2013 European Control Conference (ECC)
July 17-19, 2013, Zürich, Switzerland.

Mirror Decent Algorithm for a Multi-Armed Bandit Governed by a Stationary Finite State Markov Chain*

Alexander Nazin[1] and Boris Miller[2]

Abstract— **This article further develops an adaptive approach to the control of observable Markov chains with a finite number of states. We apply the Mirror Descent Randomized Control Algorithm (MDRCA) to a class of homogeneous finite Markov chains governed by the multi-armed bandit with unknown mean losses. The article develops the approach represented in [18]. As opposed to the partially observable Markov decision process an adaptive approach does not presuppose the knowledge of probabilistic characteristics of random perturbations and permits to obtain the control strategy with known rate of convergence to the optimal solution. We propose the concrete MDRCA and prove the explicit, non-asymptotic upper bound for the mean losses at a given (finite) time horizon. Numerical example illustrates theoretical results.**

I. INTRODUCTION

Controlled Markov Chains (CMC) constitute a universal mathematical tool applicable to a wide class of applied problems including, but not exhausting: stock and production management [3], internet congestion control [10], large dams management [14] and many others. During recent years the approach to the solution of the optimal control problems with constraints has been developed and applied to the Internet congestion control [11], [15]. Moreover, recently this approach has been extended to the class of connected controlled Markov chains, which is important for control of networks, such as wireless Internet, gas and water allocation systems [12], [14]. However, the existing theory covers mainly the systems with complete observation, where the control depends on the current state of the Markov Chains (MCs). Even if this class of controls provides the optimal solution for the problem with constraints [16], there is an important class of systems, where the state is not observable and have to be estimated on the basis of observations. These problems usually referred to as partially observable Markov decision processes (POMDP). Typical problems arising in the development of Transmission Control Protocols (TCP) in the Internet [24], and in control of autonomous vehicle in abruptly changing environment [25] belong to this class also. Recently one can observe the growing interest to POMDP, the important result of the existence of the optimal policy based on the observation of another jumping process has been proved in [2] with applications to the life testing. Another approach using so-called hybrid control is provided in [4]. Meanwhile, the problems related to TCP also invoke the stochastic control. The typical situation is that at the customer side the state of the router is unknown, and the customer sends the data packets without any guaranty of acceptance or rejection. In some very popular active queue management algorithms like Random Early Detection (RED) the number of rejected packets serves as an information related to the router state and serves as a feedback for customers to control their transmission rate and to prevent the congestion. The approach to the control of transmission rate based on the estimation of the router state is suggested in [13]. In some special cases it was proved that the separation principle is valid and the new algorithms for control of transmission rate have been proposed. It was shown that the optimal law of the increasing of the transmission rate is different from widely known so called additive increase/multiplicative-decrease (AIMD) algorithm and should be realized with the aid of concave window curve which depends on the dynamic of the conditional probabilities for non observable state of router. This algorithm provides much less variability (up to five times [13]) of the transmission rate and therefore more stable work particularly for high speed, long distance and wireless communications networks [8], [9].

Meanwhile in this article we are suggesting another adaptive approach to control where we can control only the random losses associated with the current state of the MC, but not the MC itself. The current state of a MC is assumed to be observable but not controllable, we even do not know the matrix of the transition probabilities. However, one can choose the control strategy which minimizes the average losses for sufficiently long control period and at the same time to estimate the rate of convergence to the optimal solution. Important non-asymptotic result is the dependence of this rate of convergence from the time horizon, numbers of the MC states and controls. One can obtain also *the robust* algorithm which ensures dependence of such convergence rate and does not depend on the probability properties of the MC at all.

The structure of the paper is as follows. In the next Section II we give the problem statement and basic assumptions. The main theorem which establishes the convergence and gives the estimate for the convergence rate is in Section III. In the Section IV we provide the description of the control

*This work was supported by the Australian Foundation for Scientific Research grant ARC DP 0988685 and Russian Foundation for Basic Research under grants 10-01-00710, 10-08-01068, 11-08-00223, 12-08-01245, and 13-01-00406.

[1]A. Nazin is with the Laboratory for Adaptive and Robust Control Systems, Institute of Control Sciences RAS; he performed this work as a visitor of the School of Mathematical Sciences of Monash University, Victoria, Australia nazine@ipu.ru

[2]B. Miller is with the School of Mathematical Sciences of Monash University, Victoria, Australia and the Institute of Information Transmission Problems, Russian Academy of Sciences boris.miller@Monash.edu

algorithm in the form of randomized strategy. Illustrative numerical example is represented in Section V. Section VI is conclusion where we discuss the results and directions of further research. Proofs of results are given in Appendix.

