2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

The Mirror Descent Control Algorithm for Weakly Regular Homogeneous Finite Markov Chains with Unknown Mean Losses

Alexander V. Nazin and Boris Miller

Abstract— **We address the adaptive stochastic control problem for a discrete time system described by controlled Markov chain with finite number of states. The mirror descent randomized control algorithm on the class of controlled homogeneous finite Markov chains with unknown mean losses has been proposed and studied. Here we develop the approach represented in Nazin and Miller (2011). The main assumptions are the following: processes are independent and stationary, nonnegative random losses are almost surely bounded by a given constant, and the connectivity assumption for the controlled Markov chain holds. The uncertainty is that the mean loss matrix is unknown. The novelty of the approach is in extension of the class of controlled homogeneous finite Markov chains to the chains with connectivity assumption. The main result consists in demonstration of the asymptotical upper bound (that is asymptotic by time) and in determining the explicit constant which is weakly depending on the logarithm of the number of states.**

I. Introduction

Controlled Markov chains play an important role in the optimal control of stochastic systems. One of the most important advantages of this class of models is that they admit the complete numerical solution which can be obtained by solving the system of difference equations (dynamic programming equation) [2], perhaps of very large dimension. Solution of this system gives the complete characterization of the cost functions and the optimal control of Markov type as a function of the current state and time. However, the application of the methodology of controlled Markov chains requires the knowledge of complete information about the cost functions and the properties of controlled transition matrices of the Markov chain. This is rather rare in real systems and usually it becomes necessary to simplify the general model by using the representation of unknown state by a finite number of possible ones like in the theory of Hidden Markov Models [2]. Typical example where complete information is not available, is the Internet Congestion Control [3], [7], [14], [15], where the transition rate at the user side is determined by the router feedback (unknown to the user) and provided to the user just as a flow of rejected demands. The real state of the router is covered in the flow intensity, so the user has either to evaluate it [8] or to adjust your behavior to the state of router and to unknown state of the transmission line [9].

The idea of this work is to extend the methodology introduced in [9] to more general class of weakly regular controlled Markov chain. Here we develop the approach to control of homogeneous finite Markov chains with unknown mean losses which was first introduced in [10]. We study a Mirror Descent Algorithm in the spirit of [4], [9] as an extension of this control methodology [10].

We assume that for given control action, the matrix of state transition probabilities is known a priori, but the current random losses are statistically undefined.

The main result of the article is as follows: under assumption of nonnegative losses being a.s. bounded by known constant $\sigma > 0$ we demonstrate that the expected excess bound of losses for large enough horizon T implies the convergence rate $O(T^{-1/3})\sigma\,(N\ln(KN))^{1/3}$ where N means the number of control actions, K stands for the number of states.

The structure of the paper is as follows: we give the problem statement in Section II, define the asymptotic upper bound in Section III and define the randomized strategy in Section IV. All proofs are presented in Appendix.

II. Statement of Problem

This section is essentially adopted from chapter 5 in [10]; c.f. [9]. First, denote vector $e_N^0 \triangleq (1,\dots,1)^\top \in \mathbb{R}^N$ and standard simplex in $\mathbb{R}^m$

$$S_m \triangleq \left\{ (x_1,\dots,x_m)^\top \,\middle|\, x_i \geq 0, \sum_{i=1}^m x_i = 1 \right\}. \qquad (1)$$

A. Preliminary assumptions

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a given probability space. Let a discrete-time stochastic control system be modeled as a homogeneous controlled finite Markov chain where

- the set of states $Z \triangleq \{z(1),\dots,z(K)\}$ is given, $K \geq 2$;
- the system state $z_t \in Z$ at current time $t \in \{0,1,\dots\}$ is observable;
- the given set $U \triangleq \{u(1),\dots,u(N)\}$ represents the set of possible control inputs, $N \geq 2$;

This work was supported in part by Australian Research Council Grant DP0988685. The work of the first author was supported in part by Russian Basic Research Foundation Grants 10-08-01068 and 11-08-00223. The second author was supported in part by Russian Basic Research Foundation Grant 10-01-00710.

A.V. Nazin is with the Laboratory for Adaptive and Robust Control Systems, Institute of Control Sciences RAS, 65 Profsoyuznaya str., 117997 Moscow, Russia (Tel: +7 495 334 7641; e-mail: nazine@ipu.ru) and the School of Mathematical Sciences, Monash University, Clayton, Victoria, Vic 3800, Australia.

