Mirror Descent Algorithm for Homogeneous Finite Controlled Markov Chains with Unknown Mean Losses *

Alexander V. Nazin∗ Boris M. Miller **

Abstract: We consider the adaptive stochastic problem for a system described by a controlled Markov chain (CMC) with a finite number of states. The novelty of the approach consists in adaptation technique for optimization of the system with unknown distribution of the cost function. This approach is applicable to the Internet congestion control with active users, motivating our study. In fact, we consider the homogenous finite CMC, introduce the assumptions of processes independence, stationarity of positive random losses and the bounds for their second moments, and the model regularity. The uncertainty is that the mean loss matrix is unknown. The related convex stochastic optimization problem as well as ideas of Mirror Descent Algorithm (MDA) give us the ability to design the control algorithm and to prove the upper bound for the difference between the expectation of time-averaged losses and their theoretical minimum. Asymptotically, the upper bound equals \( \left( \frac{\ln T}{T} \sigma^2 N \ln(NK) \right)^{1/3} \) and weakly depends on the state number \( K \) (standing inside the logarithm); here \( N \) and \( T \) stand for the number of control actions and the time horizon, respectively, parameter \( \sigma^2 \) bounds the conditional second moments of the losses.

1. INTRODUCTION

Controlled Markov chains (CMC) play an important role in the modelling and optimization of stochastic systems [Elliott et al. (1995)]. 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 ordinary differential equations (dynamic programming equation [Miller (2009b)]), 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 Markov chain. This situation occurs very rarely in real systems and in this case it becomes necessary to simplify the model, for example, by representation of unknown state by a finite number of possible ones like in the theory of Hidden Markov Models [Elliott et al. (1995)].

Typical example of the situation where the complete information is not available, is the example of the Internet Congestion Control [Floyd and Jacobson (1993)], [Srikant (2004)], [Welzl (2005)]. Usually it is assumed that the state of a communication link is described by a controlled hidden Markov process with a finite state set, while the loss flow is described by a counting process with intensity depending on a current transmission rate and an unobserved link state. The control is the transmission rate, and it has to be chosen as a nonanticipating process depending on the observation of the loss process. The aim of the control is to achieve the maximum of some utility function that takes into account losses of the transmitted information. Originally, the problem belongs to the class of stochastic control problems with incomplete information; however, optimal filtering equations that provide estimation of the current link state based on observations of the loss process allow one to reduce the problem to a standard stochastic control problem with full observations [Miller et al. (2005)]. Another important approach uses the idea of active users which choose the transmission rate as a function of the rejection probability established by a provider at current time and state of the transmission line. This case is typical for so called RED (Random Early Detection) transmission protocols [Low et al. (2002)], where the provider assigns the profile of the package rejection probability in order to prevent congestion and at

* The work of the first author was supported in part by Russian Basic Research Foundation Grants 10-08-01068 and by Australian Research Council Grant DP0988685. His work was partially carried out during his stay in March–April, 2011 at the School of Mathematical Sciences, Monash University, Clayton, 3800, Victoria, Australia.

** The work of the second author was supported in part by Australian Research Council Grant DP0988685 and by Russian Basic Research Foundation Grant 10-01-00710.

Copyright by the International Federation of Automatic Control (IFAC)
the same time to keep the reasonable transmission rate. If the probability of the package rejection is known at the user side, one can control the transmission rate very effectively [Miller (2009a)]. However, the profile of the rejection probability is unknown at the user side, where it is possible to observe only the number of rejected packages, thereby at the user side there exists only the possibility to estimate the average effectiveness and to adjust the transmission rate to unknown probability of rejection.

