Online Decision Making under Stochastic Constraints

Mehrdad Mahdavi
Dept. of Computer Science
Michigan State University
mahdavim@cse.msu.edu

Tianbao Yang
Machine Learning Lab
GE Global Research
tyang@ge.com

Rong Jin
Dept. of Computer Science
Michigan State University
rongjin@cse.msu.edu

Abstract

This paper proposes a novel algorithm for solving discrete online learning problems under stochastic constraints, where the leaner aims to maximize the cumulative reward given that some additional constraints on the sequence of decisions need to be satisfied on average. We propose Lagrangian exponentially weighted average (LEWA) algorithm, which is a primal-dual variant of the well known exponentially weighted average algorithm, and inspired by the theory of Lagrangian method in constrained optimization. We establish expected and high probability bounds on the regret and the violation of the constraint in full information and bandit feedback models for LEWA algorithm.

1 Introduction

Many practical problems such as online portfolio management [1], prediction from expert advice [2] [3], and online shortest path problem [4] involve making repeated decisions in an unknown and unpredictable environment (see, e.g., [5] for a comprehensive review). These situations can be formulated as a repeated game between the decision maker (i.e., the learner) and the adversary (i.e., the environment). At each round of the game, the learner selects an action from a fixed set of actions and then receives feedback (i.e., reward) for the selected action. The analysis of online learning algorithms focuses on establishing sub-linear bounds on the *regret* that is the difference between the reward of the best fixed action with the hindsight knowledge of the observed sequence and the cumulative reward of the learner.

In many current literature, the application of online learning is mostly limited to problems without constraints on the sequence of decisions made by the learner. However, in most scenarios, beyond maximizing the cumulative reward, there are some restrictions on the decisions that need to be satisfied on average. Therefore, one might desire algorithms for a much more ambitious framework, where we need to maximize total reward under the constraints. As an illustrative example, let us consider a wireless communication system where the agent chooses an appropriate transmission power in order to transmit a message successfully. In this case, the goal of the agent may be to maximize average throughput, while keeping the average power consumption under some required threshold. Attempts for such extension were made in [6], where the online learning with path constraints has been addressed and algorithms with *asymptotically* vanishing bound have been proposed.

An algorithm addressing this problem has to balance between maximizing the adversary rewards and satisfying the constraint. If the algorithm be too aggressive to satisfy the constraint, then there would be less hope to attain satisfactory cumulative reward at the end of the game and on the other hand, just trying to maximize the cumulative reward will end up in a situation in which the constraints vanish linearly in terms of the number of rounds. To affirmatively address the problem, we provide a general framework for repeated games with constraint, and propose a simple randomized algorithm called **Lagrangian exponentially weighted average (LEWA)** algorithm for a particular class of these games. The proposed formulation is inspired by the theory of Lagrangian method in constrained optimization and is based on primal-dual formulation of the exponentially weighted

average (EWA) algorithm [3] [7]. We establish expected and high probability bounds on the regret and the violation of the constraints on average for LEWA algorithm, and extend the results to the bandit setting where only partial feedback about the rewards and the constraint are available. To the best of our knowledge, this is the first time that a Lagrangian style relaxation has been proposed for this type of problem.

Notations. Let us introduce some notations used in this paper. Vectors are indicated in lower case bold letters such as $\mathbf{x}$ where $\mathbf{x}^\top$ denotes it transpose. By default, all vectors are column vectors. For a vector $\mathbf{x}$, x_i denotes its ith coordinate. We use superscripts to index rounds of the game. Component-wise multiplication between vectors is denoted by $\circ$. We use $[K]$ as a shorthand for the set of integers $\{1, 2, \ldots, K\}$. Throughout the paper we denote by $[\cdot]_+$ the projection onto the positive orthant. We shall use $\mathbf{1}$ to denote the vector of all ones. Finally, for a K-dimensional vector $\mathbf{x}$, $(\mathbf{x})^2$ represents $(x_1^2, \ldots, x_K^2)$.