B. Miller is with the School of Mathematical Sciences, Monash University, Clayton, Victoria, Vic 3800, Australia (Tel: +61 3 9905 5870; e-mail: boris.miller@sci.monash.edu.au) and Institute for Information Transmission Problems RAS, 19, B. Karetny, GSP-4, Moscow, Russia.

- the transition probabilities of the system state $z_t \in Z$ at current time $t \in \{0,1,\dots\}$ to the next state $z_{t+1} \in Z$ under the applied control $u_t \in U$ are given by the *given* conditional probabilities: $\forall t$,

$$\mathbb{P}\{z_{t+1} = z(j) \,|\, z_t = z(i), u_t = u(\ell), \mathscr{F}_t\} = \pi_{ij}^{\ell}; \quad (2)$$

here σ-algebra $\mathscr{F}_t$ is generated by the prehistory of observations up to time t; in other words, equation (2) relates to the system Markov property;

- the losses $\xi_t \triangleq \xi_t(z_t, u_t, \omega)$ at current time $t \in \{0,1,\dots\}$ are observable and statistically depend only on the state z_t and the applied control u_t with *unknown* conditional distributions; the random variables $\xi_t \triangleq \xi_t(z(i), u(\ell), \omega)$ form i.i.d. sequences by time t for all $i \in \{\overline{1,K}\}$ and $\ell \in \{\overline{1,N}\}$;
- the time-mean losses Φ_t on time interval $\{\overline{1,t}\}$ are defined by

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

Introduce the further assumptions A1-A2-A3.

A1. For each $t = 1,2,\dots$ the totalities 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 *unavailable* expectations are time-invariant:

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

A3. The losses $\xi_t(z(i), u(\ell), \omega)$ are a.s. bounded by given constant $\sigma > 0$, i.e.

$$\xi_t(z(i), u(\ell), \omega) \le \sigma < \infty. \quad (5)$$

B. Control strategies

Consider arbitrary control strategy $\mathscr{U}$ which is a sequence of (randomized, generally speaking) rules $u_t : \tau_t \to U$ with prehistory sets τ_t of all possible values of sequences $\{z_t, z_s, u_s, \xi_s \,|\, s = \overline{1,t-1}\}$, $t \ge 0$. Given strategy $\mathscr{U}$, define σ-algebras $\mathscr{F}_t = \sigma\{z_t, z_s, u_s, \xi_s \,|\, s = \overline{1,t-1}\}$. Then

$$\begin{aligned} &\mathbb{P}\{z_{t+1} = z(j) \,|\, z_t = z(i), \mathscr{F}_t\} \quad (6)\\ &= \sum_{\ell=1}^{N} \mathbb{P}\{z_{t+1} = z(j) \,|\, z_t = z(i), u_t = u(\ell), \mathscr{F}_t\} \\ &\qquad \cdot \mathbb{P}\{u_t = u(\ell) \,|\, z_t = z(i), \mathscr{F}_t\} \\ &= \sum_{\ell=1}^{N} \pi_{ij}^{\ell} d_t^{i\ell} \quad (7) \end{aligned}$$

where

$$d_t^{i\ell} \triangleq \mathbb{P}\{u_t = u(\ell) \,|\, z_t = z(i), \mathscr{F}_t\} \quad (8)$$

represent the conditional probabilities of control $u_t = u(\ell)$ at instant t under the state $z_t = z(i)$ and the prehistory $\{z_s, u_s, \xi_s \,|\, s = \overline{1,t-1}\}$.

The conditional expectation of losses

$$\begin{aligned} &\mathbb{E}\{\xi_t \,|\, z_t\} = \sum_{i=1}^{K} \mathbf{1}_{\{z_t = z(i)\}} \mathbb{E}\{\xi_t \,|\, z_t = z(i)\} \quad (9)\\ &= \sum_{i=1}^{K} \mathbf{1}_{\{z_t = z(i)\}} \sum_{\ell=1}^{N} \mathbb{E}\{\mathbf{1}_{\{u_t = u(\ell)\}} \\ &\qquad \cdot \mathbb{E}\{\xi_t \,|\, z_t = z(i), u_t = u(\ell)\} \,|\, z_t = z(i)\} \quad (10)\\ &= \sum_{i=1}^{K} \mathbf{1}_{\{z_t = z(i)\}} \sum_{\ell=1}^{N} a_{i\ell} \mathbb{E}\{d_t^{i\ell} \,|\, z_t = z(i)\} \quad (11) \end{aligned}$$

holds with arbitrary control strategy. In particular, a stationary control strategy $\mathscr{U}_{\mathrm{St}}$ (with the stationary state distribution $d \triangleq \|d^{i\ell}\|$) leads to the expectation of losses