The idea of this work is to provide the methodology of the optimization of controlled Markov chains which is based on adaptation rather than on stochastic optimization, which requires more a priori inaccessible information. Specifically, we develop the approach to stochastic setup for homogeneous controlled finite Markov chains with unknown mean losses which was first introduced in Nazin and Poznyak (1986) (cf. Poznyak et al. (2000)). We assume that given a control action, the matrix of state transition probabilities is known a priori, but the current random losses are statistically undefined. Frankly speaking, it narrows our set up in contrary to Nazin and Poznyak (1986) (where the matrices of state transition probabilities have been unknown there in), but we concentrate ourselves on the difficulties of high dimension. The aim here is to extend the approach of Juditsky et al. (2008), where the set up can be represented as a degenerate case of Markov chains with a unique/fixed state. Under assumption of nonnegative losses and finite variances we demonstrate that the explicit expected excess bound of losses for any (but large enough) horizon T implies the convergence rate of the order \( \left( \frac{\ln T}{T} \right)^{2N \ln(NK)} \) and weakly depends on the state number K (standing inside the logarithm); here N and T stand for the number of control actions and the time horizon, respectively, parameter \( \sigma^2 \) bounds the conditional second moments of the losses.

Notice that our extension mentioned above is far from being straightforward! We reformulate the initial stochastic optimization problem then we propose and study a Mirror Descent Algorithm (MDA) in the spirit of Juditsky et al. (2007).

2. STATEMENT OF PROBLEM

This section is essentially adopted from chapter 5 in Nazin and Poznyak (1986).

First, introduce the notation for standard simplex in \( \mathbb{R}^m \),

\[ S_m \triangleq \left\{ (x_1, \ldots, x_m) \in \mathbb{R}^m \mid x_i \geq 0, \sum_{i=1}^m x_i = 1 \right\}. \]

2.1 Preliminary assumptions

Let \( (\Omega, \mathcal{F}, \mathbb{P}) \) be a given probability space. Let stochastic control system be modeled as a homogeneous controlled Markov chain where

- the set of states \( Z \triangleq \{z(1), \ldots, z(K)\}, K \geq 2 \), and a system state \( z_t \in Z \) at current time \( t \in \{0, 1, \ldots\} \) are observable;
- the given set \( U \triangleq \{u(1), \ldots, u(N)\} \) represents the set of possible control inputs, \( N \geq 2 \);
- the transition probabilities of the state \( z_t \in Z \) at current time \( t \in \{0, 1, \ldots\} \) to the next state \( z_{t+1} \in Z \) under the applied control \( u_t \in U \) are given by the given conditional probabilities: \( \forall t \),

\[ \mathbb{P}\{z_{t+1} = z(j) \mid z_t = z(i), u_t = u(\ell), \mathcal{F}_t\} = \pi_{ij}^{\ell} \] (1)

here \( \sigma \)-algebra \( \mathcal{F}_t \) is generated by the prehistory of observations up to time \( t \); in other words, equation (1) relates to the system Markov property;
- the losses \( \xi_t \triangleq \xi_t(z_t, u_t, \omega) \) at current time \( t \in \{0, 1, \ldots\} \) are observable and statistically depend only on the state \( z_t \) and the applied control \( u_t \) with unknown conditional distributions; the random variables \( \xi_t \triangleq \xi_t(z_t, u_t, \omega) \) form i.i.d. sequences by time \( t \) for all \( i \in \{1, K\} \) and \( t \in \{1, N\} \);
- the time-mean losses \( \Phi_t \) on time interval \( \{1, T\} \) are defined by

\[ \Phi_t = \frac{1}{t} \sum_{s=1}^t \xi_s. \]

Introduce further assumptions:

A1. For each \( t = 1, 2, \ldots \) the totalities \( \{\xi_t(z, u, \omega) \mid z \in Z, u \in U\} \) and \( \{\xi_t(z, u, \omega) \mid z, u_k \in Z, u \in U, \sigma = 1, t, t-1, k = 1, K\} \) are independent.

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

\[ \mathbb{E}\{\xi_t(z, u(\ell), \omega)\} \leq a_{i\ell} \quad \forall t. \]

A3. The second moments of losses are a.s. bounded by constant \( \sigma^2 \), i.e.