Notation. Denote vector $e_N^0 \triangleq (1,\dots,1)^\top \in \mathbb{R}^N$, unit-vectors $e_m(k) = (0,\dots,0,1,0,\dots,0)^\top \in \mathbb{R}^m$ (with 1 on k-th position and 0 elsewhere), and standard simplex in $\mathbb{R}^N$

$$S_N \triangleq \left\{ (x^{(1)},\dots,x^{(N)})^\top \,\middle|\, x^{(i)} \geq 0, \sum_{i=1}^N x^{(i)} = 1 \right\}. \qquad (1)$$

Given a matrix $\Xi \in \mathbb{R}^{K\times N}$, its transposed i-th row represents vector $\Xi^{(i)} \in \mathbb{R}^N$, and its (ik)-entry is denoted by $\Xi^{(ik)}$.

II. Statement of Problem

This section is partially adopted from [20], chapter 5; c.f. [18].

A. Preliminaries

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a given probability space. Let a discrete-time stochastic system with discrete control be modeled as a multi-armed bandit governed by a stationary homogeneous finite Markov Chain where

- the set of states $Z \triangleq \{z(1),\dots,z(K)\}$ is given, $K \geq 2$;
- system state $z_t \in Z$ at time $t \in \{0,1,\dots\}$ is observable;
- the given set $U \triangleq \{u(1),\dots,u(N)\}$ represents the set of control inputs (or bandit arms, in other words), $N \geq 2$;
- the transition probabilities of the system state $z_t \in Z$ at each time $t \in \{0,1,\dots\}$ to the next state $z_{t+1} \in Z$ are presented by unknown conditional probabilities: $\forall t$,

$$\mathbb{P}\{z_{t+1} = z(j) \,|\, z_t = z(i)\} = \pi_{ij} \qquad (2)$$

(notice that the transition probabilities (2) are independent of the applied control);

- the random losses $\xi_t \triangleq \xi_t(z_t, u_t, \omega)$ at current time $t \in \{0,1,\dots\}$ are observable and statistically depend on the state z_t and the applied control u_t, with *unknown* conditional distributions;
- the time-mean losses Φ_t on time interval $\{\overline{1,t}\}$ are

$$\Phi_t = \frac{1}{t}\sum_{s=1}^t \xi_s. \qquad (3)$$

Introduce the further assumptions.

A1. For each $t = 1,2,\dots$ the sets of random variables

$$\{\xi_t(z,u,\omega) \,|\, z \in Z, u \in U\}$$

and

$$\{\xi_s(z,u,\omega), z_k, u_k \,|\, z \in Z, u \in U, s = \overline{1,t-1}, k = \overline{1,t}\}$$

are independent.

A2. For each $z(i) \in Z$, $u(\ell) \in U$, and $t = 1,2,\dots$ the losses $\xi_t(z(i),u(\ell),\omega)$ are non-negative a.s. and their *a priori unknown* expectations are time-invariant:

$$\mathbb{E}\{\xi_t(z(i),u(\ell),\omega)\} \triangleq a_{i\ell} \quad \forall t. \qquad (4)$$

A3. The losses $\xi_t(z(i),u(\ell),\omega)$ are bounded in the mean square sense, i.e.

$$\mathbb{E}\{\xi_t^2(z(i),u(\ell),\omega)\} \leq \sigma^2 < \infty. \qquad (5)$$

A4. The Markov chain of transition probability matrix Π is ergodic (i.e., the state set Z represents a unique ergodic class).

Remark 1: Assumption A4 implies the existence of the unique stationary state probabilities $q_i > 0$, $i = \overline{1,K}$. □

A5. MC initial distribution is assumed to be stationary.

B. Control strategies

Introduce

$$d_t^{(i\ell)} \triangleq \mathbb{P}\{u_t = u(\ell) \,|\, z_t = z(i), \mathscr{F}_{t-1}\} \qquad (6)$$

that are the conditional probabilities of control $u_t = u(\ell)$ at time t under the state $z_t = z(i)$ and the prehistory, i.e., $\mathscr{F}_t = \sigma\{z_s, u_s, \xi_s \,|\, s = \overline{1,t}\}$.

