Метод зеркального спуска в задачах о многоруком бандите

А.В. НАЗИН

Институт проблем управления РАН

nazine@ipu.ru, anazine@rambler.ru

На основе совместных работ с

А.Б. Юдицким, А.Б. Цыбаковым и Н. Ваятисом

НМУ, Москва, 27 апреля 2013 г.

План

1. Краткое введение. Идея МЗС (в непрерывном времени) и некоторые его свойства.
Роль преобразования Лежандра, функция Ляпунова, усреднение траектории исходного пространства (в оптимизации).
Оценка скорости сходимости по оптимизируемой функции. Некоторые выводы.

План (продолжение)

2. Общие понятия, объекты и конструкции: исходная и двойственная норма, прокси-функция на заданном выпуклом компакте и ее сопряженная (преобразование Лежандра-Фенхеля), их свойства (при условии сильной выпуклости).
Примеры: “евклидовые” случаи как во всем пространстве, так и в шаре, и энтропийная прокси-функция на стандартном симплексе и распределение Гиббса.

3. Приложение МЗС к задаче о многоруком бандите.

4. Краткий список литературы.

Introduction

Mirror Descent Method (MDA) is a gradient-type recursive method for convex optimization, i.e. primal-dual method performing the descent in a dual space and mapping the resulted points to a primal space. See the following references:

1. Nemirovski and Yudin (1979/1983): [1]

2. Ben-Tal, Margalit, and Nemirovski (2001): [2]

3. Beck and Teboulle (2003): [3]

4. Nesterov (2005, 2007): [4], [5]

5. Juditsky, Nazin, Tsybakov, and Vayatis (2005): [6]

6. Juditsky, Lan, Nemirovski, and Shapiro (2007): [7]

1 Idea behind MDM (continuous time) [1]

Consider a primal-dual method, that is MDM:

$$\dot{\xi}(t) = -\nabla_x f(x(t)) \,, \quad \xi(0) = \xi_0 \,, \tag{1}$$

$$x(t) = \nabla_\xi W(\xi(t)) \,, \quad t \geq 0 \,. \tag{2}$$

Here:

- f is a convex function to be minimized in Banach space E,
- W is a uniform differentiable, convex function on dual space E^* .

As an example, "Euclidean" case of

$$W(\xi) = \frac{1}{2}\|\xi\|_2^2$$

gives a well-known standard gradient method

$$\dot{x}(t) = -\nabla_x f(x(t))\,.$$

Let us look at a simple analysis as follows.

Assume

$$x^* = \arg\min f(x).$$

Then we have a candidate Lyapunov function

$$W_*(\xi) \triangleq W(\xi) - < \xi, x^* >,$$

since

$$\begin{aligned}
\frac{dW_*(\xi(t))}{dt} &= < \dot{\xi}(t), \nabla_\xi W(\xi(t)) - x^* > && (3)\\
&= - < \nabla_x f(x(t)), x(t) - x^* > && (4)\\
&\leq f(x^*) - f(x(t)) && (5)\\
&\leq 0\,, && (6)
\end{aligned}$$

that is function $W_*(\xi)$ decreases along the trajectory $\{\xi(t)\}$.

Furthermore, (3)–(5) lead to

$$f(x(t)) - f(x^*) \;\; \leq \;\; < \dot{\xi}(t), x^* > - \frac{dW(\xi(t))}{dt} \, , \qquad (7)$$

and, assuming that

$$\xi(0) = 0, \quad W(0) = 0,$$

and integrating by $t \in [0, T]$, we get

$$\begin{aligned} \int_0^T f(x(t))dt - Tf(x^*) &\leq \ < \xi(T), x^* > -W(\xi(T)) \qquad (8) \\ &\leq \ V(x^*) \qquad (9) \end{aligned}$$

with the Legendre transformation

$$V(x) \triangleq \sup_{\xi}\{< \xi, x > -W(\xi)\}.$$

Now, introduce the average estimate

$$\widehat{x}(T) \triangleq \frac{1}{T}\int_0^T x(t)dt\,.$$

By Jensen's inequality, due to convexity of $f(x)$, eqs (8)–(9) lead to

$$f(\widehat{x}(T)) - f(x^*) \quad \leq \quad \frac{1}{T}V(x^*). \tag{10}$$

Remark: The rate $O(1/T)$ in the upper bound above changes for that of $O(1/\sqrt{T})$ when working with discreet time gradient observations.