2 Statement of the Problem

We consider the general decision-theoretic framework for online learning and extend it to capture the constraints. In original online decision making, the learner is given access to a pool of K actions. In each round $t \in [T]$, the learner chooses a probability distribution $\mathbf{p}_t = (p_1^t, ..., p_K^t)$ over the actions $[K]$ and chooses an action i randomly based on $\mathbf{p}_t$. In the scenario of full information feedback model, at each iteration, the adversary reveals a reward vector $\mathbf{r}_t = (r_1^t, \cdots, r_K^t)$. Choosing an action i results in receiving a reward r_i^t, which we shall assume without loss of generality to be bounded in $[0, 1]$. In the partial feedback model or bandit setting only the cost of selected action is revealed by the adversary. The learner competes with the best fixed action in hindsight and his/her goal is to minimize the regret defined as $\max_{\mathbf{p}} \sum_t \mathbf{p}^\top \mathbf{r}_t - \sum_t \mathbf{p}_t^\top \mathbf{r}_t$. This problem is a well studied problem and there are algorithms which attain an optimal regret bound of $O(\sqrt{T \ln K})$ after T rounds of the game. In this paper we focus on the exponentially weighted average (EWA) algorithm, which will be used later as the baseline of the proposed algorithm. The EWA algorithm maintains a weight vector $\mathbf{w}_t = (w_1^t, \cdots, w_K^t)$ which is used to define the probabilities over actions. After receiving the reward vector $\mathbf{r}_t$ at round t, the EWA algorithm updates the weight vector according to $w_i^{t+1} = w_i^t \exp(\eta r_i^t)$ where η is the learning rate.

In the new setting addressed in this paper, which we refer to as *constrained regret minimization*, in addition to the rewards, there exist some constraints on the decisions that need to be satisfied. In particular, for the decision $\mathbf{p}$ made by the learner, there is an additional constraint $\mathbf{p}^\top \mathbf{c} \geq c_0$ where $\mathbf{c}$ is a constraint vector for specifying the constraint. We note that, in general, the reward vector $\mathbf{r}_t$ and the constraint vector $\mathbf{c}$ are different and can not be combined as a single objective. The learner's goal is to maximize the total reward with respect to the optimal decision in hindsight under the constraint $\mathbf{p}^\top \mathbf{c} \geq c_0$, i.e., $\min_{\mathbf{p}_1, \ldots, \mathbf{p}_T} \max_{\mathbf{p}^\top \mathbf{c} \geq c_0} \sum_{t=1}^T \mathbf{p}^\top \mathbf{r}_t - \sum_{t=1}^T \mathbf{p}_t^\top \mathbf{r}_t$, and simultaneously satisfy the constraint. Note that the comparator class includes fixed decision $\mathbf{p}$ that attains maximal cumulative reward had he known the rewards beforehand, while satisfying the additional constraint.

Within our setting, we consider repeated games with *adversarial* rewards and *stochastic* constraint. More precisely, let $\mathbf{c} = (c_1, \cdots, c_K)$ be the constraint vector defined over actions. In stochastic setting the vector $\mathbf{c}$ is **unknown** to the learner and at each round $t \in [T]$, beyond the reward feedback, the learner receives a random realization $\mathbf{c}_t = (c_1^t, \cdots, c_K^t)$ of the constraint vector $\mathbf{c}$ where $\mathrm{E}[c_i^t] = c_i$. The learner's goal is to choose a sequence of decisions $\mathbf{p}_t, t \in [T]$ to minimize the regret with respect to the optimal decision in hindsight under the constraint $\mathbf{p}^\top \mathbf{c} \geq c_0$. Without loss of generality we assume $\mathbf{c}_t \in [0, 1]^K$ and $c_0 \in [0, 1]$. Formally, the goal of the learner is to attain a gradually vanishing regret as

$$\text{Regret}_T = \max_{\mathbf{p}^\top \mathbf{c} \geq c_0} \sum_t \mathbf{p}^\top \mathbf{r}_t - \sum_t \mathbf{p}_t^\top \mathbf{r}_t \leq O(T^{1-\beta_1}). \tag{1}$$

Furthermore, the decisions $\mathbf{p}_t, t = 1, \cdots, T$ made by the learner are required to attain sub-linear bound on the violation of the constraint in the long run, i.e.,

$$\text{Violation}_T = \left[\sum_{t=1}^T \left(c_0 - \mathbf{p}_t^\top \mathbf{c} \right) \right]_+ \leq O(T^{1-\beta_2}). \tag{2}$$

We refer to the above bound as the *violation of the constraint*. The two questions we seek to answer are how to modify EWA algorithm to take the constraint under consideration and what would be the bounds on the regret as well as the violation of the constraint attainable by the modified algorithm.

Related Works. There is a rich body of literature that deals with the online decision making problem without constraints and there exist a number of online algorithms that have the optimal regret bound. The most well-known and successful work is probably the Hedge algorithm [7], which was a direct generalization of Littlestone and Warmuth's Weighted Majority (WM) algorithm [3]. Other recent studies include the improved theoretical bounds and the parameter-free hedging algorithm [8] and adaptive Hedge [9] for decision-theoretic online learning. We refer readers to the [5] for an in-depth discussion of this subject.