In order to simplify the preliminary analysis, let us consider a stationary control strategy $\mathscr{U}_{\rm St}$ (with the stationary randomized control strategy $d \triangleq \|d^{(i\ell)}\|$) leading to the loss expectation

$$\begin{aligned} \mathbb{E}\{\xi_t\} &= \mathbb{E}\left\{\sum_{i=1}^K \mathbf{1}_{\{z_t = z(i)\}} \sum_{\ell=1}^N a_{i\ell}\, d^{(i\ell)}\right\} && (7)\\ &= \sum_{i=1}^K q_i \sum_{\ell=1}^N a_{i\ell}\, d^{(i\ell)} && (8)\\ &\triangleq A(d) && (9) \end{aligned}$$

where

$$q_i \triangleq \mathbb{P}\{z_t = z(i)\} \qquad (10)$$

represent the stationary probabilities of the Markov states (see Assumptions A4, A5), the matrix of conditional probabilities

$$d^{(i\ell)} = \mathbb{P}\{u_t = u(\ell) \,|\, z_t = z(i)\} \qquad (11)$$

relates to a stationary randomized control strategy $\mathscr{U}_{\rm St}$, stochastic matrix $d = \|d^{(i\ell)}\| \in D$,

$$D \triangleq \left\{ d \,\middle|\, d^{(i\ell)} \geq 0, \sum_{\ell=1}^N d^{(i\ell)} = 1 \ (i = \overline{1,K}, \ell = \overline{1,N}) \right\}.$$

The stochastic vector $q = (q_1,\dots,q_K)^\top$ solves the stationary distribution equation for the stationary Markov chain that is $q = \Pi^\top q, \quad q \in S_K$; the transition probability matrix Π has the (ij)-entry π_{ij}.

Denote set of non-degenerate stationary randomized control strategies $\mathscr{U}_{\rm St}^+$ that is

$$D_+ \triangleq D \bigcap \left\{ d = \|d^{(i\ell)}\| \,\middle|\, d^{(i\ell)} > 0, (i = \overline{1,K}, \ell = \overline{1,N}) \right\}.$$

Remark 2: If the matrix of mean losses $\|a_{i\ell}\|$ is known a priori, one could find the optimal control strategy without knowledge of the state probabilities q_i. Indeed,

$$\min_{d\in D} A(d) = \sum_{i=1}^K q_i \min_{\ell=\overline{1,N}} a_{i\ell},$$

and each state index $i = \overline{1,K}$ gives optimal pure strategy $(d_{\rm opt}^{(i1)},\dots,d_{\rm opt}^{(iN)})^T = e_N(\ell_i^*)$ with 1 at ℓ_i^*-th entry,

$$\ell_i^* \in \mathrm{Argmin}_{\ell=\overline{1,N}}\, a_{i\ell}.$$

Since matrix $\|a_{i\ell}\|$ is unknown, the designing adaptive control strategy represents a non-trivial optimization control problem under uncertainty. □

The idea of designing the randomized control strategy is to minimize the mean loss function $A(d)$ in (9) on set D

$$A_{\min} \triangleq \min_{d\in D} A(d). \tag{12}$$

III. Main Results

Below we propose the online control strategy: at each time t, given the observation of MC state $z_t = z(i)$, the control action $u_t \in U$ is randomly drawn according to a conditional distribution $d_t^{(i)} = (d_t^{(i1)},\dots,d_t^{(iN)})^\top \in S_N$,

$$d_t^{(i\ell)} \triangleq \mathbb{P}(u_t = u(\ell)\,|\,z_t = z(i), \mathscr{F}_t), \quad \forall (i,\ell). \tag{13}$$

The update rule of the distributions $d_t^{(i)}$ over time is given by the algorithm described in Section IV and uses stochastic gradient for $A(d)$, i.e. random matrix entries

$$\Xi_{t+1}^{(i\ell)} \triangleq \xi_{t+1}\mathbf{1}_{\{z_t=z(i),u_t=u(\ell)\}}/d_t^{(i\ell)}. \tag{14}$$

