u_t)\right] \\
&\overset{\text{(ii)}}{=} \mathbb{E}\left[\frac{1}{\mathbb{E}[L]} \mathbb{E}\left[\sum_{t=1}^{L} r(u_t)\right]\right] \\
&= \frac{1}{\mathbb{E}[L]} \mathbb{E}\left[\sum_{t=1}^{L} r(u_t)\right] \\
&\overset{\text{(iii)}}{=} \frac{1}{\mathbb{E}[L]} \mathbb{E}\left[\sum_{t=1}^{L} r(\bar{x}p(1-p)^{t-1})\right] \\
&= p \sum_{k=1}^{\infty} p(1-p)^{k-1} \sum_{t=1}^{k} r(\bar{x}p(1-p)^{t-1}) \\
&\overset{\text{(iv)}}{=} \sum_{i=1}^{\infty} p(1-p)^{i-1} r(\bar{x}p(1-p)^{i-1}) \qquad (14) \\
&\overset{\text{(v)}}{\geq} \sum_{i=1}^{\infty} p(1-p)^{i-1} \cdot (1-p)^{i-1} r(\bar{x}p) \\
&= r(\bar{x}p) \sum_{i=1}^{\infty} p(1-p)^{2i-2} \\
&= \frac{r(\bar{x}p)}{2-p} \\
&\overset{\text{(vi)}}{\geq} \frac{1}{2} r(\bar{x}p) \\
&= \frac{1}{2} r(\mathbb{E}[w_t]), \qquad (15)
\end{aligned}$$

where (i) is by Fatou's lemma [28, Theorem 1.5.4]; (ii) is due to (13); (iii) is by definition of the policy (6); (iv) is by changing the order of summations and evaluating the sum over k; (v) is because concavity of $r(u)$ along with the fact that $r(0) = 0$ imply $r(\alpha u) \geq \alpha r(u)$ for any $0 \leq \alpha \leq 1$; and (vi) is because $p \geq 0$. □

B. Additive Gap for Binary Disturbance: Proof of Proposition 4

By Assumption 1, we have the inequality

$$r(\bar{x}p(1-p)^{i-1}) \geq r(\bar{x}p) - \eta \log(1-p)^{-(i-1)}.$$

Substituting in (14) from the previous section:

$$\begin{aligned} J_\pi &\geq \sum_{i=1}^{\infty} p(1-p)^{i-1}\big(r(\bar{x}p) + (i-1)\eta\log(1-p)\big) \\ &= r(\bar{x}p) + \eta\frac{1-p}{p}\log(1-p) \\ &\geq r(\bar{x}p) - \eta\log e \\ &= r(\mathbb{E}[w_t]) - \eta\log e, \end{aligned} \tag{16}$$

where the last inequality is because $\frac{1-p}{p}\log(1-p)$ attains its minimum in the interval $[0,1]$ at $p=0$. □

C. General Disturbance: Proof of Theorem 2

We will now use the result of the previous sections to lower bound the expected average reward of the Fixed Fraction Policy for general i.i.d. disturbances. We will show that under all distributions of w_t with the same mean, the lowest expected average reward is obtained when w_t is a binary random variable, taking the values 0 or $\bar{x}$.

We begin with a few notations and definitions. Recall that the Fixed Fraction Policy is given by $\mu(x_t) = qx_t$, where $q = \mathbb{E}[w_t]/\bar{x}$. Under this policy:

$$x_{t+1} = \min\big((1-q)x_t + w_t,\, \bar{x}\big), \qquad t = 1,2,\ldots,$$

where, as in the previous sections, we assume $x_1 = \bar{x}$.

In what follows, we consider the performance of this policy under different disturbance distributions and different initial states. Therefore, we define the expected N-horizon total reward for initial state $x \in [0,\bar{x}]$ under the disturbance w_t:

$$J_N^\pi(x) \triangleq \sum_{t=1}^{N} \mathbb{E}[r(qx_t)\,|\,x_1 = x].$$

Note that the long-term expected average reward is given by $J_\pi = \liminf_{N\to\infty}\frac{1}{N}J_N^\pi(\bar{x})$.