Résumé:

- Function $W : E^* \to \mathbb{R}$ is a parameter of MDM which ensures the Lyapunov function $W_* : E^* \to \mathbb{R}$; in particular, MDM reduces to standard gradient method; therefore, this additional degree of freedom may improve the accuracy algorithm, at least potentially.

- MDM leads to the average estimate $\widehat{x}(t)$, i.e. time-average to current estimates over the time interval $[0, t]$.

- Non-asymptotical upper bound on difference between current estimation function $f(\widehat{x}(t))$ and function minimum $f(x^*)$ is ensured; this upper bound is of type $O(T^{-1})$, and it is directly depending on $V(x^*)$; therefore, the given class function has to ensure the finite upper bound $\sup V(x)$. (Thus, further consideration is reduced to function minimization over a given compact convex set.)
- The previous consideration shows the role of Legendre transformation.

2 A Generalized View-Point

Proxy functions. Denote by E the space $\mathbb{R}^M$ with a norm $\|z\|$ and by E^* the dual space which is $\mathbb{R}^M$ equipped with the conjugate (dual) norm

$$\|z\|_* = \max_{\|\theta\|=1} z^T\theta\,, \quad \forall\, z \in E^*.$$

Let Θ be a convex, closed set in E. For a given parameter $\beta > 0$ and a convex function $V : \Theta \to \mathbb{R}$, we call β-*conjugate* function of V the Legendre–Fenchel type transform of βV:

$$\forall\, z \in E^*, \quad W_\beta(z) = \sup_{\theta\in\Theta} \left\{ -z^T\theta - \beta V(\theta) \right\} . \qquad (11)$$

Assumption (L). *A convex function* $V : \Theta \to \mathbb{R}$ *is such that its* β*-conjugate* W_β *is continuously differentiable on* E^* *and its gradient* ∇W_β *satisfies*

$$\|\nabla W_\beta(z) - \nabla W_\beta(\,\tilde{z}\,)\| \le \frac{1}{\alpha\beta}\|z - \tilde{z}\|_*\,, \quad \forall\, z, \tilde{z} \in E^*,\ \beta > 0,$$

where $\alpha > 0$ *is a constant independent of* β*.*

Assumption (L) relates to the strong convexity w.r.t. *initial norm* $\|\cdot\|$:

$$V(sx+(1-s)y) \le sV(x)+(1-s)V(y)-\frac{\alpha}{2}s(1-s)\|x-y\|^2 \tag{12}$$

for all $x, y \in \Theta$ and any $s \in [0,1]$.

The following proposition sums up some properties of β-conjugates and, in particular, yields a sufficient condition for Assumption (L).

Proposition 1. *Let function* $V : \Theta \to \mathbb{R}$ *be convex and* $\beta > 0$. *Then, the* β*-conjugate* W_β *of* V *has the following properties.*

1. *The function* $W_\beta : E^* \to \mathbb{R}$ *is convex and has a conjugate* βV, *i.e.,*

$$\forall\, \theta \in \Theta, \quad \beta V(\theta) = \sup_{z \in E^*} \left\{ -z^T \theta - W_\beta(z) \right\} .$$

2. *If function* V *is* α*-strongly convex with respect to the initial norm* $\|\cdot\|$ *then*

   (i) *Assumption (L) holds true,*

   (ii) $\displaystyle \operatorname*{argmax}_{\theta \in \Theta} \left\{ -z^T \theta - \beta V(\theta) \right\} = -\nabla W_\beta(z) \in \Theta$.

Definition 1. *We call* $V : \Theta \to \mathbb{R}_+$ proxy function *if it is convex, and*

(i) there exists a point $\theta_* \in \Theta$ *such that* $\min\limits_{\theta \in \Theta} V(\theta) = V(\theta_*)$,

(ii) Assumption (L) holds true.