Indeed, under stationarity assumption of control strategy $d \equiv d_t \in D_+$ and, for all (i,ℓ),

$$\mathbb{E}\left(\Xi_{t+1}^{(i\ell)}\right) = \mathbb{E}\left\{\mathbb{E}\left(\frac{\xi_{t+1}\mathbf{1}_{\{z_t=z(i),u_t=u(\ell)\}}}{d^{(i\ell)}}\,|\,\mathscr{F}_t\right)\right\} \tag{15}$$

$$= \mathbb{E}\left\{\frac{a_{i\ell}}{d^{(i\ell)}}\,\mathbf{1}_{\{z_t=z(i),u_t=u(\ell)\}}\right\} = q_i\,a_{i\ell} = \frac{\partial A(d)}{\partial d^{(i\ell)}}. \tag{16}$$

The expected average loss equals to the average over time of $\mathbb{E}A(d_t)$, that is

$$\mathbb{E}(\Phi_T) = \frac{1}{T}\sum_{t=1}^T \mathbb{E}(\mathbb{E}(\xi_t\,|\,z_{t-1},\mathscr{F}_{t-1})) \tag{17}$$

$$= \frac{1}{T}\sum_{t=1}^T \mathbb{E}(A(d_{t-1})). \tag{18}$$

Theorem 1: Let assumptions A1–A5 be satisfied and let the conditional distributions $(d_t^{(i)})_{t\ge 0}$, $i=\overline{1,K}$, be defined by the randomized control algorithm of Section IV with parameters (23). Then, for a given horizon $T \ge 1$,

$$\mathbb{E}\,(\Phi_T) - A_{\min} \le \sigma\sqrt{\frac{2N\ln N}{T}}\sum_{i=1}^K \sqrt{q_i}. \tag{19}$$

■

IV. Definition of the randomized strategy

In this section we introduce our online strategy (cf. [7] and [6]). We refer to [21] and [1] for the general idea of mirror descent and its development in non-stochastic optimization, as well as to [22] for the pioneering extension to a stochastic setup.

First we introduce a Gibbs distribution defined by the probability vector-function

$$G_\beta(\zeta) = [S_\beta(\zeta)]^{-1}\left(e^{-\zeta^{(1)}/\beta},\dots,e^{-\zeta^{(N)}/\beta}\right)^\top$$

where $S_\beta(\zeta) = \sum_{j=1}^N e^{-\zeta^{(j)}/\beta}$ for arbitrary fixed $\zeta \in \mathbb{R}^N$ and parameter $\beta > 0$. Note, that ζ represents a dual vector variable, see (26) below in the Appendix.

Thus, we fix horizon T, temperature parameters $\beta_i > 0$, $i=\overline{1,K}$, and define the control randomized strategy as follows.

1) *Fix the initial matrix d_0 with equal entries, i.e., $d_0^{(ij)} \equiv 1/N$, and zero dual matrix $\zeta_0 = 0 \in \mathbb{R}^{K\times N}$;*
   a) *for each $t \ge 0$, by having the observed state $z_t = z(i_t)$, draw control action $u_t = u(\ell_t)$ with random $\ell_t \in \{\overline{1,N}\}$ distributed according to stochastic vector $(d_t^{(i_t1)},\dots,d_t^{(i_tN)})^\top$;*
   b) *compute a stochastic gradient*

$$\Xi_{t+1} = \frac{\xi_{t+1}}{d_t^{i_t\ell_t}}\,e_K(i_t)e_N^\top(\ell_t); \tag{20}$$

   c) *update both dual and initial variables*

$$\zeta_{t+1} = \zeta_t + \Xi_{t+1}, \tag{21}$$

$$d_{t+1}^{(i)} = \begin{cases} d_t^{(i)}, & i \ne i_t, \\ G_{\beta_i}(\zeta_{t+1}^{(i)}), & i = i_t. \end{cases} \tag{22}$$

2) *At horizon T, output sequences of states $(z_0,\dots,z_T)$, control actions $(u_0,\dots,u_T)$, matrices $(d_0,\dots,d_T)$, and the observed losses $(\xi_1,\dots,\xi_{T+1})$ and Φ_T.*

Remark 3: Notice that matrix Ξ_{t+1} in (20) contains a unique (random) nonzero entry. Therefore, the rows of dual variable ζ_t change at time t for the related unique index $i = i_t$. The Gibbs distribution maps vector $\zeta_{t+1}^{(i_t)}$ at time $t+1$ for current state z_{i_t} resulting initial variable $d_{t+1}^{(i_t)}$.