Let $\hat{w}_t$ be i.i.d. binary random variables, specifically $\hat{w}_t \in \{0,\bar{x}\}$ and $\Pr(\hat{w}_t = \bar{x}) = q$. Note that

$$\mathbb{E}[\hat{w}_t] = \mathbb{E}[w_t].$$

Define the N-horizon total reward for initial state $x \in [0,\bar{x}]$ under the disturbance $\hat{w}_t$:

$$\hat{J}_N^\pi(x) \triangleq \sum_{t=1}^{N} \mathbb{E}[r(q\hat{x}_t)\,|\,\hat{x}_1 = x],$$

where $\hat{x}_t$ is the state evolved following the disturbance $\hat{w}_t$, that is

$$\hat{x}_{t+1} = \min\big((1-q)\hat{x}_t + \hat{w}_t,\, \bar{x}\big), \qquad t = 1,2,\ldots$$

In the following proposition, we claim that the N-horizon expected total reward for any disturbance distribution is always better than the expected total reward obtained for binary disturbance with the same mean, for any N and any initial state x.

Proposition 5. *For any $x \in [0,\bar{x}]$ and any integer $N \geq 1$:*

$$J_N^\pi(x) \geq \hat{J}_N^\pi(x).$$

In the proof of this proposition, we will make use of the following lemma from [1]:

Lemma 2. *Let $f(z)$ be a concave function defined on the interval $[0,\bar{x}]$, and let Z be a random variable confined to the same interval, i.e. $0 \leq Z \leq \bar{x}$. Let $\hat{Z} \in \{0,\bar{x}\}$ be a binary valued random variable with $\Pr(\hat{Z} = \bar{x}) = \mathbb{E}[Z]/\bar{x}$. Then*

$$\mathbb{E}[f(Z)] \geq \mathbb{E}[f(\hat{Z})].$$

Proof of Proposition 5. We will give a proof by induction. Clearly for $N=1$ we have

$$J_1^\pi(x) = \hat{J}_1^\pi(x) = r(qx).$$

Observe that this is a non-decreasing concave function of x. This will in fact be true for every $\hat{J}_N^\pi(x)$, $N \geq 1$, and we will use this in the induction step.

Assume that $J_{N-1}^\pi(x) \geq \hat{J}_{N-1}^\pi(x)$ for all $x \in [0,\bar{x}]$, and also that $\hat{J}_{N-1}^\pi(x)$ is monotone non-decreasing and concave in x.

For the induction step, observe that:

$$J_N^\pi(x) = r(qx) + \mathbb{E}[J_{N-1}^\pi(x_2)],$$

where the expectation is over the RV $x_2 = \min\{(1-q)x + w_2, \bar{x}\}$. This is due to the process x_t being a time-homogeneous Markov chain. By the induction hypothesis, we have:

$$J_N^\pi(x) \geq r(qx) + \mathbb{E}[\hat{J}_{N-1}^\pi(x_2)], \tag{17}$$

where still $x_2 = \min\big((1-q)x + w_2,\, \bar{x}\big)$. Now,

$$\begin{aligned} \hat{J}_{N-1}^\pi(x_2) &= \hat{J}_{N-1}^\pi\big(\min((1-q)x + w_2,\, \bar{x})\big) \\ &= \min\big(\hat{J}_{N-1}^\pi((1-q)x + w_2),\, \hat{J}_{N-1}^\pi(\bar{x})\big) \end{aligned}$$

where the second equality is because $\hat{J}_{N-1}^\pi(\,\cdot\,)$ is non-decreasing, due to the induction hypothesis. Next, we claim that the function $f_1(z) \triangleq \hat{J}_{N-1}^\pi((1-q)x + z)$ is concave. This is true again by the induction hypothesis that $\hat{J}_{N-1}^\pi(\,\cdot\,)$ is concave. Therefore, since $\hat{J}_{N-1}^\pi(\bar{x})$ is simply a constant, the function $f_2(z) \triangleq \hat{J}_{N-1}^\pi\big(\min((1-q)x + z,\, \bar{x})\big)$ is a minimum of two concave functions, hence it is itself concave. We can now apply Lemma 2 to obtain:

$$\begin{aligned} \mathbb{E}[\hat{J}_{N-1}^\pi(x_2)] &= \mathbb{E}[f_2(w_2)] \\ &\geq \mathbb{E}[f_2(\hat{w}_2)] \\ &= \mathbb{E}[\hat{J}_{N-1}^\pi(\hat{x}_2)], \end{aligned}$$

where $\hat{x}_2 = \min\big((1-q)x + \hat{w}_2,\, \bar{x}\big)$. Substituting this into (17):

$$\begin{aligned} J_N^\pi(x) &\geq r(qx) + \mathbb{E}[\hat{J}_{N-1}^\pi(\hat{x}_2)] \\ &= \hat{J}_N^\pi(x). \end{aligned}$$