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

2.2 General control strategies

Consider an arbitrary control strategy \( U \) which is a sequence of (randomized, generally speaking) rules \( u_t : \tau \rightarrow U \) with prehistory sets \( \tau_t \) of all possible values of sequences \( \{z_t, z_s, u_s, \xi_s \mid s = 1, T-1\}, t \geq 0 \). Formally, our control optimization problem is as follows:

\[ \lim_{t \to \infty} \sup \mathbb{E}\Phi_t = \inf_{U} \mathbb{E}\Phi_t \]

(2)

where infimum is over the class of all the control strategies. Besides, given strategy \( U \), define \( \sigma \)-algebras \( \mathcal{F}_t = \sigma\{z_t, z_s, u_s, \xi_s \mid s = 1, T-1\} \). Then

\[ \mathbb{P}\{z_{t+1} = z(j) \mid z_t = z(i), \mathcal{F}_t\} \]

\[ = \sum_{\ell=1}^N \mathbb{P}\{z_{t+1} = z(j) \mid z_t = z(i), u_t = u(\ell), \mathcal{F}_t\} \]

\[ = \sum_{\ell=1}^N \pi_{ij}^{\ell} d_{ij}^{\ell} \]

(3)

(4)

where

\[ d_{ij}^{\ell} \triangleq \mathbb{P}\{u_t = u(\ell) \mid z_t = z(i), \mathcal{F}_t\} \]

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

Denote a characteristic function \( 1_{(i)} \) of a given event. The conditional expectation of losses

\[ \mathbb{E}\{\xi_t \mid z_t = z(i)\} = \sum_{i=1}^K \mathbb{E}\{\xi_t \mid z_t = z(i)\} \]
where

\[ p_i(d) \triangleq \mathbb{P}(z_t = z(i)) \]

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

\[ d_{i\ell} = \mathbb{P}(u_t = u(\ell) \mid z_t = z(i)) \]

may be treated as a stationary randomized control strategy \( U_{St} \), stochastic matrix \( d \triangleq \|d_{i\ell}\| \in D \),

\[ D \triangleq \left\{ d \mid d_{i\ell} \geq 0, \sum_{i=1}^{K} d_{i\ell} = 1 \right\}. \]

As a consequence of (3)–(4), the stochastic vector \( p(d) = (p_1(d), \ldots, p_K(d))^T \) solves the stationary distribution equation for the stationary controlled Markov chain that is

\[ p(d) = \Pi^T(d)p(d), \quad p(d) \in S_K. \]

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

2.4 Regularity assumptions and optimization problem

A4. We assume that the controlled Markov chain is regular [Sragovich (2006)], i.e. the transition matrix \( \Pi(d) \) has a single ergodic class without any cyclic subclasses for any \( d \in D \).

Remark 1. Assumption A4 implies that the matrix \( \Pi(d) \) is indecomposable and aperiodic for any nondegenerate stationary control strategy \( d \in D \). Besides, A4 implies the existence of a unique solution \( p(d) \) to (6) for any \( d \in D \), and \( p_i(d) > 0, i = 1, K \) [Sragovich (2006)].

The idea to design control strategy is to minimize function \( A(d) \) in (5) on set \( D \)

\[ A_{\min} \triangleq \min_{d \in D} A(d). \]

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

\[ c^\ell \triangleq d_{i\ell} p_i(d), \quad i = 1, K, \quad \ell = 1, N; \]

observe the correctness of their existence on set \( D \) mapping the latter onto

\[ C \triangleq \left\{ c = \|c^\ell\| \mid c^\ell \geq 0, \sum_{i=1}^{K} \sum_{\ell=1}^{N} c^\ell = 1 \right\}, \]

\[ \sum_{\ell=1}^{N} c^\ell = \sum_{i=1}^{K} \sum_{\ell=1}^{N} \pi_{i\ell} c^\ell, \forall (i, \ell) \}. \]

Notice that the assumption A4 implies the positiveness of all \( p_i(d) \) in (8) subject to any \( d \in D \). Therefore, the matrix mapping (8) which transits \( d \) from set \( D \) onto \( C \) is non-degenerate due to all

\[ \sum_{\ell=1}^{N} c^\ell = p_i(d) > 0, \quad (10) \]

and inverse mapping gives the explicit formulas for \( d \in D \),

\[ d^\ell = c^\ell / \sum_{k=1}^{N} c^k, \quad i = 1, K, \quad \ell = 1, N, \quad c \in C. \]