The presented algorithm explains its structure: at each time t, by obtaining stochastic gradient Ξ_{t+1}, we make a step in the dual space by transposed line-row applying the result to Gibbs distribution $G_{\beta_{i_t}}$ and obtaining d_{t+1}. □

The algorithm parameters β_i depend on the horizon T as follows:

$$\beta_i = \sigma\sqrt{\frac{TNq_i}{2\ln N}}, \quad i=\overline{1,K}. \tag{23}$$

They assume the state probabilities q_i; in case they are not available one could apply an adaptive idea to recursively update the temperature parameters β_i in time, see, e.g. [19].

Remark 4: Notice the bounds $1 \le \sum_{i=1}^K \sqrt{q_i} \le \sqrt{K}$ for $q \in S_K$ in (19). The left inequality attains the equality on each symplex vertex, when (19) implies the related bound for the multi-armed bandit problem, e.g. [6]; the right inequality attains at the symplex center when state probabilities become equal, $q_i \equiv 1/K$. The worst upper bound in (19), i.e.

$$\mathbb{E}\,(\Phi_T) - A_{\min} \le \sigma\sqrt{\frac{2KN\ln N}{T}}, \tag{24}$$

may be proved for a simpler case of our algorithm (which is additionally robust with respect to unknown state probability distribution q) having the identical temperature parameters

$$\beta_i \equiv \beta \triangleq \sigma\sqrt{\frac{TN}{2K\ln N}}. \tag{25}$$

That may be directly demonsrated at the end of Theorem 1 proof, see Appendix.

□

V. Numerical example

To illustrate the algorithm and the degree of divergence between the upper bound of Theorem 1 and l.h.s. of (19), consider the numerical example with the number of states $K = 7$ and number of control inputs $N = 5$. The cyclic transition probability matrix is fixed as follows

$$\|\pi_{ij}\| = \begin{pmatrix} 1/4 & 1/2 & 0 & 0 & 0 & 0 & 1/4 \\ 1/4 & 1/4 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/4 & 1/4 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 1/4 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/4 & 1/4 & 1/2 \\ 1/2 & 0 & 0 & 0 & 0 & 1/4 & 1/4 \end{pmatrix}$$

resulting the row of equal stationary state probabilities $(1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1)/7$; therefore, parameters β_i coincide due to both (24) and (25). The mean loss matrix is as follows

$$\|a_{i\ell}\| = \begin{pmatrix} 0.1 & 0.3 & 0.5 & 0.7 & 0.9 \\ 0.55 & 0.15 & 0.25 & 0.35 & 0.45 \\ 0.325 & 0.375 & 0.175 & 0.225 & 0.275 \\ 0.2375 & 0.2625 & 0.2875 & 0.1875 & 0.2125 \\ 0.175 & 0.225 & 0.325 & 0.375 & 0.275 \\ 0.15 & 0.25 & 0.35 & 0.55 & 0.45 \\ 0.1 & 0.7 & 0.9 & 0.3 & 0.5 \end{pmatrix}$$

having the minimum mean losses $A_{\min} = 0.1482$; therefore, parameter $\sigma = 1$, and optimal pure strategy is represented by vector of integers $\ell^* = (1,\ 2,\ 3,\ 4,\ 1,\ 1,\ 1)^T$. Random losses $\xi_t(z(i),u(\ell),\omega)$ at state $z(i)$ and control input $u(\ell)$ represent the i.i.d. Bernoully random variables with probability $\mathbb{P}(\xi_t(z(i),u(\ell),\omega) = 1) = a_{i\ell}$. We put horizon $T = 10000$ and apply equal parameters $\beta_i = 47.1068$.

There are log-log scale plots for losses versus time $t = 100,\ldots,10000$ in Fig. 1. The top solid (almost linear) line represents the upper bound of Theorem 1, that is r.h.s. of (19). The two dashed (red) "noised" lines, both upper and lower ones, demonstrate envelope of 100 algorithm realizations, time-mean losses Φ_t (3); the mean of these 100 realizations is represented by (green) bold line, which is about three times less than the upper bound at $T = 10000$.

The results of the example show that for sufficiently large time t, say $t \geq 1500$, the l.h.s. of (19) which is related to (green) bold line in Fig. 1 behave similarly as r.h.s. of (19) with additional factor $1/3$. Therefore, the degree of divergence between the upper bound of Theorem 1 and l.h.s. of (19) is small in the above sense.