$$\begin{aligned} \mathbb{E}\{\xi_t\} &= \mathbb{E}\Big\{\sum_{i=1}^{K} \mathbf{1}_{\{z_t = z(i)\}} \sum_{\ell=1}^{N} a_{i\ell} d^{i\ell}\Big\} \quad (12)\\ &= \sum_{i=1}^{K} p_i(d) \sum_{\ell=1}^{N} a_{i\ell} d^{i\ell} \quad (13)\\ &\triangleq A(d) \quad (14) \end{aligned}$$

where

$$p_i(d) \triangleq \mathbb{P}\{z_t = z(i)\} \quad (15)$$

defines stationary probabilities of the stationary controlled Markov states, the matrix of conditional probabilities

$$d^{i\ell} = \mathbb{P}\{u_t = u(\ell) \,|\, z_t = z(i)\} \quad (16)$$

may be treated as a stationary randomized control strategy $\mathscr{U}_{\mathrm{St}}$, stochastic matrix $d = \|d^{i\ell}\| \in D$,

$$D \triangleq \Big\{ d \,\Big|\, d^{i\ell} \ge 0, \sum_{i=1}^{K} d^{i\ell} = 1 \,(i = \overline{1,K}, \ell = \overline{1,N}) \Big\}. \quad (17)$$

As a consequence of (6)–(7), the stochastic vector $p(d) = (p_1(d), \dots, p_K(d))^{\top}$ solves the stationary distribution equation for the stationary controlled Markov chain that is

$$p(d) = \Pi^{\top}(d) p(d), \quad p(d) \in S_K. \quad (18)$$

Here the transition probability matrix $\Pi(d)$ has the (ij)-entry $\sum_{\ell=1}^{N} \pi_{ij}^{\ell} d^{i\ell}$.

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

$$D_+ \triangleq D \bigcap \Big\{ d = \|d^{i\ell}\| \,\Big|\, d^{i\ell} > 0, (i = \overline{1,K}, \ell = \overline{1,N}) \Big\}. \quad (19)$$

C. Weak regularity assumption

A4. The controlled Markov chain is weakly regular, i.e., for any non-degenerate matrix $d \in D_+$, the related Markov chain having the transition probability matrix $\Pi(d)$ is regular (or, in other words, the state set Z represents a unique ergodic class).

Remark 1: Assumption A4 implies that the Markov chain with the transition matrix $\Pi(d)$ is irreducible for any non-degenerate stationary control strategy $d \in D_+$. Hence, A4 implies the existence of a unique solution $p(d)$ to (18) for any $d \in D_+$ and $p_i(d) > 0$ for all $i = \overline{1,K}$. However, the minimum

$$c_- \triangleq \inf_{d \in D_+} \min_{i = \overline{1,K}} p_i(d)$$

may be zero; this extends the assumption in [9] where positive c_- was assumed and used in the algorithm. □

The idea in designing the randomized control strategy is to minimize the mean loss function $A(d)$ in (14) on set D_+

$$A_{\min} \triangleq \inf_{d \in D_+} A(d). \quad (20)$$

However, the direct minimization problem is non-convex, since function $p(d)$ is non-linear. To cope with this objection we introduce another variables

$$c^{i\ell} \triangleq d^{i\ell} p_i(d), \quad i=\overline{1,K},\ \ell=\overline{1,N}; \tag{21}$$

observe the correctness of their existence on set D_+ mapping the latter onto

$$C_+ \triangleq \Big\{ c=\|c^{i\ell}\| \,\Big|\, c^{i\ell}>0, \sum_{i=1}^K \sum_{\ell=1}^N c^{i\ell}=1, \tag{22}$$

$$\sum_{\ell=1}^N c^{j\ell} = \sum_{i=1}^K \sum_{\ell=1}^N \pi_{ij}^{\ell} c^{i\ell}, \forall (i,j,\ell) \Big\}. \tag{23}$$

Notice that assumption A4 implies the positiveness of all $p_i(d)$ in (21) subject to any $d \in D_+$. Therefore, the matrix mapping (21) which transits d from set D_+ onto C_+ is non-degenerate due to all

$$\sum_{\ell=1}^N c^{i\ell} = p_i(d) > 0, \tag{24}$$

and inverse mapping gives the explicit formulas for $d \in D_+$,

$$d^{i\ell} = c^{i\ell} \Big/ \sum_{k=1}^N c^{ik},\ i=\overline{1,K},\ell=\overline{1,N},\ \ c \in C_+. \tag{25}$$