It is left to verify that $\hat{J}_N^\pi(x)$ is concave and non-decreasing in x. Writing it explicitly:

$$\hat{J}_N^\pi(x) = r(qx) + q\hat{J}_{N-1}^\pi(\bar{x}) + (1-q)\hat{J}_{N-1}^\pi((1-q)x),$$

we see that it is a sum of non-decreasing concave functions of x, hence it is a non-decreasing concave function of x. □

As an immediate result of Proposition 5, we obtain

$$\liminf_{N\to\infty} \frac{1}{N} J_N^\pi(\bar{x}) \geq \liminf_{N\to\infty} \frac{1}{N} \hat{J}_N^\pi(\bar{x}). \tag{18}$$

Now we can apply the results of the previous section. From (15) we have

$$\begin{aligned}\liminf_{N\to\infty} \frac{1}{N} \hat{J}_N^\pi(\bar{x}) &\geq \frac{1}{2} r(\mathbb{E}[\hat{w}_t]) \\ &= \frac{1}{2} r(\mathbb{E}[w_t]),\end{aligned}$$

and if Assumption 1 holds we have from (16):

$$\begin{aligned}\liminf_{N\to\infty} \frac{1}{N} \hat{J}_N^\pi(\bar{x}) &\geq r(\mathbb{E}[\hat{w}_t]) - \eta \log e \\ &= r(\mathbb{E}[w_t]) - \eta \log e.\end{aligned}$$

Substituting in (18) completes the proof of Theorem 2.

VI. Conclusion

We proposed a policy for a general class of MDPs with concave reward function, and proved that it is within constant additive and multiplicative gaps to optimality for any disturbance distribution and any state space size $\bar{x}$. This allowed us to develop a simple and insightful approximation for the optimal expected average reward.

An important step in our proof was to show that binary disturbance constitute the worst case for our proposed policy among all i.i.d. disturbance processes with the same mean, i.e. the expected average reward achieved by our proposed policy is smallest when the process is a binary random variable. Whether i.i.d. binary disturbances are also the worst case in terms of the optimal average reward is an interesting question.

Appendix A
Optimal Policy for Binary Disturbance: Proof of Theorem 1

It can be argued [24, Theorem 6.4] that the there exists a stationary policy, i.e. there exists a function $\mu^*(x)$, satisfying $0 \leq \mu^*(x) \leq x$ for $0 \leq x \leq \bar{x}$, s.t. the optimal average reward is given by $J^* = \liminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^n \mathbb{E}[r(\mu^*(x_t))]$. Under such a stationary policy, the state x_t is a regenerative process (see Definition 1). The regeneration times $\{T(n)\}_{n=0}^\infty$ are the times in which $w_t = \bar{x}$, i.e. $w_{T(n)} = \bar{x}$, or equivalently $x_{T(n)} = \bar{x}$. Applying Lemma 1 from Section V-A, we obtain:

$$J^* = \frac{1}{\mathbb{E}L} \mathbb{E}\left[\sum_{t=1}^L r(mu^*(x_t))\right],$$

where $L = T(1) - T(0)$ is a Geometric(p) RV, which follows from the fact that w_t are i.i.d. Bernoulli(p). Observe that for $2 \leq t \leq L$ the disturbance is $w_t = 0$ by definition of L. Hence, we have the following deterministic recursive relation:

$$\begin{aligned} &x_1 = \bar{x}, \\ &x_t = x_{t-1} - \mu^*(x_{t-1}) \quad , t = 2, \ldots, L. \end{aligned} \tag{19}$$

Since L can take any positive integer, this defines a sequence $\{\gamma_i^*\}_{i=1}^\infty$ such that $\mu^*(x_i) = \gamma_i^*$. We can therefore write

$$\begin{aligned} J^* &= \frac{1}{\mathbb{E}L} \mathbb{E}\left[\sum_{i=1}^L r(\gamma_i^*)\right] \\ &= p \sum_{k=1}^\infty p(1-p)^{k-1} \sum_{i=1}^k r(\gamma_i^*) \\ &= \sum_{i=1}^\infty p(1-p)^{i-1} r(\gamma_i^*). \end{aligned}$$

Moreover, by the constraint $\mu^*(x_t) \leq x_t$ and the recursive relation (19), we must have $\sum_{i=1}^\infty \gamma_i^* \leq \bar{x}$, in addition to $\gamma_i^* \geq 0$ for all $i \geq 1$.