VI. Conclusions

We stated the optimization problem for homogeneous finite Markov chains governed by the multi-armed bandit with unknown mean losses on a given horizon T. Basing on the adaptive approach, we proposed the mirror descent randomized algorithm which ensures explicit non-asymptotic rate of convergence for mean losses to the minimum. The algorithm does not know the MC state transit probabilities, but stationary state probabilities are supposed to be available. The algorithm parameters and the rate of convergence depend

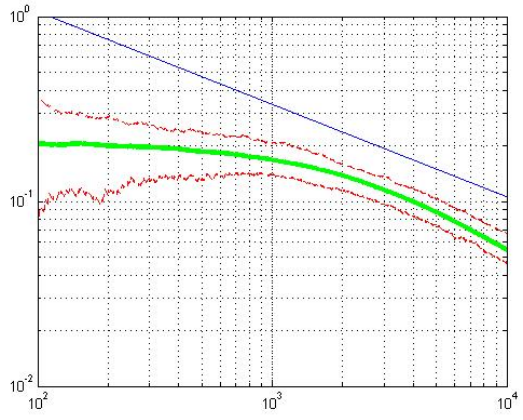

Fig. 1. The results of the numerical example with $K = 7$ and $N = 5$ are represented by log-log scale plots of the losses versus time $t = 100,\ldots,10000$.

explicitly on the stationary state probabilities $q_1,\ldots,q_K$, N number of bandit arms, T current time, and σ the bound on square-root moment for random losses.

In further research we are going to develop the robust algorithms and to obtain the convergence rate if the stationary probabilities are unknown (see Remark 4).

References

[1] A. Ben-Tal and A. S. Nemirovski. *The conjugate barrier mirror descent method for non-smooth convex optimization.* Minerva optimization center, Technion Institute of Technology, 1999.

[2] C. Ceci, A. Gerardi, and P. Tardelli. *Existence of Optimal Controls for Partially Observed Jump Processes.* Acta Applicandae Mathematicae, v. 74, 2002, pp. 155-175.

[3] M. H. A. Davis. *Markov Models and Optimization.* London: Chapman and Hall, 1993.

[4] R. J. Elliott, L. Aggoun, and J. B. Moore. *Hidden Markov Models. Estimation and Control.* New York: Springer Verlag, 1995, 2008.

[5] R. Howard. *Dynamic Programming and Markov Processes.* New York, London: John Wiley & Sons, Inc, 1960.

[6] A. Juditsky, A. V. Nazin, A. Tsybakov, and N. Vayatis. *Gap-free bounds for stochastic multi-armed bandit.* 17th IFAC World Congress, Seoul, Korea, 6–11 July, 2008.

[7] A. B. Juditsky, A. V. Nazin, A. B. Tsybakov, and N. Vayatis. *Recursive aggregation of estimators by the mirror descent algorithm with averaging.* Problems of Information Transmission, 41(4):368–384, 2005. Translated from Problemy Peredachi Informatsii, No.4, 2005, pp.78–96.

[8] Shao Liu, Tamer Başar, and R. Srikant. *TCP-Illinois: A loss- and delay-based congestion control algorithm for high-speed networks.* Performance Evaluation v. 65, 2008, pp. 417-440.

[9] D. Kliazovich, F. Granelli, and D. Miorandi. *Logarithmic window increase for TCP Westwood+ for improvement in high speed, long distance networks.* Computer Networks, Vol. 52, pp. 2395-2410, 2008.

[10] S. H. Low, F. Paganini, and J. C. Doyle. *Internet Congestion Control.* IEEE Control Systems Magazine, 2002, pp. 28–43.

[11] A. Miller and B. Miller. *Congestion avoidance with the aid of stochastic control.* Proceedings of the 49th IEEE CDC 2010, Atlanta, Geogia, USA, December, pp. 552–558, 2010.

[12] A. Miller and B. Miller. *Control of connected Markov chains. Application to congestion avoidance in the Internet.* Proceedings of 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC) Orlando, FL, USA, December 12-15, 2011, pp. 7242–7248.

[13] B. Miller, K. Avrachenkov, K. Stepanyan, and G. Miller. *The problem of the optimal stochastic control of a data flow with incomplete information.* Problems of Information Transmission, v. 41 n 2, 2005, pp. 150–170.

[14] B. Miller and D. McInnes. *Optimal Management of a Two Dam System via Stochastic Control: Parallel Computing Approach.* Proceedings of 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC) Orlando, FL, USA, December 12-15, 2011, pp. 1417-1423.

[15] B. Miller, G. Miller, and K. Siemenikhin. *Torwards the optimal control of Markov chains with constraints.* Automatica, vol. 46, pp. 1495–1502, 2010.