As the first seminal paper in adversarial constrained decision making, Mannor et al. [6] introduced the online learning with simple path constraints. They considered the infinitely repeated two player games with stochastic rewards where for every joint action of the players, there is an additional stochastic constraint vector that is accumulated by the decision maker. We note that the analysis in [6] is asymptotic while the bounds to be established in this work are applicable to finite repeated games. In [10] the budget limited MAB was introduced where polling an arm is costly where the cost of each arm is fixed in advance. In this setting both the exploration and exploitation phases are limited by a global budget. This setting matches the stochastic rewards with deterministic constraints without violation game discussed before. It has been shown that existing MAB algorithms are not suitable to efficiently deal with costly arms. They proposed the $\epsilon - first$ algorithm that dedicates the first ϵ fraction of the total budget exclusively for exploration and the remaining $(1-\epsilon)$ fraction for exploitation. [11] improves the bound obtained in [10] by proposing a Knapsack based UCB [12] algorithm which extends the UCB algorithm by solving a Knapsack problem at each round to cope with the constraints. We note that Knapsack based UCB does not make explicit distinction between exploration and exploitation steps as done in $\epsilon - first$ algorithm. In both [11] and [10] the algorithm proceeds as long as sufficient budget existing to play the arms.

3 Full Information Constrained Regret Minimization

A straightforward approach to tackle the problem is to modify the reward functions of the learner to include constraint term with a penalty coefficient that reduces the reward when the constraint is violated. This approach circumvents the problem of a constrained online learning by turning it into an unconstrained problem, but a simple analysis shows that, in the adversarial setting, this simple penalty based approach fails to attain gradually vanishing bounds for regret and the violation of constraints. The main difficulty arises from the fact that an adaptive adversary can play with the penalty coefficient associated with constraint in order to weaken the influence of the penalty parameter which results in linear bound on at least one of the measures, i.e. either regret bound or violation of the constraints.

Alternatively, since the constraint in our setting is stochastic, one possible solution is to take an exploration and exploitation scheme, i.e., to burn a small portion ϵ of the rounds to estimate the constraint vector $\mathbf{c}$ by $\widetilde{\mathbf{c}}$ and then in the remaining $(1-\epsilon)T$ rounds follow the existing algorithms with restricted decisions, i.e., $\mathbf{p} \in \Delta_K \cap \mathbf{p}^\top \widetilde{\mathbf{c}} \geq c_0$, where Δ_K is the simplex over $[K]$. The parameter ϵ balances the accuracy of estimating $\mathbf{c}$ and the number of rounds for exploitation to increase the total reward. One may hope that by careful adjustment of ϵ, it would be possible to get satisfactory bounds on regret and the violation of the constraint. But unfortunately this naive approach suffers from two main drawbacks. First, the number of rounds T is not known in advance. Second, the decisions are made by projecting into an estimated domain $\mathbf{p}^\top \widetilde{\mathbf{c}} \geq c_0$ instead of the true domain $\mathbf{p}^\top \mathbf{c} \geq c_0$ which is problematic as follows. In order to show the regret bound, we need to relate the best cumulative reward in the estimated domain to that in the true domain, which however requires imposing a regularity condition on reward and constrain vectors to be solvable [13]. Basically, we can make the algorithm adaptive to T by using a similar idea to *epoch greedy* [14] algorithm that runs exploration/exploitation in epochs, but it still suffers from the second drawback. Additionally, projection to the inaccurate estimated constraint $\widetilde{\mathbf{c}}$ does not exclude the possibility that the solution will be infeasible.

Here, we take a different path to solve the problem. The proposed formulation is inspired by the theory of Lagrangian method in constrained optimization. The intuition behind the proposed al-

LEWA (η and δ)
initialize: $\mathbf{w}_1 = \mathbf{1}$ and $\lambda_1 = 0$
iterate $t = 1, 2, \ldots, T$
Draw an action accordingly to the probability $\mathbf{p}_t = \dfrac{\mathbf{w}_t}{\sum_j w_j^t}$
Receive reward $\mathbf{r}_t$ and a realization of constraint $\mathbf{c}_t$
Update $\mathbf{w}_{t+1} = \mathbf{w}_t \circ \exp(\eta(\mathbf{r}_t + \lambda_t \mathbf{c}_t))$
Update $\lambda_{t+1} = [(1 - \delta\eta)\lambda_t - \eta(\mathbf{p}_t^\top \mathbf{c}_t - c_0)]_+$
end iterate