To find $\{\gamma_i^\star\}_{i=1}^\infty$ we need to solve the following infinite-dimensional optimization problem:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^\infty p(1-p)^{i-1} r(\gamma_i) \\ \text{subject to} \quad & \gamma_i \geq 0, \quad i = 1, 2, \ldots, \\ & \sum_{i=1}^\infty \gamma_i \leq \bar{x}. \end{aligned} \tag{20}$$

Let $\{\gamma_i^*\}_{i=1}^\infty$ and J^* be the optimal sequence and optimal objective, respectively, of (20). We will show that (20) can be solved by the limit as $N \to \infty$ of the following N-dimensional optimization problem:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^N p(1-p)^{i-1} r(\gamma_i) \\ \text{subject to} \quad & \gamma_i \geq 0, \quad i = 1, 2, \ldots, N, \\ & \sum_{i=1}^N \gamma_i \leq \bar{x}. \end{aligned} \tag{21}$$

Denote by J_N the optimal objective of (21). Clearly J_N is non-decreasing and $J_N \leq J^*$. Observe that the first N values of the infinite-dimensional solution, $\{\gamma_i^*\}_{i=1}^N$, are a feasible solution for (21). Therefore,

$$\begin{aligned} J_N &\geq \sum_{i=1}^N p(1-p)^{i-1} r(\gamma_i^*) \\ &= J^* - \sum_{i=N+1}^\infty p(1-p)^{i-1} r(\gamma_i^*) \\ &\overset{(*)}{\geq} J^* - \sum_{i=N+1}^\infty p(1-p)^{i-1} r(\bar{x}) \\ &= J^* - (1-p)^N r(\bar{x}), \end{aligned}$$

where $(*)$ is because $\gamma_i^* \leq \bar{x}$ and $r(u)$ is non-decreasing. Along with the inequality $J_N \leq J^*$, this implies

$$J^* = \lim_{N\to\infty} J_N.$$

We continue with the explicit solution of (21). We write the Lagrangian and solve using KKT conditions:

$$\mathscr{L} = \sum_{i=1}^{N} p(1-p)^{i-1} r(\gamma_i) + \sum_{i=1}^{N} \lambda_i \gamma_i - \nu \left(\sum_{i=1}^{N} \gamma_i - \bar{x} \right).$$

Taking derivative:

$$\frac{\partial \mathscr{L}}{\partial \gamma_i} = p(1-p)^{i-1} r'(\gamma_i) + \lambda_i - \nu = 0,$$

along with complementary slackness conditions $\lambda_i \gamma_i = 0$ and $\nu(\sum_{i=1}^{N} \gamma_i - \bar{B}) = 0$. For non-zero γ_i we have $\lambda_i = 0$, which gives $\nu = p(1-p)^{i-1} r'(\gamma_i)$ for all i. The reward function $r(u)$ is non-decreasing and strictly concave, hence $r'(u)$ is strictly decreasing and therefore invertible. We get

$$\gamma_i = (r')^{-1} \left(\frac{\nu}{p(1-p)^{i-1}} \right).$$

Since $(r')^{-1}(\,\cdot\,)$ is decreasing, the sequence γ_i is decreasing. Applying the function $r'(\,\cdot\,)$ on both sides of the inequality $\gamma_i \geq 0$, using the monotonicity of $r'(u)$:

$$r'(\gamma_i) \leq r'(0),$$
$$\frac{\nu}{p(1-p)^{i-1}} \leq r'(0),$$
$$i \leq 1 + \frac{\log \nu - \log(pr'(0))}{\log(1-p)},$$

where the last inequality holds for all i for which $\gamma_i > 0$. Denoting $\tilde{N} = \min \left(\left\lfloor 1 + \frac{\log \nu - \log(pU'(0))}{\log(1-p)} \right\rfloor,\, N \right)$, we get that $\gamma_i > 0$ for $i = 1, \ldots, \tilde{N}$ and $\gamma_i = 0$ otherwise. Along with the fact that the second constraint must hold with equality, i.e. $\sum_{i=1}^{\infty} \gamma_i = \bar{x}$, since increasing γ_i can only increase the reward, we obtain the following equation for ν:

$$\bar{x} = \sum_{i=1}^{\tilde{N}} (r')^{-1} \left(\frac{\nu}{p(1-p)^{i-1}} \right).$$

Taking $N \to \infty$ completes the proof. □

References

[1] D. Shaviv and A. Özgür, "Universally near optimal online power control for energy harvesting nodes," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3620–3631, Dec. 2016.

[2] J. Yang and S. Ulukus, "Optimal packet scheduling in an energy harvesting communication system," *IEEE Trans. Commun.*, vol. 60, no. 1, pp. 220–230, 2012.