By the construction under assumption A4 the set C_+ represents a non-empty convex set. Thus, the minimization problem in (20) subject to (18) is equivalent to that of

$$\widetilde{A}(c) \triangleq \sum_{i=1}^K \sum_{\ell=1}^N a_{i\ell} c^{i\ell} \to \inf_{c \in C_+}. \tag{26}$$

III. Main Results

Below we propose the online decision strategy in which, at every step $t+1$, the control action $u_t \in U$ is randomly drawn according to a conditional distribution $d_t = \|d_t^{i\ell}\| \in D$ where

$$d_t^{i\ell} \triangleq \mathbb{P}(u_t = u(\ell) \,|\, z_t = z(i), \mathscr{F}_t), \quad \forall (i,\ell). \tag{27}$$

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

$$\Xi_{t+1}^{i\ell} \triangleq \xi_{t+1} \mathbf{1}_{\{z_t=z(i), u_t=u(\ell)\}} / c_t^{i\ell}, \tag{28}$$

where matrices c_t and d_t correspond to each other by one-to-one mappings (21), (25). Indeed, under stationarity assumption of control strategy $c \equiv c_t$, $d \equiv d_t$, and, for all (i,ℓ),

$$\mathbb{E}\left(\Xi_{t+1}^{i\ell}\right) = \mathbb{E}\left\{ \mathbb{E}\left(\frac{\xi_{t+1} \mathbf{1}_{\{z_t=z(i),u_t=u(\ell)\}}}{c^{i\ell}} \,|\, \mathscr{F}_{t+1} \right) \right\} \tag{29}$$

$$= \mathbb{E}\left\{ \frac{a_{i\ell}}{c^{i\ell}} \mathbf{1}_{\{z_t=z(i),u_t=u(\ell)\}} \right\} = \frac{a_{i\ell}}{c^{i\ell}} d^{i\ell} p_i(d)$$

$$= a_{i\ell} = \frac{\partial \widetilde{A}(c)}{\partial c^{i\ell}}. \tag{30}$$

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

$$\mathbb{E}(\Phi_T) = \frac{1}{T} \sum_{t=1}^T \mathbb{E}(\mathbb{E}(\xi_t \,|\, z_{t-1}, \mathscr{F}_{t-1})) \tag{31}$$

$$= \frac{1}{T} \sum_{t=1}^T \mathbb{E}(A(d_{t-1})). \tag{32}$$

Theorem 1: Let assumptions A1-A2-A3-A4 be satisfied and let the conditional distributions $(d_t)_{t\geq 0}$ be defined by the randomized control algorithm of Section IV with parameters (42), (43), and $\Delta t_s = t_{s+1} - t_s \to \infty$ as $s \to \infty$. Then

$$\sup_{T\to\infty} \lim \sqrt[3]{T} \left(\mathbb{E}(\Phi_T) - A_{\min} \right) \leq O(1) \sigma \left(N \ln(KN) \right)^{1/3} \tag{33}$$

where $O(1)$ stands for an absolute constant. ■

IV. Definition of the randomized strategy

In this section we introduce our online strategy (cf. [6] and [5]). We refer to [11] and [1] for the general idea of mirror descent and its development in non-stochastic optimization, as well as to [12] for the pioneering extension to a stochastic setup.

First we introduce a Gibbs distribution defined by the probability vector

$$G_\beta(z) = [S_\beta(z)]^{-1} \left(e^{-z^{(1)}/\beta}, \ldots, e^{-z^{(m)}/\beta} \right)^\top$$

where $S_\beta(z) = \sum_{j=1}^m e^{-z^{(j)}/\beta}$ for arbitrary fixed $z \in \mathbb{R}^m$ and some parameter $\beta > 0$. We will also use the notation $e_m(k) = (0,\ldots,0,1,0,\ldots,0)^\top$ for vectors in $\mathbb{R}^m$ with 1 on k-th position and 0 elsewhere. Note, that z represents a dual vector variable, see (44) below in the Appendix.