[16] B. M. Miller, G. B. Miller, and K. V. Semenikhin. *Methods to Design Optimal Control of Markov Process with Finite State Set in the Presence of Constraints.* Automation and Remote Control, 2011, Vol. 72, No. 2, pp. 323-341.

[17] G. E. Monahan. *Survey of Partially Observable Markov Decision Processes: Theory, Models and Algorithms.* Management Science, Vol. 28, No. 1, pp. 1–16, 1982.

[18] A. V. Nazin and B. M. Miller. *The Mirror Descent Control Algorithm for Weakly Regular Homogeneous Finite Markov Chains with Unknown Mean Losses.* Proceedings of 50th IEEE CDC ECC, Orlando, Florida, USA, 2011, December 12-15.

[19] A. V. Nazin and B. T. Polyak, *Adaptive Randomized Algorithm for Finding Eigenvector of Stochastic Matrix with Application to PageRank.* Proc. combined 48th IEEE Conf. Decision Control and 28th Chinese Control Conf., Shanghai, China, Dec. 2009.

[20] A. V. Nazin and A. S. Poznyak. *Adaptive Choice of Variants*. Nauka, Moscow, 1986. (In Russian).

[21] A. S. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

[22] Yu. Nesterov. *Primal-dual subgradient methods for convex problems.* Core discussion paper 2005/67. Louvain-la-Neuve, Belgium: Center for Operation Research and Econometrics, 2005.

[23] U. Rieder and Jens Winter. *Optimal control of Markovian jump processes with partial information and applications to a parallel queueing model.* Math Meth Oper Res, Vol. 70, pp. 567-596, 2009.

[24] R. Srikant *The mathematics of Internet congestion control.* Boston : Birkhuser, 2004.

[25] D. Sworder, John E. Boyd. *Estimation problems in hybrid systems.* Cambridge; New York: Cambridge University Press, 1999.

[26] Jens Winter. *Finite Horizon Control Problems Under Partial Information.* NET-COOP 2007, T. Chahed and B. Tuffin (Eds.): LNCS 4465, pp. 120-128, 2007.

APPENDIX

A. Gibbs distribution.

For the convenience of reader, recall the properties of function $G_\beta(\cdot)$ (cf., e.g., [7]). We have $G_\beta(\zeta) = -\nabla W_\beta(\zeta)$,

$$W_\beta(\zeta) = \beta \ln\left(\frac{1}{N}\sum_{k=1}^N e^{-\zeta^{(k)}/\beta}\right), \quad z \in \mathbb{R}^N.$$

Function W_β and the entropy type function

$$V(\theta) \triangleq \ln N + \sum_{j=1}^N \theta^{(j)} \ln \theta^{(j)} \geq 0, \quad \theta \in S_N,$$

are related to each other via convex duality formula:

$$W_\beta(\zeta) = \sup_{\theta \in S_N}\left\{-\zeta^\top \theta - \beta V(\theta)\right\}, \quad \zeta \in \mathbb{R}^N. \qquad (26)$$

Recall $\nabla W_\beta(\zeta) \equiv -G_\beta(\zeta)$.

B. Proof of Theorem 1

Let us fix arbitrary $i \in \{\overline{1,K}\}$ and simply write β instead of β_i, skipping the low index for a while. Note that $W_\beta(\zeta_{t+1}^{(i)}) = W_\beta(\zeta_t^{(i)})$ for all $i \neq i_t$, otherwise

$$W_\beta(\zeta_{t+1}^{(i)}) - W_\beta(\zeta_t^{(i)}) = \beta \ln\left(\frac{\sum_{k=1}^N e^{-\zeta_{t+1}^{(ik)}/\beta}}{\sum_{k=1}^N e^{-\zeta_t^{(ik)}/\beta}}\right)$$
$$= \beta \ln((d_t^{(i)})^\top v_{t+1}^{(i)}), \quad \text{for} \quad i = i_t$$

where k-th entry of i-th vector $v_t^{(i)}$ equals $v_t^{(ik)} = e^{-\Xi_t^{(ik)}/\beta}$. Since $e^x \leq 1 + x + x^2/2$ for $x \leq 0$, we get

$$v_t^{(ik)} \quad \leq \quad 1 - \frac{\Xi_t^{(ik)}}{\beta} + \frac{(\Xi_t^{(ik)})^2}{2\beta^2}.$$