By the construction under assumption A4 the set \( C \) represents a non-empty convex set. Thus, the minimization problem in (7) subject to (6) is equivalent to that of

\[ \tilde{A}(c) \triangleq \sum_{i=1}^{K} \sum_{\ell=1}^{N} a_{i\ell} c^\ell \rightarrow \min_{c \in C} \tilde{A}(c). \]

By continuity, we introduce constant \( c_- \) in (10),

\[ \sum_{\ell=1}^{N} c^\ell = p_i(d) \geq \min_{i=1, K} \min_{d \in D} p_i(d) \triangleq c_- > 0. \]

If all the entries \( a_{i\ell} \) are known, the solution of the optimal control problem in (2) may be found directly. Instead, in the case of unknown \( a_{i\ell} \) we will find the randomized adaptive strategy which in the limit gives the value of the criterion close to the minimum. Moreover, we prove the asymptotic upper bound with the explicit rate of convergence (see Theorem 1 below).

3. MAIN RESULT

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

\[ d_{i\ell}^t \triangleq \mathbb{P}(u_t = u(\ell) \mid z_t = z(i), F_t), \quad \forall (i, \ell). \]

The sequence of distributions \( d_t, t = 0, 1, \ldots \) of randomized strategies is given by the algorithm described in Section 4 below and uses stochastic gradient for \( \tilde{A}(c) \), i.e. the random matrix with entries

\[ \Xi_{t+1} = \Xi_t + 1 \{z_t = z(i), u_t = u(\ell)\} / c_{i\ell}^t, \]

where matrices \( |c_{i\ell}| \) and \( \|d_{i\ell}\| \) correspond to each other by one-to-one mappings (8), (11), as it follows from the algorithm description all \( c_{i\ell}^t > 0 \). Indeed, this is the stochastic gradient since under stationarity assumption of the control strategy, we have for all \( (i, \ell) \)

\[ \mathbb{E}(\Xi_{t+1}^\ell) = \mathbb{E}\left( \frac{\Xi_t^\ell 1 \{u_t = u(\ell)\}}{c_{i\ell}^t} \right) \mid z_t = z(i), F_t \]

\[ = \mathbb{E}(\frac{\Xi_t^\ell + d_t^\ell p_i(d_t)}{c_{i\ell}^t} \mid z_t = z(i), F_t) \]

\[ = a_{i\ell} = \frac{\partial A(c)}{\partial c_{i\ell}}. \]

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

\[ \mathbb{E}(\Phi_T) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(\mathbb{E}(\xi_t \mid z_{t-1}, F_{t-1})) \]

\[ = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(A(d_{t-1})). \]
Proposition 1. By the regularity assumption A4, the set \( C \) in (9) represents a polyhedron with \( m \) vertices \( \theta_k \in \mathbb{R}^{(m)} \), with \( K > 1 \) diminishes the rate of convergence from \( T^{-1/2} \) of Juditsky et al. (2008) to that of \( T^{-1/3} \) (up to a logarithmic factor).

Remark 2. Calculation of vertices \( \theta_k \), \( k = 1, \ldots, m \), can be seen from the Proposition proof, see the Appendix.

Function (12) leads to the linear operator
\[
\Psi \triangleq \nabla \theta \theta^T (\theta) = (\theta^T_1, \ldots, \theta^T_m)^T.
\]

Now take into account that Markov chain is a dynamical system. To adequately observe its properties, we fix decision rules \( d_t \equiv d_{t,s} \) of control actions \( u_t = u(\ell_t) \) between a priori given sequential instances \( t_s < t_{s+1}, s \geq 0, 1, \ldots, t_0 = 0 \), and change them only at the instances \( t_s \). By assumption A4 constants \( \rho, \mu, c \) and \( \alpha > 0 \) provide for all \( t_s < t < t_{s+1} \)
\[
|\mathbb{P}[i_t = i | \mathcal{F}_{t_s}] - \alpha_i (d_t)\| \leq \mu e^{-\rho(t-t_s)},
\]

Theorem 1. Let assumptions A1–A4 be satisfied and let the conditional distributions \( (d_t)_{t \geq 0} \) be defined by the algorithm of Section 4 with parameters (20), (21). Then
\[
\limsup_{T \to \infty} \frac{1}{T} \mathbb{E}[\Phi_T - A_{\text{min}}] \leq \left(N \sigma^2 \ln (NK)\right)^{1/3}.
\]