Proposition 1: Notice that the closed set $\overline{C}_+$ in (22) represents a convex polyhedron with m vertices $\overline{c}_k \in \overline{C}_+$ situated in the unique hyperplane, and $(N-1)K < m \leq NK$. It can be parameterized on standard simplex S_m by $\theta = (\theta^{(1)},\ldots,\theta^{(m)})^\top \in S_m$, i.e.

$$\overline{C}_+ = \{ c = \psi(\theta) \,|\, \theta \in S_m \} \tag{34}$$

with function

$$\psi(\theta) \triangleq \sum_{k=1}^m \overline{c}_k \theta^{(k)}, \quad \theta = (\theta^{(1)},\ldots,\theta^{(m)})^\top. \tag{35}$$

■

Remark 2: Calculation of vertices $\overline{c}_k$, $k=\overline{1,m}$, can be seen from the Proposition proof in [9]. □

Function (35) leads to the linear operator

$$\Psi \triangleq \nabla_\theta \psi^\top(\theta) = (\overline{c}_1^\top, \ldots, \overline{c}_m^\top)^\top. \tag{36}$$

Now take into account that Markov chain is a dynamical system. To adequately observe its behavior, we fix decision rules $d_t \equiv d_{t_s}$ of control actions $u_t = u(\ell_t)$ between a priori given sequential instances $t_s < t_{s+1}, s = 0,1,\ldots$, $t_0 = 0$, and change them only at the instances t_s. By assumption A4 constants μ and $\rho > 0$ provide for all $t_s \leq t < t_{s+1}$

$$|\mathbb{P}\{i_t = i \,|\, \mathscr{F}_{t_s}\} - p_i(d_{t_s})| \leq \mu\, e^{-\rho(t-t_s)}, \tag{37}$$

cf. [13]. Natural number $\overline{s}$ defines horizon $T = t_{\overline{s}}$. Thus, we introduce the positive sequences (β_t) and (ε_t) and define the control randomized strategy by the following algorithm.

1) *Fix the initial matrix $c_0 \in C$ and zero dual matrix $\zeta_0 = 0 \in \mathbb{R}^{K\times N}$. Define sequential instances $t_0 = 0$, $t_s < t_{s+1}, s = 0,1,\ldots$.*
2) *For each $s = 0,\ldots,\overline{s}-1$ and $t = t_s,\ldots,t_{s+1}-1$:*
   a) *compute matrices d_{t_s} via c_{t_s} by mapping (25) and apply $d_t \equiv d_{t_s}$ and $c_t \equiv c_{t_s}$ for $t_s \le t < t_{s+1}$; 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 $\ell_t \in \{1,\ldots,N\}$, being distributed according to stochastic vector $(1+\varepsilon_{t_s})(d_{t_s}^{i_t 1},\ldots,d_{t_s}^{i_t N})^\top + \varepsilon_{t_s} N^{-1} e_N^0$;*
   b) *compute a stochastic gradient*

$$\Xi_{t+1} = \frac{\xi_{t+1}}{c_t^{i_t \ell_t}}\, e_K(i_t) e_N^\top(\ell_t); \qquad (38)$$

   c) *applying operator Ψ in (36), update dual variables at time $t+1$ and initial variables at time t_{s+1}*

$$\zeta_{t+1} = \zeta_t + \Xi_{t+1}, \qquad (39)$$

$$c_{t_{s+1}} = \psi\left(G_{\beta_{t_s}}(\Psi \circ \zeta_{t_{s+1}})\right). \qquad (40)$$

3) *At horizon $T = t_{\overline{s}}$, output sequences of states $(z_0,\ldots,z_T)$, control actions $(u_0,\ldots,u_T)$, matrices $(c_0,\ldots,c_T)$ and $(d_0,\ldots,d_T)$, and the observed losses $(\xi_1,\ldots,\xi_{T+1})$ and Φ_T.*

Remark 3: Notice that matrix Ξ_{t+1} in (38) contains a unique nonzero entry. Thus, vectors $\Psi \circ \Xi_{t+1}$ in (38)–(40) can be easily computed by

$$\Psi \circ \Xi_{t+1} = (\overline{c}_1^{i_t \ell_t},\ldots,\overline{c}_m^{i_t \ell_t})^\top \xi_{t+1} / c_t^{i_t \ell_t}, \qquad (41)$$

simplifying real calculations in (39)–(40). However, the presented algorithm explains its understanding structure: at each time t, by obtaining stochastic gradient Ξ_{t+1}, we make a step in the dual space and map into set C_+, by applying the result to transformation $\psi \circ G_{\beta_{t_s}}$ and obtaining $c_{t_{s+1}}$. □

The tuning algorithm parameters (β_t) and (ε_t) are defined as follows: $\forall s = 0,1,\ldots$, $\forall t_s \le t < t_{s+1}$,

$$\beta_t \equiv \beta_{t_s} = \beta_0 (t_s+1)^{2/3}, \quad \varepsilon_t \equiv \varepsilon_{t_s} = \varepsilon_0 (t_s+1)^{-1/3}, \qquad (42)$$

$$\beta_0 = O(1)\frac{\sigma N^{1/3}}{(\ln(NK))^{2/3}}, \quad \varepsilon_0 = O(1)\,(N\ln(NK))^{1/3}. \qquad (43)$$

It is important to note that horizon T is not known in advance. Therefore, the algorithm is completely recursive.