Recalling $\Xi_{t+1}^{(ik)}$ (20), we obtain $(d_t^{(i)})^\top \Xi_{t+1}^{(i)} = 0$ for $i \neq i_t$, and

$$(d_t^{(i_t)})^\top \Xi_{t+1}^{(i_t)} = \sum_{k=1}^N d_t^{(i_t k)} \mathbf{1}_{\{k=\ell_t\}} \xi_{t+1} / d_t^{(i_t \ell_t)} = \xi_{t+1},$$

we bound, for $i = i_t$,

$$\beta \ln((d_t^{(i)})^\top v_{t+1}^{(i)}) \quad \leq \quad \beta \ln\left(1 - \frac{\xi_{t+1}}{\beta} + \frac{\xi_{t+1}^2}{2d_t^{(i\ell_t)}\beta^2}\right)$$
$$\leq \quad -\xi_{t+1} + \frac{\xi_{t+1}^2}{2d_t^{(i\ell_t)}\beta}.$$

Combining the previous cases one may write: $\forall\, i \in \{\overline{1,K}\}$,

$$W_\beta(\zeta_{t+1}^{(i)}) - W_\beta(\zeta_t^{(i)}) \leq \mathbf{1}_{\{i=i_t\}}\left(-\xi_{t+1} + \frac{\xi_{t+1}^2}{2d_t^{(i\ell_t)}\beta}\right).$$

Now take expectation of both sides (first over ℓ_t, conditional on i_t and d_t, then taking the full expectation) and applying assumption A2 we obtain

$$\begin{array}{rl} & \mathbb{E}\left(W_\beta(\zeta_{t+1}^{(i)}) - W_\beta(\zeta_t^{(i)})\right) \\ \leq & -\mathbb{E}(\xi_{t+1}\mathbf{1}_{\{i=i_t\}}) + \frac{N\sigma^2 q_i}{2\beta}. \end{array} \qquad (27)$$

Summing up from $t = 0$ to $t = T-1$ we obtain

$$\sum_{t=1}^T \mathbb{E}(\xi_t \mathbf{1}_{\{i=i_{t-1}\}}) \quad \leq \quad -\mathbb{E}W_\beta(\zeta_T^{(i)}) + \frac{N\sigma^2 q_i}{2\beta} T. \qquad (28)$$

The minimizer $d_{\text{opt}} \triangleq \arg\min_{d \in D} A(d)$ satisfies $A(d_{\text{opt}}) = A_{\min}$ (12). Then we apply the duality formula (26) and use the inequality

$$W_\beta(\zeta_T^{(i)}) \geq -(\zeta_T^{(i)})^\top d_{\text{opt}}^{(i)} - \beta V(d_{\text{opt}}^{(i)}).$$

Using (28), the fact that $\sup_{\theta \in S_N} V(\theta) = \ln N$, and the last display we obtain $\mathbb{E}W_\beta(\zeta_T^{(i)}) \geq -\mathbb{E}((\zeta_T^{(i)})^\top d_{\text{opt}}^{(i)}) - \beta \ln N$ with the expectation

$$\begin{array}{rcl} \mathbb{E}((\zeta_T^{(i)})^\top d_{\text{opt}}^{(i)}) & = & \sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{1}_{\{i=i_t\}} \sum_{\ell=1}^N a_{i\ell} e_N^\top(\ell) d_{\text{opt}}^{(i)}\right] \\ & = & T q_i \min_{\ell=\overline{1,N}} a_{i\ell}. \end{array} \qquad (29)$$

Therefore, (28)–(29) lead to

$$\begin{array}{rcl} \sum_{t=1}^T \mathbb{E}\left[(\xi_t \mathbf{1}_{\{i=i_{t-1}\}}) - q_i \min_{\ell=\overline{1,N}} a_{i\ell}\right] & \leq & \beta \ln N + \frac{N\sigma^2 q_i}{2\beta} T \\ & = & \sigma\sqrt{2 q_i T N \ln N} \end{array}$$

under $\beta = \beta_i$ (23). Dividing by time T, summarizing by $i = \overline{1,K}$, and bounding the sum under sequence (β_t) (23) we obtain

$$\mathbb{E}(\Phi_T) - A_{\min} \leq \sigma\sqrt{\frac{2N\ln N}{T}} \sum_{i=1}^K \sqrt{q_i}.$$

One may see that the obtained upper bound above reduces to the square root dependence like $O(\sqrt{T})$ and optimizes parameter β_0. This proves the theorem. ▲