Remark 3. The non-asymptotic upper bound (A.8) obtained at the very end of the theorem proof weakly depends on the state set number \( K \) which is a remarkable feature of the algorithm. However, it depends on all the problem parameters involved, i.e. \( \rho, \mu, c, \sigma \). For the case of non-asymptotic optimal algorithm parameters \( \beta \equiv \beta_\infty \) and \( \overline{\sigma} \) lead to optimal convergence rate \( O(T^{-1/3}) \); unfortunately, it depends on those problem parameters which is less of our interests.

4. DEFINITION OF THE STRATEGY

In this section we introduce the adaptive strategy (cf. Juditsky et al. (2005) and Juditsky et al. (2008)). We refer to Nemirovski and Juditsky (1985) and Ben-Tal and Nemirovski (1999) for the general idea of mirror descent and its development in non-stochastic optimization, as well as to Nesterov (2005) for the pioneering extension to a stochastic setup.

First we introduce a Gibbs distribution defined by the probability vector
\[
G_\beta(z) = [S_\beta(z)]^{-1} \left(e^{-z^{(1)}/\beta}, \ldots, e^{-z^{(m)}/\beta}\right)^T,
\]
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 \( e_m(k) = (0, \ldots, 0, 1, 0, \ldots, 0) \) for vectors in \( \mathbb{R}^m \) with 1 on \( k \)-th position and 0 elsewhere. Note, that \( z \) represents a dual vector variable, see (A.2) below in the Appendix.

Thus, we introduce a natural number \( \overline{s} \) by horizon \( T = t_\overline{s} \), a positive sequence \( (\beta_t) \), and define control strategy by the following algorithm.

1. Fix the initial matrix \( c_0 \in C \), compute matrix \( d_0 \) via \( c_0 \) by mapping (11), and put conjugate matrix \( \zeta_0 = 0 \in \mathbb{R}^{NK} \). Define sequential instances \( t_0 = 0, t_s < t_{s+1}, s = 0, 1, \ldots, s < 1 \) with \( t_\overline{s} = T \).
2. For each \( s = 0, \ldots, \overline{s} - 1 \) and \( t = t_s, t_s + 1 \):
   a. having the observed state \( z_t = z(t) \), draw control action \( u_t = u(\ell_t) \) with random \( \ell_t \in \{1, \ldots, N\} \), being distributed according to stochastic vector \( (d_{t_s}^{1}, \ldots, d_{t_s}^{TN})^T \).
   b. compute the stochastic gradients
\[
\Xi_{t_\overline{s}+1} = \frac{\xi_{t_\overline{s}+1}}{c_{t_\overline{s}}} e_K(\ell_t) e_N^T(\ell_t) \quad (15)
\]
and their sum over \( t_s < t < t_{s+1} \) that is
\[
\Xi_{t_\overline{s}+1} = \sum_{t=t_s}^{t_{s+1}-1} \Xi_t; \quad (16)
\]
   c. applying \( \Psi \) and \( \psi \) in (13) and (12), update
\[
\zeta_{t_\overline{s}+1} = \zeta_{t_\overline{s}} + \Xi_{t_\overline{s}+1}; \quad (17)
\]
   d. at time \( t = t_{s+1} \) define \( d_{t_{s+1}} \) via \( c_{t_{s+1}} \) by (11).
3. At horizon \( T = t_\overline{s} \), output matrix \( d_T \) and the obtained mean losses \( \Phi_T \). It finalizes the algorithm.

Remark 4. Notice that matrix \( \Xi_{t_\overline{s}+1} \) in (15) contains a unique nonzero entry. Thus, vectors \( \Psi \circ \Xi_{t_\overline{s}+1} \) in (15)–(18) can be easily computed by
\[
\Psi \circ \Xi_{t_\overline{s}+1} = \sum_{t=t_s}^{t_{s+1}} (\tau_{t}^{c_t} e_1, \ldots, \tau_{t}^{c_t} e_N \tau_{t+1}^{c_t t_s}) \tau_{t+1}^{c_t t_s}, \quad (19)
\]
simplifying calculations in (17)–(18). However, the presented algorithm explains its structure: at each time \( t_s+1 \), by obtaining stochastic gradient \( \Xi_{t_s+1} \), we make a step in the dual space and map into set \( C \), by applying the result to transformation \( \psi (G_\beta(\cdot)) \) and obtaining \( c_{t_s+1} \).