V. Conclusions

We obtained an extension of the results in [9] for the case of weakly regular controlled Markov chain. The asymptotic upper bound has the form $O(T^{-1/3})\sigma(N\ln(KN))^{1/3}$. Therefore, the upper bound dependence on the losses scale parameter σ and on the number of control actions N remain the same as in the regular case [9]; furthermore, its dependence on the horizon T essentially remains the same (up to a logarithmic term $\ln T$ in [9]). Finally, the upper bound has insignificant log-dependence on K. We suppose this is a remarkable feature of the proposed algorithm especially for large values of K.

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

[3] S. Floyd and V. Jacobson. Random early detection gateways for congestion avoidance. *IEEE/ACM Transactions on networking*, 1(4):393–413, 1993.

[4] A. Juditsky, G. Lan, A. Nemirovski, and A. Shapiro. Stochastic approximation approach to stochastic programming. 2007.

[5] 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.

[6] 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.

[7] S.H. Low, F. Paganini, and J.C. Doyle. Internet congestion control. *IEEE Control Systems Magazine*, 1(22):28–43, 2002.

[8] B.M. Miller, K.E. Avrachenkov, K.V. Stepanyan, and G.B. Miller. Flow control as a stochastic optimal control problem with incomplete information. *Problems of Information Transmission*, 41(2):150–170, 2005.

[9] A.V. Nazin and B.M. Miller. Mirror descent algorithm for controlled homogeneous finite markov chains with unknown mean losses. 18th IFAC World Congress, Milano, Italy, 28 Aug – 2 Sept, 2011.

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

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

[12] 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.

[13] Yu. Rozanov. *Probability Theory, Random Processes, and Mathematical Statistics*. Kluwer Academic Publishers, 1995.

[14] R. Srikant. *The mathematics of Internet congestion control*. Birkhauser, 2004.

[15] M. Welzl. *Network Congestion Control : Managing Internet Traffic*. John Wiley & Sons, Ltd., Chichester, 2005.

Appendix

For the convenience of reader, recall the properties of function $G_\beta(\cdot)$ (cf., e.g., [6]). We have $G_\beta(z) = -\nabla W_\beta(z)$,

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

Function W_β and the entropy type function

$$V(\theta) \triangleq \ln N + \sum\nolimits_{j=1}^m \theta^{(j)} \ln \theta^{(j)} \ge 0, \quad \theta \in S_m,$$

are related to each other via convex duality formula:

$$W_\beta(z) = \sup_{\theta \in S_m}\left\{-z^\top\theta - \beta V(\theta)\right\}, \quad z \in \mathbb{R}^m. \qquad (44)$$

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

A. Proof of Theorem 1

Introduce variables $\widetilde{\zeta}_t = \Psi \circ \zeta_t$ and write the algorithm in variables $(\theta, \widetilde{\zeta})$ instead of (c, ζ). Since $\widetilde{\Xi}_t \triangleq \Psi \circ \Xi_t$ represents the stochastic gradient by θ for function $A(\psi(\theta))$ at time $t-1$, equations (39)–(40) are written

$$\widetilde{\zeta}_{t+1} = \widetilde{\zeta}_t + \widetilde{\Xi}_{t+1}, \qquad (45)$$

$$\theta_{t_{s+1}} = G_{\beta_{t_s}}(\widetilde{\zeta}_{t_{s+1}}). \qquad (46)$$

Note that

$$W_{\beta_t}(\widetilde{\zeta}_{t+1}) - W_{\beta_t}(\widetilde{\zeta}_t) = \beta_t \ln\left(\frac{\sum_{k=1}^m e^{-\widetilde{\zeta}_{t+1}^{(k)}/\beta_t}}{\sum_{k=1}^m e^{-\widetilde{\zeta}_t^{(k)}/\beta_t}}\right) = \beta_t \ln(\theta_t^\top v_{t+1})$$

where the k-th entry of vector v_t equals $v_t^{(k)} = e^{-\widetilde{\Xi}_t^{(k)}/\beta_{t-1}}$. Since $e^x \le 1+x+x^2/2$ for $x \le 0$, we get

$$v_t^{(k)} \quad \le \quad 1 - \frac{\widetilde{\Xi}_t^{(k)}}{\beta_{t-1}} + \frac{(\widetilde{\Xi}_t^{(k)})^2}{2\beta_{t-1}^2}.$$