gorithm is to optimize one criterion (i.e., minimizing regret or maximizing the reward) subject to explicit constraint on the restrictions that the learner needs to satisfy in average for the sequence of the decisions. A challenging ingredient in this formulation is that of establishing bounds on the regret and the violation of the constraints. In particular, our algorithms will exhibit a bound in the following structure,

$$\text{Regret}_T + \frac{\text{Violation}_T^2}{O(T^{1-\alpha})} \leq O(T^{1-\beta}), \tag{3}$$

where Violation_T is a term related to the violation of constraint in long term. From (3) we can derive a bound for regret and the violation of constraints as

$$\text{Regret}_T \leq O(T^{1-\beta}) \tag{4}$$

$$\text{Violation}_T \leq \sqrt{O\left([T + T^{1-\beta}]T^{1-\alpha}\right)}, \tag{5}$$

where the last bound follows the fact $-\text{Regret}_T \leq O(T)$.

The detailed steps of the proposed algorithm are shown in LEWA. The algorithm keeps two set of variables: the weight vector $\mathbf{w}_t$ and the Lagrangian multiplier λ_t. The high level interpretation of the algorithm is as follows: if the constraint is being violated a lot, the decision maker places more weight on the constraint controlled by λ_t; but he tunes down the weight on the constraint when the constraint is satisfied reasonably. We note the LEWA is equivalent to the original EWA when the constraint is satisfied at each iteration, i.e., $\mathbf{p}_t^\top \mathbf{c}_t \geq c_0$, which gives $\lambda_1 = \cdots = \lambda_t = \ldots = 0$. It should be emphasized that in some previous works such as [10], the learner is not allowed to exceed the pre-specified threshold for the violation of the constraints and the game stops as soon as the learner violates the constraint. In contrast, within our setting similar to [15], the learner's goal is to obtain sub-linear bound on the long term violation of the constraint.

We now state the main theorem about the performance of LEWA algorithm.

Theorem 1. *Let $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_T$ be the sequence of randomized decisions over the set of actions $[K] := \{1, 2, \cdots, K\}$ produced by LEWA algorithm under the sequence of adversarial rewards $\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_T \in [0, 1]^K$ observed for these decisions. Let $\lambda_1, \lambda_2, \cdots, \lambda_T$ be the corresponding dual sequence. By setting $\eta = \sqrt{4 \ln K/(9T)}$ and $\delta = \eta/2$ we have:*

$$\max_{\mathbf{p}^\top \mathbf{c} \geq c_0} \sum_{t=1}^T \mathbf{p}^\top \mathbf{r}_t - \mathrm{E}\left[\sum_{t=1}^T \mathbf{p}_t^\top \mathbf{r}_t\right] \leq 3\sqrt{T \ln K}, \text{ and } \mathrm{E}\left[\sum_{t=1}^T (c_0 - \mathbf{p}_t^\top \mathbf{c})\right]_+ \leq O(T^{3/4}),$$

where expectation is taken over randomness in $\mathbf{c}_1, \cdots, \mathbf{c}_T$.

Remark 2. *From Theorem 1 we see that the LEWA algorithm attains the optimal bound for the regret and an $O(T^{3/4})$ bound on the violation of the constraint. We note that when deriving the bound for Violation_T, we simply use a weak lower bound on regret as $\text{Regret}_T \geq -T$. It is possible to obtain an improved bound by considering tighter bound for the Regret_T. One way to do this is to bound the regret by the variation of the reward vectors as $\text{Variation}_T = \sum_{t=1}^T \|\mathbf{r}_t - \widehat{\mathbf{r}}_T\|_\infty$, where $\widehat{\mathbf{r}}_T = (1/T)\sum_{t=1}^T \mathbf{r}_t$ denotes the mean of $\mathbf{r}_t, t \in [T]$. As shown in the full-length version of this paper [16], we can bound the violation of the constraints in terms of Variation_T as*

$$\left[\sum_{t=1}^T (c_0 - \mathbf{x}_t^\top \mathbf{c})\right]_+ \leq O(\sqrt{T}) + O(T^{1/4}\sqrt{\text{Variation}_T}).$$