Example 1: Consider Euclidean space $\mathbb{R}^M$ as set $\Theta = \mathbb{R}^M$. Then half of the squared Euclidean norm be related proxy-function

$$V(\theta) = \frac{1}{2}\,\|\theta\|^2\,, \quad \theta \in \mathbb{R}^M\,.$$

Indeed, minimum point $\theta_* = 0 \in \mathbb{R}^M$, the function is strongly convex w.r.t. the Euclidean norm, and the constant of strong convexity $\alpha = 1$. Evidently, $E^* = E$, a β-conjugate function

$$W_\beta(z) = \frac{1}{2\beta}\,\|z\|^2\,, \quad z \in \mathbb{R}^M$$

with $\nabla W_\beta(z) = z/\beta$. □

Example 2: Let set Θ in the previous Example be Euclidean r-ball with the center at the origin, $r > 0$. The same proxy-function leads to the related β-conjugate function as follows: $\forall\, z \in \mathbb{R}^M$,

$$W_\beta(z) = \begin{cases} \frac{1}{2\beta}\,\|z\|^2, & \|z\| \le r\beta, \\ r\|z\| - \frac{\beta}{2}\,r^2, & \text{otherwise.} \end{cases}$$

The gradient

$$\nabla W_\beta(z) = \begin{cases} \frac{1}{\beta}\,z, & \|z\| \le r\beta, \\ rz/\|z\|, & \text{otherwise;} \end{cases}$$

it realizes the metric projection onto ball $B_{r\beta}$. $\square$

Example 3: Consider a standard simplex $\Theta = \Theta_M$ and an entropy-type proxy function

$$V(\theta) = \ln(M) + \sum_{j=1}^{M} \theta^{(j)} \ln \theta^{(j)} \qquad (13)$$

(where $0 \ln 0 \triangleq 0$) which has a single minimizer $\theta_* = (1/M, \dots, 1/M)^T$ with $V(\theta_*) = 0$.

Let the initial norm in $\mathbb{R}^M$ be 1-norm

$$\|\theta\|_1 = \sum_{j=1}^{M} |\theta^{(j)}|, \quad \theta \in \mathbb{R}^M .$$

Therefore, the initial space is $E = \ell_1^M$, and the dual space $E^* = \ell_\infty^M$ is $\mathbb{R}^M$ equipped with the sup-norm

$$\|z\|_\infty = \max_{\|\theta\|_1=1} z^T\theta = \max_{1\le j\le M} \left|z^{(j)}\right|, \quad \forall\, z \in E^*.$$

It is directly checked that this function is α-strongly convex w.r.t. the 1-norm, with the parameter

$$\alpha = 1\,.$$

This leads to β-conjugate function to $V(\theta)$ as follows:

$$W_\beta(z) = \beta\, \ln\left(\frac{1}{M}\sum_{k=1}^{M} e^{-z^{(k)}/\beta}\right), \quad z \in \mathbb{R}^M, \qquad (14)$$

with partial derivatives relating to a Gibbs distribution on the coordinates of vector $z = (z^{(1)}, \ldots, z^{(M)})^T$, with β being a "temperature" parameter:

$$-\frac{\partial W_\beta(z)}{\partial z^{(j)}} = e^{-z^{(j)}/\beta} \left(\sum_{k=1}^{M} e^{-z^{(k)}/\beta} \right)^{-1}, \; j = 1, \ldots, M. \quad (15)$$

$\square$

Convex Stochastic Optimization Problem

$$A(\theta) \triangleq \mathbb{E}\, Q(\theta, Z) \; \to \min_{\theta \in \Theta}$$

with loss function $Q : \Theta \times \mathcal{Z} \to \mathbb{R}_+$ being such that the random function $Q(\cdot\,, Z) : \Theta \to \mathbb{R}_+$ is convex a.s., on a convex closed set $\Theta \subset \mathbb{R}^M$.

Let a learning sample be given in the form of an i.i.d. sequence $(Z_1, \ldots, Z_{t-1})$, where each Z_i has the same distribution as Z.

Denote stochastic subgradients

$$u_i(\theta) = \nabla_\theta Q(\theta, Z_i)\,, \quad i = 1, 2, \dots, \tag{16}$$

which are measurable functions on $\Theta \times \mathcal{Z}$ such that, for any $\theta \in \Theta$, the expectation $\mathbb{E}\, u_i(\theta)$ belongs to the subdifferential of the function $A(\theta)$.