Recalling $\widetilde{\Xi}_{t+1}^{(k)} = \overline{c}_k^{i_t\ell_t}\xi_{t+1}/c_t^{i_t\ell_t}$ by (41), we obtain

$$\theta_t^\top \widetilde{\Xi}_{t+1} = \left(\sum_{k=1}^m \theta_t^{(k)} \overline{c}_k^{i_t\ell_t}\right)\xi_{t+1}/c_t^{i_t\ell_t} = \xi_{t+1},$$

and introducing

$$\widetilde{\eta}_t \triangleq \sum_{k=1}^m \theta_t^{(k)} (\overline{c}_k^{i_t\ell_t})^2$$

we bound

$$\begin{aligned}\beta_t \ln(\theta_t^\top v_{t+1}) &\le \beta_t \ln\left(1 - \frac{\xi_{t+1}}{\beta_t} + \frac{\xi_{t+1}^2 \widetilde{\eta}_t}{2(c_t^{i_t\ell_t})^2\beta_t^2}\right) \\ &\le -\xi_{t+1} + \frac{\xi_{t+1}^2\widetilde{\eta}_t}{2(c_t^{i_t\ell_t})^2\beta_t}. \end{aligned} \tag{47}$$

Note that W_β is monotone decreasing in β, as follows from (44). Using this, taking expectation of both sides of (47) (first over ℓ_t, conditional on i_t and c_t, then taking the full expectation) and applying assumption A2 we obtain

$$\begin{aligned}&\mathbb{E}\left(W_{\beta_{t+1}}(\widetilde{\zeta}_{t+1}) - W_{\beta_t}(\widetilde{\zeta}_t)\right) \\ \le\ & -\mathbb{E}(\xi_{t+1}) + \tfrac{\sigma^2}{2\beta_t}\,\mathbb{E}\left(\tfrac{\widetilde{\eta}_t}{(c_t^{i_t\ell_t})^2}\right).\end{aligned} \tag{48}$$

The latter expectation in RHS (48) is bounded:

$$\mathbb{E}\left(\frac{\widetilde{\eta}_t}{(c_t^{i_t\ell_t})^2}\right) = \mathbb{E}\left\{\mathbb{E}\left(\frac{\widetilde{\eta}_t}{(c_t^{i_t\ell_t})^2}\,|\,i_t, \mathscr{F}_t\right)\right\} \tag{49}$$

$$\begin{aligned} &= \mathbb{E}\left(\sum_{\ell=1}^N \frac{1+O(\varepsilon_t)}{\sum_{\ell'=1}^N c_t^{i_t\ell'}} \cdot \frac{\sum_{k=1}^m \theta_t^{(k)}\left(\overline{c}_k^{i_t\ell}\right)^2}{\sum_{k=1}^m \theta_t^{(k)}\overline{c}_k^{i_t\ell}}\right) \\ &\le \frac{(1+O(\varepsilon_t))N}{c_- + O(\varepsilon_t)} \max_{i,\ell,k} \overline{c}_k^{i\ell} \le \frac{O(1)N}{\varepsilon_t}. \end{aligned} \tag{50}$$

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

$$\sum_{t=1}^T \mathbb{E}(\xi_t) \quad \le \quad -\mathbb{E}W_{\beta_T}(\widetilde{\zeta}_T) + \sum_{t=0}^{T-1} \frac{O(1)N\sigma^2}{\beta_t\varepsilon_t}. \tag{51}$$

The minimizer $\theta^* \triangleq \operatorname*{arg\,min}_{\theta\in S_m} \widetilde{A}(\psi(\theta))$ satisfies $\widetilde{A}(\psi(\theta^*)) = \inf_{c\in C_+}\widetilde{A}(c) = A_{\min}$ due to (26). The following idea is to apply (44) and use inequality

$$W_{\beta_T}(\widetilde{\zeta}_T) \ge -\widetilde{\zeta}_T^\top\theta^* - \beta_T V(\theta^*). \tag{52}$$