High Probability LEWA (η, δ and ϵ)
initialize: $\mathbf{w}_1 = \mathbf{1}$ and $\lambda_1 = 0$
iterate $t = 1, 2, \ldots, T$
Draw an action accordingly to the probability $\mathbf{p}_t = \dfrac{\mathbf{w}_t}{\sum_j w_j^t}$.
Receive reward $\mathbf{r}_t$ and a realization of constraint $\mathbf{c}_t$
Compute average constraint estimate $\overline{\mathbf{c}}_t = \dfrac{1}{t}\sum_{s=1}^{t} \mathbf{c}_s$
Update $\mathbf{w}_{t+1} = \mathbf{w}_t \circ \exp(\eta(\mathbf{r}_t + \lambda_t \overline{\mathbf{c}}_t))$
Update $\lambda_{t+1} = [(1-\delta\eta)\lambda_t - \eta(\mathbf{p}_t^\top \overline{\mathbf{c}}_t + \alpha_t - c_0)]_+$.
end iterate

This bound is significantly better when the variation of the reward vectors is small and in worst case it attains an $O(T^{3/4})$ bound as Theorem 1.

With a simple trick, we are able to modify the LEWA algorithm to attain high probability bounds on regret and the violation of the constraint in the same order as in expectation. To this end, we slightly change the original LEWA algorithm and instead of using $\mathbf{c}_t$ in updating λ_{t+1}, we use the average estimate and add a confidence bound to achieve a more accurate estimation of the constraint vector $\mathbf{c}$. The following theorem bounds the regret and the violation of the constrain in high probability for the modified algorithm.

Theorem 3. *Let $\alpha_t = \frac{1}{\sqrt{t}}\sqrt{(1/2)\ln(2/\epsilon)}$, $\eta = O(T^{-1/2})$, and $\delta = \eta/2$. By running LEWA, we have, with probability $1-\epsilon$*

$$\max_{\mathbf{p}^\top \mathbf{c} \geq c_0} \sum_{t=1}^{T} \mathbf{p}^\top \mathbf{r}_t - \sum_{t=1}^{T} \mathbf{p}_t^\top \mathbf{r}_t \leq \widetilde{O}(T^{1/2}) \text{ and } \left[\sum_{t=1}^{T}(c_0 - \mathbf{p}_t^\top \mathbf{c})\right]_+ \leq O(T^{3/4})$$

where $\widetilde{O}(\cdot)$ omits the log *term in T.*

4 Bandit Constrained Regret Minimization

In this section, we generalize our results to the bandit setting for both rewards and constraints. In the bandit setting, at each iteration, we are required to choose an action i_t from the pool of the actions $[K]$. Then only the reward and constraint feedback for i_t is revealed to the learner, i.e. $r_{i_t}^t, c_{i_t}^t$. In this case, we are interested in the regret bound as $\max_{\mathbf{p}^\top \mathbf{c} \geq c_0} \sum_{t=1}^{T} \mathbf{p}^\top \mathbf{r}_t - \sum_{t=1}^{T} r_{i_t}^t$. In the classical setting, i.e., without constraint, this problem can be solved in stochastic and adversarial settings by UCB and EXP3 algorithms proposed in [17] and [12], respectively. The algorithm is shown in BanditLEWA algorithm which uses the similar idea to EXP3 for exploration and exploitation. Before presenting the performance bound of the algorithm, let us introduce two vectors: $\widehat{\mathbf{r}}_t$ is all zero vector except in i_tth component which is set to be $\widehat{r}_{i_t}^t = r_{i_t}^t / p_{i_t}^t$ and similarly $\widehat{\mathbf{c}}_t$ is all zero vector except in i_tth component which is set to be $\widehat{c}_{i_t}^t = c_{i_t}^t / p_{i_t}^t$. It is easy to verify that $\mathrm{E}_{i_t}[\widehat{\mathbf{r}}_t] = \mathbf{r}_t$ and $\mathrm{E}_{i_t}[\widehat{\mathbf{c}}_t] = \mathbf{c}_t$.

BanditLEWA (η, γ, and δ)
initialize: $\mathbf{w}_1 = \mathbf{1}$ and $\lambda_1 = 0$
iterate $t = 1, 2, \ldots, T$
Set $\mathbf{q}_t = \dfrac{\mathbf{w}_t}{\sum_j w_j^t}$
Draw action i_t randomly accordingly to $\mathbf{p}_t = (1-\gamma)\mathbf{q}_t + \gamma\dfrac{\mathbf{1}}{K}$
Receive reward $r_{i_t}^t$ and a realization of constraint $c_{i_t}^t$ for action i_t
Update $w_i^{t+1} = w_i^t \exp(\eta(\widehat{r}_i^t + \lambda_t \widehat{c}_i^t))$
Update $\lambda_{t+1} = [(1-\delta\eta)\lambda_t - \eta(\mathbf{q}_t^\top \widehat{\mathbf{c}}_t - c_0)]_+$
end iterate

The following theorem shows that BanditLEWA algorithm achieves $O(T^{3/4})$ regret bound and $O(T^{3/4})$ bound on the violation of the constraints in expectation.