Mirror Descent Algorithm (MDA)

The algorithm is defined as follows:

- Fix the initial value $\zeta_0 = 0 \in \mathbb{R}^M$.
- For $i = 1, \ldots, t-1$, do the recursive update

$$\begin{aligned} \zeta_i &= \zeta_{i-1} + \gamma_i u_i(\theta_{i-1}) \, , \\ \theta_i &= -\nabla W_{\beta_i}(\zeta_i) \ . \end{aligned} \tag{17}$$

- Output at iteration t the following convex combination:

$$\widehat{\theta}_t = \sum_{i=1}^{t} \gamma_i \theta_{i-1} \left(\sum_{i=1}^{t} \gamma_i \right)^{-1} \ . \tag{18}$$

3 Multi-Armed Bandit Problem (classic).

Presented at the 17th IFAC World Congress:

- Juditsky, A., A.V. Nazin, A.B. Tsybakov, N. Vayatis. Gap-free Bounds for Stochastic Multi-Armed Bandit. *Proc. 17th IFAC World Congress, Seoul, Korea, 6–11 July 2008, pp.11560–11563*.

Let $X = \{x(1), \dots, x(N)\}$ be a set of N available actions. At each time $t = 1, 2, \dots,$ we have to choose sequentially an action $x_t \in X$. We denote by η_t the observable (instantaneous) loss for the choice of x_t, and introduce the average loss up to horizon T which is to be minimized:

$$\Phi_T = \frac{1}{T} \sum_{t=1}^{T} \eta_t \, . \tag{19}$$

A strategy $\mathcal{U}$ is a sequence of rules for the choice x_t at times $t = 1, \ldots, T$. In the stochastic setup that we consider here, the sequence of losses $(\eta_t)_{t\geq 1}$ is a stochastic process and x_t is a measurable function (random, in general) depending only on the vector of past decisions and losses
$(x_1, \ldots, x_{t-1}; \eta_1, \ldots, \eta_{t-1})$.

Any strategy $\mathcal{U}$ generates a flow of σ-algebras
$\mathcal{F}_t = \sigma\{x_1, \ldots, x_t; \eta_1, \ldots, \eta_t\}$, $t \geq 1$ (for brevity we do not indicate the dependence of $\mathcal{F}_t$ on $\mathcal{U}$). Throughout the paper we denote by $z^{(j)}$ the jth component of vector $z \in \mathbb{R}^N$.

Two basic assumptions:

A1. With probability 1, the conditional expectations satisfy

$$\mathbb{E}\{\eta_t \,|\, \mathcal{F}_{t-1}\,,\, x_t = x(k)\} = a_k,\ \ k = 1, \dots, N, \qquad (20)$$

where $a_k \in \mathbb{R}$ are unknown deterministic values.

The value a_k characterizes the expected loss for deciding to take the action $x_t = x(k)$ at time t. Assumption A1 says that this loss should not depend on t.

A2. The second conditional moment of the loss η_t is a.s. bounded by a constant:

$$\mathbb{E}\{\eta_t^2 \,|\, \mathcal{F}_{t-1}\,,\, x_t\} \le \sigma^2 < \infty\,. \qquad (21)$$

It is easy to prove (see, e.g., [8]) that under these assumptions all the limiting points of the average loss sequence $(\Phi_t)_{t\geq 1}$ cannot be almost surely (a.s.) less than

$$a_{\min} \triangleq \min_{k=1,\ldots,N} a_k \,.$$

Thus, the problem is to design a strategy $\mathcal{U}^*$ which has the asymptotically minimal average loss:

$$\Phi_T \to a_{\min} \quad \text{as} \quad T \to \infty \,, \tag{22}$$

in an appropriate probability sense.

We study here *convergence in mean*, trying to get the rate of convergence

$$\mathbb{E}(\Phi_T) \to a_{\min}$$

as fast as possible.
In particular, we provide *non-asymptotic* upper bounds for the expected excess risk $\mathbb{E}(\Phi_T) - a_{\min}$ that are close, up to logarithmic factors, to the lower bound of the order $\sqrt{N/T}$ proved for arbitrary N by (see Theorem 6.11 in [10]).

We will suppose that the following assumption on the loss sequence $(\eta_t)_{t\geq 1}$ holds:

A3. The losses are nonnegative: $\eta_t \geq 0$ a.s.