The tuning parameter \( \beta_\overline{s} \) involved into the algorithm is defined for Theorem 1 as follows: for each \( s = 0, 1, \ldots, \overline{s} - 1 \) and \( t = t_s, \ldots, t_{s+1} - 1 \),
\[
\beta_s \equiv \left. \left( N \sigma^2 T^2 \ln T \right)^{1/3} \right|_{T = \overline{s}}, \quad \Delta t_s = \frac{T}{\overline{s}}, \quad (20)
\]
and
\[
\overline{s} = \left( N \sigma^2 T^2 T^2 \ln T \right)^{1/3}, \quad \overline{t}_s = \left( \frac{T}{\overline{s}} \right)^{1/3}, \quad (21)
\]

Remark 5. The algorithm parameters (20)–(21) depend on the horizon \( T \). However, one may prove Theorem 1 with a horizon-free parameters
\[
\beta_s \sim \left( t_s^2 \ln t_s \right)^{1/3}, \quad \Delta t_s \sim t_s^{1/3},
\]
when the algorithm becomes completely recursive.

5. CONCLUSIONS

The problem considered above and the proved Theorem can be treated as an extension to those of Juditsky et al. (2008). Indeed, a multi-armed bandit problem may be treated as a trivial controlled Markov chain having the unique state; however, the proved inequality (14) with \( K > 1 \) diminishes the rate of convergence from \( T^{-1/2} \) of Juditsky et al. (2008) to that of \( T^{-1/3} \) (up to a logarithmic
term ln T). It looks nice in the asymptotic result (14) that its RHS weakly depends on the state number K which is standing inside the logarithm. From our point of view this is the principal advantage of the proposed algorithm.

REFERENCES


Appendix A. PROOFS

A.1 Proof of Proposition 1

First, observe that K linear system equations in (9) is degenerate, since they sum to identity 1 ≡ 1. For each j, the related equation in (9) may be rewritten as

$$\sum_{t=1}^{N} (1 - \pi_{jj}^t) c_{jl} = \sum_{t=1}^{K} \sum_{t=1}^{N} \pi_{ij}^t c_{jl},$$  \hspace{1cm} (A.1)

by separating the variables $c_{jl}^t$ in LHS and those of $c_{jl}^t$ in RHS. At least one coefficient in LHS is positive: otherwise, $\pi_{jj}^t = 1$ for all $\ell$ imply zero RHS for any $c \in C$, therefore, $\pi_{ij}^t = 0$ for all $i \neq j$ and all $\ell$, which contradict A4. Thus, for all $j \in \{1, K\}$, there exist $\ell_j \in \{1, K\}$ such that $\pi_{jj}^t < 1$, and equations (A.1) are equivalent to

$$c_{jj}^t = \sum_{i \neq j}^{K} \sum_{t=1}^{N} \pi_{ij}^t c_{jl} - \sum_{\ell \neq \ell_j}^{K} \sum_{t=1}^{N} \pi_{jj}^t c_{jl}.$$  \hspace{1cm} (A.2)

This eliminate variables $c_{jj}^t, \ldots, c_{Kn}$ from entries of matrix $\|c_{jl}^t\|$, and all the rest variables parameterize set C by leaving the only hyperplane in (9). Since C is closed convex set, this easily ends the proof.