Theorem 4. *Let* $\gamma = O(T^{-1/2}), \eta = \frac{\gamma}{K}\frac{\delta}{\delta+1}$, *by running BanditLEWA algorithm, we have*

$$\max_{\mathbf{p}^\top \mathbf{c} \geq c_0} \sum_{t=1}^{T} \mathbf{p}^\top \mathbf{r}_t - \mathrm{E}\left[\sum_{t=1}^{T} r_{i_t}^t\right] \leq O(T^{3/4}) \;\; \textit{and} \;\; \mathrm{E}\left[\sum_{t=1}^{T}(c_0 - \mathbf{p}_t^\top \mathbf{c})\right]_+ \leq O(T^{3/4}).$$

5 Conclusions and Future Works

In this extended abstract we propose an efficient algorithm for regret minimization under stochastic constraints. The proposed algorithm that is called LEWA, is a primal dual variant of the exponentially weighted average algorithm. We establish expected and high probability bounds on the regret and the long term violation of the constraint in full information and bandit settings. In particular, in full information setting, LEWA algorithms attains optimal $\tilde{O}(\sqrt{T})$ regret bound and $O(T^{3/4})$ bound on the violation of the constraints in expectation, and with a simple trick in high probability. The present work leaves open a number of interesting directions for future work. In particular, extending the framework to handle multi-criteria online decision making is left to future work. Turning the proposed algorithm to the one which exactly satisfies the constraint in the long run is also an interesting problem. Finally, it would be interesting to see if it is possible to improve the bound obtained for the violation of the constraint.

References

[1] E. Hazan, S. Kale, On stochastic and worst-case models for investing, in: NIPS, 2009, pp. 709–717.

[2] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, M. K. Warmuth, How to use expert advice, J. ACM 44 (3) (1997) 427–485.

[3] N. Littlestone, M. K. Warmuth, The weighted majority algorithm, Inf. Comput. 108 (2) (1994) 212–261.

[4] E. Takimoto, M. K. Warmuth, Path kernels and multiplicative updates, Journal Machine Learnning Research 4 (2003) 773–818.

[5] N. Cesa-Bianchi, G. Lugosi, Prediction, Learning, and Games, Cambridge University Press, 2006.

[6] S. Mannor, J. N. Tsitsiklis, J. Y. Yu, Online learning with sample path constraints, Journal of Machine Learning Research 10 (2009) 569–590.

[7] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139.

[8] K. Chaudhuri, Y. Freund, D. Hsu, A parameter-free hedging algorithm, in: NIPS, 2009, pp. 297–305.

[9] T. van Erven, P. Grunwald, W. M. Koolen, S. de Rooij, Adaptive hedge, in: NIPS, 2011, pp. 1656–1664.

[10] L. Tran-Thanh, A. C. Chapman, E. M. de Cote, A. Rogers, N. R. Jennings, Epsilon-first policies for budget-limited multi-armed bandits, in: AAAI, 2010.

[11] L. Tran-Thanh, A. C. Chapman, A. Rogers, N. R. Jennings, Knapsack based optimal policies for budget-limited multi-armed bandits, in: AAAI, 2012.

[12] P. Auer, N. Cesa-Bianchi, P. Fischer, Finite-time analysis of the multiarmed bandit problem, Machine Learning 47 (2-3) (2002) 235–256.

[13] S. M. Robinson, A characterization of stability in linear programming, Operations Research 25 (3) (1977) 435–447.

[14] J. Langford, T. Zhang, The epoch-greedy algorithm for multi-armed bandits with side information, in: NIPS, 2007.

[15] M. Mahdavi, R. Jin, T. Yang, Trading regret for efficiency: online convex optimization with long term constraints, JMLR 13 (2012) 2465–2490.

[16] M. Mahdavi, T. Yang, R. Jin, Efficient constrained regret minimization, CoRR abs/1205.2265.

[17] P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire, The nonstochastic multiarmed bandit problem, SIAM J. Comput. 32 (1) (2002) 48–77.

[18] M. Zinkevich, Online convex programming and generalized infinitesimal gradient ascent, in: ICML, 2003, pp. 928–936.

Appendix A. Proof of Theorem 1

In order to prove Theorem 1, we state two lemmas that pave the way to the proof of theorem.