Therefore, cf. (28)–(30) and (37), (41), and for $t_s \le t < t_{s+1}$,

$$\begin{aligned}\sum_{k=1}^m \mathbb{E}\left(\theta^{*(k)}\mathbb{E}\{\widetilde{\Xi}_{t+1}^{(k)}|\mathscr{F}_{t_s}\}\right) &= \mathbb{E}\sum_{k=1}^m \theta^{*(k)} \sum_{\ell=1}^N \frac{a_{i\ell}\,(1+O(\varepsilon_t))\,\overline{c}_k^{i_t\ell}}{\sum_{\ell'=1}^N c_t^{i_t\ell'}} \\ &= \mathbb{E}\sum_{\ell=1}^N\sum_{i=1}^K \frac{a_{i\ell}\,c^{*i\ell}}{\sum_{\ell'=1}^N c_{t_s}^{i\ell'}}\,\mathbb{P}\{i_t = i\,|\,\mathscr{F}_{t_s}\}\,(1+O(\varepsilon_t)) \\ &\le A_{\min} + \sigma O(\varepsilon_t) + \mathbb{E}\sum_{i=1}^K\sum_{\ell=1}^N a_{i\ell}\,c^{*i\ell}\left|\frac{\mathbb{P}\{i_t=i\,|\,\mathscr{F}_{t_s}\}}{\sum_{\ell'=1}^N c_{t_s}^{i\ell'}} - 1\right| \\ &\quad + \mathbf{1}_{\{t=t_{s+1}\}}\,\mathbb{E}\sum_{i=1}^K\sum_{\ell=1}^N a_{i\ell}\,c^{*i\ell}\,\frac{\sum_{\ell'=1}^N |c_{t_{s+1}}^{i\ell'} - c_{t_s}^{i\ell'}|}{\sum_{\ell'=1}^N c_t^{i\ell'}} \\ &\le A_{\min} + \sigma O(\varepsilon_t) + \sigma\mu\, e^{-\rho(t-t_s)}O(\varepsilon_t^{-1}) \qquad (53) \\ &\quad + \mathbf{1}_{\{t=t_{s+1}\}}\mathbb{E}\|\theta_{t_{s+1}} - \theta_{t_s}\|_1 \sigma O(\varepsilon_t^{-1}). \qquad (54)\end{aligned}$$

The last term with 1-norm is bounded by (L)-property [6]

$$\|\nabla W_\beta(z) - \nabla W_\beta(z')\|_1 \le \beta^{-1}\|z - z'\|_\infty$$

and the following formula, for $0 < \beta \le \beta'$,

$$\|\nabla W_\beta(z) - \nabla W_{\beta'}(z)\|_1 \le \beta^{-2}\|z\|_\infty(\beta' - \beta).$$

So, we set $\Delta\beta_{t_s} \triangleq \beta_{t_{s+1}} - \beta_{t_s}$ and get by (46)

$$\begin{aligned}\|\theta_{t_{s+1}} - \theta_{t_s}\|_1 &\le \|\nabla W_{\beta_{t_s}}(\widetilde{\zeta}_{t_{s+1}}) - \nabla W_{\beta_{t_{s-1}}}(\widetilde{\zeta}_{t_s})\|_1 \\ &\le \frac{\Delta\beta_{t_s}}{\beta_{t_s}^2}\|\widetilde{\zeta}_{t_s}\|_\infty + \frac{N\sigma\,\Delta t_s}{O(\varepsilon_{t_s})\beta_{t_s}}.\end{aligned}$$

Using (44), (51)–(54), the fact that $\sup_{\theta\in S_m} V(\theta) = \ln m$, and the last display we obtain

$$\begin{aligned}\mathbb{E}W_{\beta_T}(\widetilde{\zeta}_T) &\ge -\mathbb{E}(\widetilde{\zeta}_T^\top\theta^*) - \beta_T\ln m \\ &= -\sum_{t=0}^{T-1}\sum_{k=1}^m \mathbb{E}\left(\widetilde{\Xi}_{t+1}^{(k)}\theta^{*(k)}\right) - \beta_T\ln m \\ &\ge -TA_{\min} - \beta_T\ln m - \sum_{s=0}^{\overline{s}-1}\sigma O(\varepsilon_{t_s})\left(\Delta t_s + \frac{\sigma\,\mu}{1-e^{-\rho}}\right) \\ &\quad - O(1)N\sigma^2\sum_{s=0}^{\overline{s}-1}\frac{1}{\varepsilon_{t_s}}\left(\frac{\Delta\beta_{t_s}}{\beta_{t_s}^2}t_s + \frac{\Delta t_s}{\beta_{t_s}}\right),\end{aligned}$$

and, by applying the algorithm parameters (42) and evaluating the sums by the integrals, we get

$$\mathbb{E}(\Phi_T) - A_{\min} \le O(1)T^{-1/3}\left(\beta_0\ln m + \sigma\varepsilon_0 + \frac{N\sigma^2}{\varepsilon_0\beta_0}\right).$$

The result of the theorem easily follows by optimizing parameters β_0 and ε_0 in RHS. ▲