Now, recall some properties of the function $G_\delta(z)$ (cf., e.g., Juditsky et al. (2005)). We have $G_\delta(z) = -\nabla W_\delta(z)$ where

$$W_\delta(z) = \beta \ln \left( \frac{1}{N} \sum_{t=1}^{N} e^{-z^T \theta / \beta} \right), \hspace{1cm} z \in \mathbb{R}^m.$$  \hspace{1cm} (A.3)

Function $W_\delta$ and the entropy type function $V(\theta) \equiv \ln N + \sum_{j=1}^{m} \theta^{(j)} \ln \theta^{(j)} \geq 0, \hspace{1cm} \theta \in S_m$, are related to each other via convex duality formula:

$$W_\delta(z) = \sup_{\theta \in S_m} \{-z^T \theta - \beta V(\theta)\} \hspace{1cm} (z \in \mathbb{R}^m).$$  \hspace{1cm} (A.4)

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

A.2 Auxiliary Lemma

The following lemma holds under Theorem 1.

**Lemma 1.** Let the assumptions of Theorem 1 fulfill. Then, for each $s = 0, \ldots, s-1$, $i = 1, \ldots, K$, $t = 1, \ldots, N$, and $k = 1, \ldots, m$, the following hold true for (16) a.s.:

$$E \left\{ \left( \Psi \circ \Xi_t \right)^{(k)} | F_t \right\} = \sum_{\ell = 1}^{\ell_j - 1} E \left\{ \frac{\pi_{ij}^t}{c_{ij}^t} \xi_{\ell+1} | F_t \right\},$$  \hspace{1cm} (A.3)

$$E \left\{ \left( \Psi \circ \Xi_t \right)^{(k)} | F_t \right\} \leq \frac{\sigma^2 \Delta_t}{c_{-}} \sum_{\ell = 1}^{\ell_j - 1} E \left\{ \sum_{t=1}^{N} \frac{\pi_{ij}^t}{c_{ij}^t} | F_t \right\}$$  \hspace{1cm} (A.4)

Proof: Equation (A.3) follows directly from (19). The evaluations

$$E \left\{ \left( \Psi \circ \Xi_t \right)^{(k)} | F_t \right\}$$  \hspace{1cm} (A.5)
≤ Δₜσ² \sum_{t=tₜ}^{tᵣ+1} \mathbb{E} \left\{ \left( \frac{\sigma_{\xi_{t+1}}}{c_{l_{t,1}}/2} \right)^2 F_{tₜ} \right\} ≤ Δₜσ² \sum_{t=tₜ}^{tᵣ+1} \mathbb{E} \left\{ \left( \frac{\sigma_{\xi_{t+1}}}{c_{l_{t,1}}/2} \right)^2 F_{tₜ} \right\} \]

finalizes the proof.

### A.3 Proof of Theorem 1

We first treat general case of monotone nondecreasing sequence parameter (βₖ), i.e. βₖ₊₁ ≥ βₖ, s ≥ 0.

Introduce variables \( \tilde{\zeta}_t = Ψ \circ \zeta_t \) and write the algorithm in variables \((\theta, \tilde{\zeta})\) instead of \((c_t, \zeta_t)\). Since \( \Xi \tilde{\zeta}_t = Ψ \circ \Xi \zeta_t \) represent the stochastic gradient by θ for function \( A(\psi(\theta)) \) at time \( t = 1 \), equations (17)–(18) are written

\[
\begin{align*}
\tilde{\zeta}_{t+1} &= \tilde{\zeta}_t + \Xi \zeta_{t+1}, \\
\theta_{t+1} &= G_\beta(\tilde{\zeta}_{t+1}),
\end{align*}
\]

(A.5)

where \( \Xi \zeta_{t+1} = Ψ \circ \Xi \zeta_{t+1} \), see (16) and (19). Note that

\[
W_\beta(\tilde{\zeta}_{t+1}) - W_\beta(\tilde{\zeta}_t) = \beta_s \ln \left( \frac{\sum_{k=1}^{m} e^{-\tilde{\zeta}_{t+1}(k)/\beta_s}}{\sum_{k=1}^{m} e^{-\tilde{\zeta}_t(k)/\beta_s}} \right) = \beta_s \ln (\theta_{t,1} v_{t+1}).
\]

k-th entry of vector \( v_{t+1} \) equals \( v_{t,1} = e^{-\tilde{\zeta}_{t+1}(k)/\beta_s} \). Since \( e^x ≥ 1 + x + x²/2 \) for \( x ≤ 1 \), and \( x > 0 \) for \( x ≤ 1 \), we get

\[
v_{t,1} ≤ 1 - \frac{\Xi \zeta_{t+1} + (\Xi \zeta_{t+1})}{\beta_s}.
\]