Lemma 5. [Primal Inequality] *Let* $\mathbf{R}_t = \mathbf{R}_t^1 + \lambda_t \mathbf{R}_t^2$, *where* $\mathbf{R}_t^1, \mathbf{R}_t^2 \in \mathbb{R}_+^K$, $\mathbf{w}_{t+1} = \mathbf{w}_t \circ \exp(\eta \mathbf{R}_t)$, *and* $\mathbf{p}_t = \mathbf{w}_t / \mathbf{w}_t^\top \mathbf{1}$. *Assuming* $\max(\|\mathbf{R}_t^1\|_\infty, \|\mathbf{R}_t^2\|_\infty) \leq s$, *we have the following primal equality*

$$\sum_{t=1}^{T} (\mathbf{p} - \mathbf{p}_t)^\top \mathbf{R}_t \leq \frac{\ln K}{\eta} + s^2 \left(\frac{\eta T}{4} + \frac{\eta}{4} \sum_{t=1}^{T} \lambda_t^2 \right). \tag{6}$$

Proof. Let $W_t = \sum_{i=1}^K w_i^t$. We first show an upper bound and a lower bound on $\ln W_{T+1}/W_1$, followed by combining the bounds together. We have

$$\begin{aligned}
&\sum_{t=1}^{T} \ln \frac{W_{t+1}}{W_t} = \ln \frac{W_{T+1}}{W_1} \\
&= \ln \sum_{i=1}^{K} w_i^{T+1} - \ln K \geq \ln \sum_{i=1}^{K} p_i w_i^{T+1} - \ln K \geq \eta \mathbf{p}^\top \sum_{t=1}^{T} \mathbf{R}_t - \ln K,
\end{aligned}$$

where the last inequality follows from the concavity of the log function. By following Lemma 2.2 in [5], we obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \ln \frac{W_{t+1}}{W_t} = \sum_{t=1}^{T} \sum_{i=1}^{K} \frac{w_i^t \exp(\eta R_i^t)}{\sum_{j=1}^K w_j^t} \\
&\leq \eta \sum_{t=1}^{T} \sum_{i=1}^{K} \frac{w_i^t}{\sum_{j=1}^K w_j^t} R_i^t + \frac{\eta^2}{8} s^2 (1+\lambda_t)^2 \leq \eta \sum_{t=1}^{T} \mathbf{p}_t^\top \mathbf{R}_t + \frac{\eta^2}{8} \sum_{t=1}^{T} s^2 (1+\lambda_t)^2
\end{aligned}$$

Combining the lower and upper bounds and using the inequality $(a+b)^2 \leq 2(a^2+b^2)$, we obtain the desired inequality in (6). □

Lemma 6. [Dual Inequality] *Let* $g_t(\lambda) = \frac{\delta}{2}\lambda^2 + \lambda(\beta_t - c_0)$, $\lambda_{t+1} = [(\lambda_t - \eta \nabla g_t(\lambda_t)]_+$, *and* $\lambda_1 = 0$. *Assuming* $\eta > 0, 0 \leq \beta_t \leq \beta_0$, *we have*

$$\sum_{t=1}^{T} (\lambda_t - \lambda)(\beta_t - c_0) + \frac{\delta}{2} \sum_{t=1}^{T} (\lambda_t^2 - \lambda^2) \leq \frac{\lambda^2}{2\eta} + (c_0^2 + \beta_0^2)\eta T. \tag{7}$$

Proof. First we note that

$$\lambda_{t+1} = [\lambda_t - \eta \nabla g_t(\lambda_t)]_+ = [(1-\delta\eta)\lambda_t - \eta(\beta_t - c_0)]_+ \leq [(1-\delta\eta)\lambda_t + \eta c_0]_+.$$

By induction on λ_t, one can easily show that $\lambda_t \leq \frac{c_0}{\delta}$. Applying the standard analysis of online gradient descent [18] yields

$$\begin{aligned}
|\lambda_{t+1} - \lambda|^2 &= |\Pi_+[\lambda_t - \eta(\delta\lambda_t + \beta_t - c_0)] - \lambda|^2 \\
&\leq |\lambda_t - \lambda|^2 + |\eta(\delta\lambda_t - c_0) + \eta\beta_t|^2 - 2(\lambda_t - \lambda)(\eta \nabla g_t(\lambda_t)) \\
&\leq |\lambda_t - \lambda|^2 + 2\eta^2 c_0^2 + 2\eta^2 \beta_0^2 + 2\eta(g_t(\lambda) - g_t(\lambda_t)).
\end{aligned}$$