Below we propose a randomized decision strategy in which, at each step $t+1$, the action x_{t+1} is drawn according to a distribution $p_t \triangleq \left(p_t^{(1)}, \ldots, p_t^{(N)}\right)^\top$ over X where:

$$p_t^{(k)} \triangleq \mathbb{P}(x_{t+1} = x(k) \,|\, \mathcal{F}_t)\,, \quad k = 1, \ldots, N\,. \tag{23}$$

The update of the distribution p_t over time is given by the MDA.

Denote by Θ the simplex of all probability vectors over X:

$$\Theta \triangleq \left\{ p \in \mathbb{R}_+^N \;\middle|\; \sum\nolimits_{k=1}^{N} p^{(k)} = 1 \right\} . \tag{24}$$

We then define the mean (over the set of actions) loss function A on Θ:

$$A(p) = \sum_{k=1}^{N} a_k p^{(k)} = a^\top p \,, \quad p \in \Theta \ , \tag{25}$$

where $a = (a_1, \ldots, a_N)^\top$. Since p_t is a random vector, the quantity $A(p_t)$ is a random variable. The update rule for the probability distribution p_t uses a stochastic gradient of A.

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

$$\mathbb{E}(\Phi_T) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(\mathbb{E}(\eta_t \mid x_t\,, \mathcal{F}_{t-1})) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(A(p_{t-1}))\ . \tag{26}$$

Theorem. *Let assumptions A1–A3 be satisfied and let the conditional distributions $(p_t)_{t\geq 0}$ be defined by the MDA. Then, for any horizon $T \geq 1$,*

$$\mathbb{E}\left(\Phi_T\right) - a_{\min} \leq 2\,\sigma\,\frac{\sqrt{(T+1)N\ln N}}{T}\,. \tag{27}$$

The MD algorithm for multi-armed bandit.

1. Fix $p_0 = (N^{-1}, \ldots, N^{-1})^T$ and $\zeta_0 = 0 \in \mathbb{R}^N$.
2. For $t = 1, \ldots, T$:
   (a) draw an action $x_t = x(k_t)$ with random k_t distributed according to p_{t-1};
   (b) compute the stochastic gradient
   $$u_t(p_{t-1}) = \frac{\eta_t}{p_{t-1}^{(k_t)}} \, e_N(k_t)\,; \tag{28}$$
   (c) update the dual and probability vectors
   $$\begin{aligned} \zeta_t &= \zeta_{t-1} + \gamma_t u_t(p_{t-1})\,, && (29) \\ p_t &= -\nabla W_{\beta_t}(\zeta_t)\,. && (30) \end{aligned}$$

3. At horizon $t = T$, output a sequence of actions $(x_1, \ldots, x_T)$.

The tuning parameters γ_t and β_t are as follows: $\forall t \geq 1$,

$$\gamma_t \equiv 1, \quad \beta_{t-1} = \beta_0 \sqrt{t}, \quad \beta_0 = \sigma \sqrt{N/(\ln N)}\,. \tag{31}$$

Notice that

$$\mathbb{E}\left\{ \frac{\eta_t}{p_{t-1}^{(k_t)}}\, e_N(k_t) \,\middle|\, \mathcal{F}_{t-1} \right\} = a = \nabla A(p_{t-1})\,. \tag{32}$$

Here is the corrected information lower bound from [10], Theorem 6.11. Let $\Phi_{i,T}$ be mean losses under fixed i-th arm, i.e., $x_t \equiv x(i)$.

Theorem. *Let $T, N \geq 1$ be such that $T > N/(4\ln(4/3))$. There exists a loss function such that for any, possibly randomized, control strategy*

$$\sup_{\mathcal{Z}} \mathbb{E}\left(\Phi_T - \min_{i=1,\dots,N} \Phi_{i,T}\right) \geq \frac{\sqrt{N/T}}{32\sqrt{\ln 4/3}}\,, \qquad (33)$$

where $\sup$ *is over set of all multi-armed bandit problems with losses η_t with values from interval $[0,1]$ a.s.*

Remarks: The information lower bound above (see [10], Theorem 6.11) differs from the upper bound (27) by logarithmic term $\sqrt{\ln N}$.

The wrong constant in the lower bound of Theorem 6.11 [10]

$$\frac{\sqrt{2}-1}{\sqrt{32\ln 4/3}} \approx 0.1365 \tag{34}$$

is more than that of Theorem

$$\frac{1}{32\sqrt{\ln 4/3}} \approx 0.0583\,. \tag{35}$$

Unfortunately, the constant (34) is uncorrectly calculated in [10], page 165.

