An Efficient Algorithm for Learning with Semi-bandit Feedback

Gergely Neu[1] and Gábor Bartók[2]

[1] Department of Computer Science and Information Theory
Budapest University of Technology and Economics
gergely.neu@gmail.com
[2] Department of Computer Science, ETH Zürich
bartok@inf.ethz.ch

Abstract. We consider the problem of online combinatorial optimization under semi-bandit feedback. The goal of the learner is to sequentially select its actions from a combinatorial decision set so as to minimize its cumulative loss. We propose a learning algorithm for this problem based on combining the Follow-the-Perturbed-Leader (FPL) prediction method with a novel loss estimation procedure called Geometric Resampling (GR). Contrary to previous solutions, the resulting algorithm can be efficiently implemented for any decision set where efficient offline combinatorial optimization is possible at all. Assuming that the elements of the decision set can be described with d-dimensional binary vectors with at most m non-zero entries, we show that the expected regret of our algorithm after T rounds is $O(m\sqrt{dT \log d})$. As a side result, we also improve the best known regret bounds for FPL in the full information setting to $O(m^{3/2}\sqrt{T \log d})$, gaining a factor of $\sqrt{d/m}$ over previous bounds for this algorithm.

Keywords: Follow-the-perturbed-leader, bandit problems, online learning, combinatorial optimization.

1 Introduction

In this paper, we consider a special case of online linear optimization known as online combinatorial optimization (see Figure 1). In every time step $t = 1, 2, \dots, T$ of this sequential decision problem, the learner chooses an *action* $\boldsymbol{V}_t$ from the finite action set $\mathcal{S} \subseteq \{0,1\}^d$, where $\|\boldsymbol{v}\|_1 \leq m$ holds for all $\boldsymbol{v} \in \mathcal{S}$. At the same time, the environment fixes a loss vector $\boldsymbol{\ell}_t \in [0,1]^d$ and the learner suffers loss $\boldsymbol{V}_t^\top \boldsymbol{\ell}_t$. We allow the loss vector $\boldsymbol{\ell}_t$ to depend on the previous decisions $\boldsymbol{V}_1, \dots, \boldsymbol{V}_{t-1}$ made by the learner, that is, we consider *non-oblivious* environments. The goal of the learner is to minimize the cumulative loss $\sum_{t=1}^T \boldsymbol{V}_t^\top \boldsymbol{\ell}_t$. Then, the performance of the learner is measured in terms of the total expected *regret*

$$R_T = \max_{\boldsymbol{v} \in \mathcal{S}} \mathbb{E}\left[\sum_{t=1}^T (\boldsymbol{V}_t - \boldsymbol{v})^\top \boldsymbol{\ell}_t\right] = \mathbb{E}\left[\sum_{t=1}^T \boldsymbol{V}_t^\top \boldsymbol{\ell}_t\right] - \min_{\boldsymbol{v} \in \mathcal{S}} \mathbb{E}\left[\sum_{t=1}^T \boldsymbol{v}^\top \boldsymbol{\ell}_t\right], \qquad (1)$$

S. Jain et al. (Eds.): ALT 2013, LNAI 8139, pp. 234–248, 2013.

Parameters: set of decision vectors $\mathcal{S} = \{\boldsymbol{v}(1), \boldsymbol{v}(2), \ldots, \boldsymbol{v}(N)\} \subseteq \{0,1\}^d$ satisfying $\|\boldsymbol{v}\|_1 \leq m$ for all $\boldsymbol{v} \in \mathcal{S}$, number of rounds T;
For all $t = 1, 2, \ldots, T$**, repeat**

1. The learner chooses a probability distribution $\boldsymbol{p}_t$ over $\{1, 2, \ldots, N\}$.
2. The learner draws an action I_t randomly according to $\boldsymbol{p}_t$. Consequently, the learner plays decision vector $\boldsymbol{V}_t = \boldsymbol{v}(I_t)$.
3. The environment chooses loss vector $\boldsymbol{\ell}_t$.
4. The learner suffers loss $\boldsymbol{V}_t^\top \boldsymbol{\ell}_t$.
5. The learner observes some feedback based on $\boldsymbol{\ell}_t$ and $\boldsymbol{V}_t$.

Fig. 1. The protocol of online combinatorial optimization

Note that, as indicated in Figure 1, the learner chooses its actions randomly, hence the expectation.

The framework described above is general enough to accommodate a number of interesting problem instances such as path planning, ranking and matching problems, finding minimum-weight spanning trees and cut sets. Accordingly, different versions of this general learning problem have drawn considerable attention in the past few years. These versions differ in the amount of information made available to the learner after each round t. In the simplest setting, called the *full-information* setting, it is assumed that the learner gets to observe the loss vector $\boldsymbol{\ell}_t$ regardless of the choice of $\boldsymbol{V}_t$. However, this assumption does not hold for many practical applications, so it is more interesting to study the problem under *partial information*, meaning that the learner only gets some limited feedback based on its own decision. In particular, in some problems it is realistic to assume that the learner observes the vector $(V_{t,1}\ell_{t,1}, \ldots, V_{t,d}\ell_{t,d})$, where $V_{t,i}$ and $\ell_{t,i}$ are the i^{th} components of the vectors $\boldsymbol{V}_t$ and $\boldsymbol{\ell}_t$, respectively. This information scheme is called *semi-bandit* information. An even more challenging variant is the *full bandit* scheme where all the learner observes after time t is its own loss $\boldsymbol{V}_t^\top \boldsymbol{\ell}_t$.

The most well-known instance of our problem is the (adversarial) *multi-armed bandit* problem considered in the seminal paper of Auer et al. [4]: in each round of this problem, the learner has to select one of N arms and minimize regret against the best fixed arm, while only observing the losses of the chosen arm. In our framework, this setting corresponds to setting $d = N$ and $m = 1$, and assuming either full bandit or semi-bandit feedback. Among other contributions concerning this problem, Auer et al. propose an algorithm called Exp3 (Exploration and Exploitation using Exponential weights) based on constructing loss estimates $\hat{\ell}_{t,i}$ for each component of the loss vector and playing arm i with probability proportional to $\exp(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,i})$ at time t $(\eta > 0)$[1]. This algorithm is known as the Exponentially Weighted Average (EWA) forecaster in the full information case. Besides proving

[1] In fact, Auer et al. mix the resulting distribution with a uniform distribution over the arms with probability $\gamma > 0$. However, this modification is not needed when one is concerned with the total expected regret.