Then, by rearranging the terms we get

$$g_t(\lambda_t) - g_t(\lambda) \leq \frac{1}{2\eta} \left(|\lambda_{t+1} - \lambda|^2 - |\lambda_t - \lambda|^2 \right) + \eta(c_0^2 + \beta_0^2)$$

Expanding the terms on l.h.s and taking the sum over t, we obtain the inequality as desired. □

Proof. [of Theorem 1] Applying $\mathbf{R}_t = \mathbf{r}_t + \lambda_t \mathbf{c}_t$ to the primal inequality in Lemma 5, where $\max(\|\mathbf{r}_t\|_\infty, \|\mathbf{c}_t\|_\infty) \leq 1$, we have

$$\sum_{t=1}^{T} (\mathbf{p} - \mathbf{p}_t)^\top (\mathbf{r}_t + \lambda_t \mathbf{c}_t) \leq \frac{\ln K}{\eta} + \frac{\eta T}{4} + \frac{\eta}{4} \sum_{t=1}^{T} \lambda_t^2.$$

Applying $\beta_t = \mathbf{p}_t^\top \mathbf{c}_t$ to the dual inequality in Lemma 6, where $\beta_t \leq 1, c_0 \leq 1$, we have

$$\sum_{t=1}^{T} (\lambda_t - \lambda)(\mathbf{p}_t^\top \mathbf{c}_t - c_0) + \frac{\delta}{2} \sum_{t=1}^{T} (\lambda_t^2 - \lambda^2) \leq \frac{\lambda^2}{2\eta} + 2\eta T.$$

Combining the above two inequalities gives

$$\begin{aligned}
&\sum_{t=1}^{T} (\mathbf{p}^\top \mathbf{r}_t - \mathbf{p}_t^\top \mathbf{r}_t) + \sum_{t=1}^{T} \lambda (c_0 - \mathbf{p}_t^\top \mathbf{c}_t) - \left(\frac{\delta T}{2} + \frac{1}{2\eta}\right) \lambda^2 \\
&\leq \frac{\ln K}{\eta} + \frac{9\eta T}{4} + \left(\frac{\eta}{4} - \frac{\delta}{2}\right) \sum_{t=1}^{T} \lambda_t^2 + \sum_{t=1} \lambda_t (c_0 - \mathbf{p}^\top \mathbf{c}_t).
\end{aligned}$$

Taking expectation over $\mathbf{c}_t, t = 1, \cdots, T$, by using $\mathrm{E}[\mathbf{c}_t] = \mathbf{c}$ and noting that $\mathbf{p}_t$ and λ_t are independent of $\mathbf{c}_t$, we have

$$\begin{aligned}
&\mathrm{E}\left[\sum_{t=1}^{T} (\mathbf{p}^\top \mathbf{r}_t - \mathbf{p}_t^\top \mathbf{r}_t) + \sum_{t=1}^{T} \lambda (c_0 - \mathbf{p}_t^\top \mathbf{c}) - \left(\frac{\delta T}{2} + \frac{1}{2\eta}\right) \lambda^2\right] \\
&\leq \frac{\ln K}{\eta} + \frac{9}{4}\eta T + \mathrm{E}\left[\left(\frac{\eta}{4} - \frac{\delta}{2}\right) \sum_{t=1}^{T} \lambda_t^2\right] + \mathrm{E}\left[\sum_{t=1}^{T} \lambda_t (c_0 - \mathbf{p}^\top \mathbf{c})\right].
\end{aligned}$$

Let $\mathbf{p}$ be the solution satisfying $\mathbf{p}^\top \mathbf{c} \geq c_0$. Noting that $\frac{\eta}{4} - \frac{\delta}{2} \leq 0$ and taking maximization over $\lambda > 0$ in L.H.S, we get

$$\mathrm{E}\left[\max_{\mathbf{p}^\top \mathbf{c} \geq c_0} \sum_{t=1}^{T} \mathbf{p}^\top \mathbf{r}_t - \mathbf{p}_t^\top \mathbf{r}_t\right] + \mathrm{E}\left[\frac{\left[\sum_{t=1}^{T} (c_0 - \mathbf{p}_t^\top \mathbf{c})\right]_+^2}{2(\delta T + 1/\eta)}\right] \leq \frac{\ln K}{\eta} + \frac{9}{4}\eta T.$$

By plugging the values of η and δ, and noting the similar structure of above inequality as in (3) and writing in (4) and (5) formats, we obtain the desired bound for the regret and the violation of the constraint in a long run. □