4 Multi-Armed Bandit Governed by a Stationary Finite Markov Chain.

To be presented at the ECC2013:

- Nazin, A.V., B.M. Miller. Mirror Decent Algorithm for a Multi-Armed Bandit Governed by a Stationary Finite State Markov Chain. *The 12th European Control Conference, ECC13, July 17–19, 2013, Zurich, Switzerland.*

In addition to the classic case of Multi-Armed Bandit Problem, assume that instantaneous losses η_t depend now on both chosen arm $x_t \in X$ and current state $z_t \in Z$ of *unknown* stationary finite Markov Chain (MC), $Z = \{z(1), \ldots, z(K)\}$. The main new assumption is as follows:

- the transition probabilities of the state $z_t \in Z$ at each time $t \in \{0, 1, \ldots\}$ to the next state $z_{t+1} \in Z$ are presented by unknown conditional probabilities: $\forall\, t$,

$$\mathbb{P}\{z_{t+1} = z(j) \,|\, z_t = z(i)\} = \pi_{ij}; \qquad (36)$$

- MC state z_t is observable at current time $t \geq 0$.

Further assumptions:

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

$$\{\eta_t(z, u, \omega) \,|\, z \in Z, u \in U\} \quad \text{and}$$

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

are independent.

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

$$\mathbb{E}\{\eta_t(z(i), u(\ell), \omega)\} \triangleq a_{i\ell} \quad \forall\, t\,. \tag{37}$$

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

$$\mathbb{E}\{\eta_t^2(z(i), u(\ell), \omega)\} \le \sigma^2 < \infty \,. \tag{38}$$

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

A5. The initial distribution of MC assumed to be stationary. The stationary distribution of the MC states is assumed to be unknown.

Introduce randomized strategy by

$$d_t^{(\,i\ell)} \triangleq \mathbb{P}\{u_t = u(\ell)\,|\,z_t = z(i),\, \mathcal{F}_{t-1}\}\,. \tag{39}$$

Under a stationary strategy $\mathcal{U}_{\mathrm{St}}$ with $d \triangleq \|d^{(i\ell)}\|$, the loss expectation lead to the loss function

$$\begin{aligned} \mathbb{E}\{\eta_t\} &= \sum\nolimits_{i=1}^{K} q_i \sum\nolimits_{\ell=1}^{N} a_{i\ell}\, d^{\,(i\ell)} && (40) \\ &\triangleq A(d)\,, \quad d \in D, && (41) \end{aligned}$$

with stationary state probabilities

$$q_i \triangleq \mathbb{P}\{z_t = z(i)\} \tag{42}$$

and the set stochastic matrix

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

Denote

$$A_{\min} \triangleq \min_{d \in D}\ A(d)\,. \tag{43}$$

Theorem. *Let assumptions A1–A5 be satisfied and let the conditional distributions* $(d_t^{(i)})_{t \geq 0}$, $i = \overline{1,K}$*, be defined by the randomized control algorithm (see below) with parameters (48). Then, for any time* $T \geq 1$*,*

$$\mathbb{E}\,(\Phi_T) - A_{\min} \leq 2\sigma\,\sqrt{KN \ln N}\,\frac{\sqrt{T+1}}{T}\,. \tag{44}$$

■

Thus, we fix the increasing temperature parameter sequence $(\beta_t)_{t\geq 0}$ and introduce 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* $t \geq 0$*, by having the observed state* $z_t = z(i_t)$*, draw arm* $x_t = x(\ell_t)$ *with random* $\ell_t \in \{\overline{1,N}\}$ *distributed according to* $(d_t^{(i_t 1)}, \ldots, d_t^{(i_t N)})^\top$*;*

(b) *compute a stochastic gradient*

$$\Xi_{t+1} = \frac{\eta_{t+1}}{d_t^{i_t \ell_t}} \, e_K(i_t) e_N^\top(\ell_t)\,; \tag{45}$$

(c) *update both dual and initial variables*

$$\zeta_{t+1} = \zeta_t + \Xi_{t+1}\,, \tag{46}$$

$$d_{t+1}^{(i)} = G_{\beta_t}(\zeta_{t+1}^{(i)})\,, \quad \forall\, i = \overline{1,K}\,. \tag{47}$$