Applying (A.3) and (A.4), we obtain

\[
\mathbb{E} \left\{ \theta_{t,1} \Xi \zeta_{t+1} \mid F_{tₜ} \right\} = \sum_{tₜ}^{tᵣ+1} \mathbb{E} \left\{ \left( \frac{\sigma_{\xi_{t+1}}}{c_{l_{t,1}}/2} \right)^2 F_{tₜ} \right\} \leq \frac{Nσ²}{c_-} (Δₜσ²).
\]

Apply inequality \( \ln x ≤ x - 1, \forall x > 0 \); therefore,

\[
\mathbb{E} \left\{ \beta_s \ln (\theta_{t,1} v_{t+1}) \mid F_{tₜ} \right\} ≤ - \sum_{tₜ}^{tᵣ+1} \mathbb{E} \left\{ \left| \zeta_{t+1} \mid F_{tₜ} \right\} \right. + \frac{Nσ²}{2β_c} (Δₜσ²).
\]

Note that \( W_\beta \) is monotone decreasing in \( β \), as follows from (A.2). Using this we obtain

\[
\mathbb{E} \left\{ W_{\beta_s}(\tilde{\zeta}_{t+1}) - W_{\beta_s}(\tilde{\zeta}_{t}) \right\} ≤ - \sum_{tₜ}^{tᵣ+1} \mathbb{E} \left\{ \left| \zeta_{t+1} \mid F_{tₜ} \right\} \right. + \frac{Nσ²}{2β_c} (Δₜσ²).
\]

Summing up from \( s = 0 \) to \( s = \bar{s} - 1 \), we obtain (due to (20), (21))

\[
\sum_{tₜ}^{T-1} \mathbb{E} \left\{ \zeta_{t,1} \right\} ≤ - EW_{β_s}(\tilde{\zeta}_{T}) + \sum_{s=0}^{\bar{s}-1} \frac{Nσ²}{2β_c} (Δₜσ²). \tag{A.6}
\]

The minimizer \( θ^* \) arg min \( \tilde{A}(\psi(\theta^*)) \) satisfies \( \tilde{A}(\psi(\theta^*)) = \min_{c \in C} \tilde{A}(c) = A_{\text{min}} \). Applying (A.2) and (A.5) we continue

\[
EW_{β_s}(\tilde{\zeta}_{T}) ≥ -\mathbb{E} \sum_{s=0}^{\bar{s}-1} \Xi \zeta_{s+1} - θ^* - β_s \ln m.
\]

Recall σ-algebras \( \mathcal{F}_t = \sigma (z_s, z_s, u_s, ξ_s \mid s = 1, t - 1) \). By applying Lemma 1, we continue the expectation

\[
\mathbb{E} \sum_{k=1}^{m} \left( \theta^*(k) \Xi \zeta_{t,1} \mid \mathcal{F}_t \right) \leq A_{\text{min}} Δₜ + \frac{O(1)σμ}{(1 - e^{x})c_-} \tag{A.7}
\]

Using (A.2), (A.6)–(A.7), the fact that \( \sup_{θ \in \Theta_m} V(θ) = \ln m \), we obtain

\[
EW_{β_s}(\tilde{\zeta}_{T}) \geq -\mathbb{E} \sum_{s=0}^{\bar{s}-1} \Xi \zeta_{s+1} - θ^* - β_s \ln m
\]

Thus, by applying (20) we get

\[
EW_{β_s}(\tilde{\zeta}_{T}) \geq -β_s \ln m - \sum_{s=0}^{\bar{s}-1} Δₜ + \frac{O(1)σμ}{(1 - e^{x})c_-}
\]

and by applying (A.6) we obtain

\[
\sum_{tₜ}^{T} ( \mathbb{E} \zeta_{t,1} - A_{\text{min}} ) \leq \frac{O(1)σμ}{(1 - e^{x})c_-}
\]

and passing to \( T \rightarrow \infty \) with the related multiplication factor we arrive at (14). ▲

12426