[3] K. Tutuncuoglu and A. Yener, "Optimum transmission policies for battery limited energy harvesting nodes," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 1180–1189, 2012.

[4] O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, and A. Yener, "Transmission with energy harvesting nodes in fading wireless channels: Optimal policies," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1732–1743, 2011.

[5] M. Zafer and E. Modiano, "Optimal rate control for delay-constrained data transmission over a wireless channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 9, pp. 4020–4039, 2008.

[6] C. K. Ho and R. Zhang, "Optimal energy allocation for wireless communications powered by energy harvesters," in *IEEE Int. Symp. Information Theory (ISIT)*, 2010, pp. 2368–2372.

[7] A. Sinha and P. Chaporkar, "Optimal power allocation for a renewable energy source," in *National Conf. Commun. (NCC)*. IEEE, 2012, pp. 1–5.

[8] P. Blasco, D. Gunduz, and M. Dohler, "A learning theoretic approach to energy harvesting communication system optimization," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1872–1882, 2013.

[9] C. K. Ho and R. Zhang, "Optimal energy allocation for wireless communications with energy harvesting constraints," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4808–4818, 2012.

[10] S. Mao, M. H. Cheung, and V. W. Wong, "An optimal energy allocation algorithm for energy harvesting wireless sensor networks," in *IEEE Int. Conf. Commun. (ICC)*, 2012, pp. 265–270.

[11] X. Wang, J. Gong, C. Hu, S. Zhou, and Z. Niu, "Optimal power allocation on discrete energy harvesting model," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, pp. 1–14, 2015.

[12] A. Kazerouni and A. Özgür, "Optimal online strategies for an energy harvesting system with Bernoulli energy recharges," in *13th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2015, pp. 235–242.

[13] Y. Dong, F. Farnia, and A. Özgür, "Near optimal energy control and approximate capacity of energy harvesting communication," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 540–557, 2015.

[14] M. B. Khuzani and P. Mitran, "On online energy harvesting in multiple access communication systems," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1883–1898, 2014.

[15] C. M. Vigorito, D. Ganesan, and A. G. Barto, "Adaptive control of duty cycling in energy-harvesting wireless sensor networks," in *4th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON'07)*, 2007, pp. 21–30.

[16] V. Sharma, U. Mukherji, V. Joseph, and S. Gupta, "Optimal energy management policies for energy harvesting sensor nodes," *IEEE Trans. Wireless Commun.*, vol. 9, no. 4, pp. 1326–1336, 2010.

[17] R. Rajesh, V. Sharma, and P. Viswanath, "Capacity of fading Gaussian channel with an energy harvesting sensor node," in *IEEE Global Telecommunications Conference (GLOBECOM 2011)*, 2011, pp. 1–6.

[18] R. Srivastava and C. E. Koksal, "Basic performance limits and tradeoffs in energy-harvesting sensor nodes with finite data and energy storage," *IEEE/ACM Transactions on Networking (TON)*, vol. 21, no. 4, pp. 1049–1062, 2013.

[19] Q. Wang and M. Liu, "When simplicity meets optimality: Efficient transmission power control with stochastic energy harvesting," in *Proc. IEEE INFOCOM*, 2013, pp. 580–584.

[20] F. Amirnavaei and M. Dong, "Online power control strategy for wireless transmission with energy harvesting," in *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2015, pp. 6–10.

[21] J. Xu and R. Zhang, "Throughput optimal policies for energy harvesting wireless transmitters with non-ideal circuit power," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 2, pp. 322–332, 2014.

[22] B. T. Bacinoglu, E. Uysal-Biyikoglu, and C. E. Koksal. (2017) Finite horizon energy-efficient scheduling with energy harvesting transmitters over fading channels. [Online]. Available: https://arxiv.org/abs/1702.06390

[23] S. Satpathi, R. Nagda, and R. Vaze, "Optimal offline and competitive online strategies for transmitter–receiver energy harvesting," *IEEE Trans. Inf. Theory*, vol. 62, no. 8, pp. 4674–4695, 2016.

[24] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus, "Discrete-time controlled Markov processes with average cost criterion: a survey," *SIAM J. Control Optim.*, vol. 31, no. 2, pp. 282–344, 1993.

[25] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 2nd ed. Athena Scientific, 2001, vol. 2.

[26] S. Asmussen, *Applied probability and queues*. Springer Science & Business Media, 2008, vol. 51.

[27] S. M. Ross, *Introduction to probability models*. Academic press, 2014.

[28] R. Durrett, *Probability: theory and examples*. Cambridge university press, 2010.