2. *At time T of interest, output the observed sequences of states $(z_0, \ldots, z_T)$, control actions $(u_0, \ldots, u_T)$, matrices $(d_0, \ldots, d_T)$, and the observed losses $(\eta_1, \ldots, \eta_{T+1})$ and Φ_T.*

The tuning algorithm parameter β_t is defined as follows:
$\forall\, t = 0, 1, \ldots,$

$$\beta_t = \beta_0 \sqrt{t+1}, \quad \beta_0 = \sigma \sqrt{N/(K \ln N)}\,. \tag{48}$$

Вычислительный пример: $K = 7$ и $N = 5$;

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

далее, матрица средних потерь

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

и $A_{\min} = 0.1482$. Случайные потери $\eta_t(z(i), x(\ell), \omega)$ в состоянии $z(i)$ и выбранной ручке $x(\ell)$ являются н.о.р. с.в. Бернулли с вероятностями $\mathbb{P}\left(\eta_t(z(i), x(\ell), \omega) = 1\right) = a_{i\ell}$.

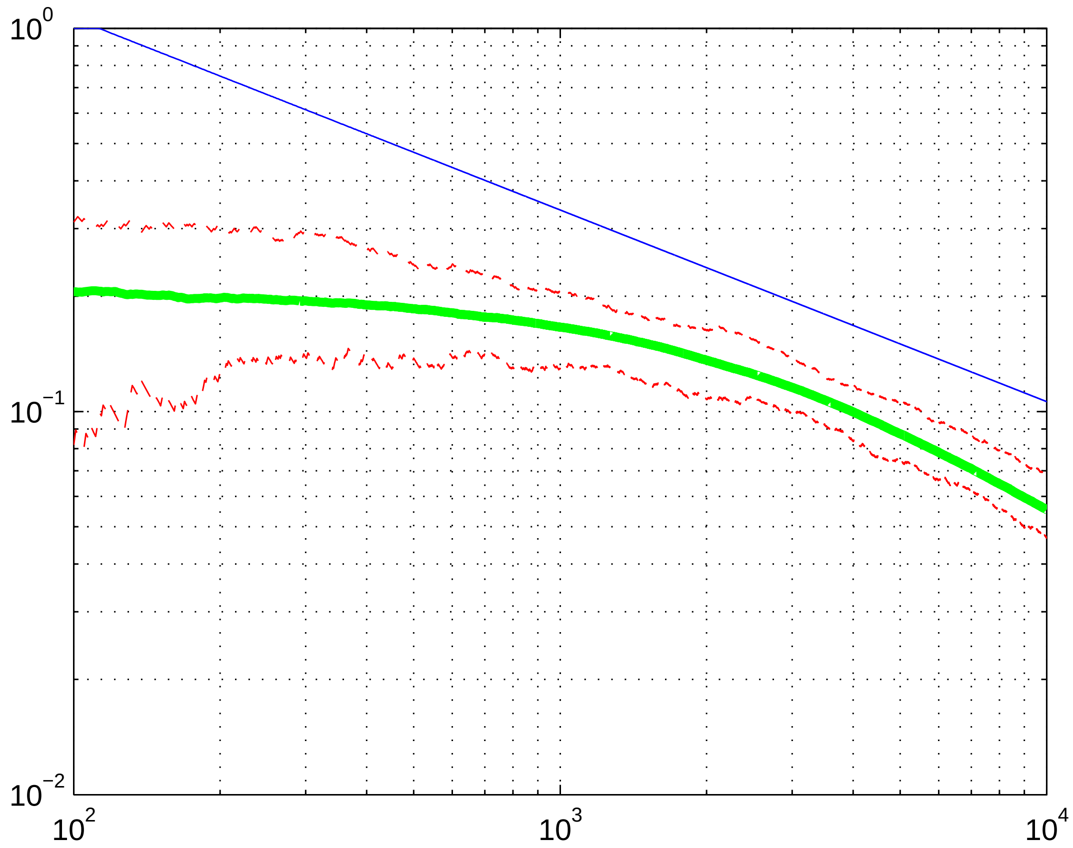

Рис. 1: Результаты вычислительного примера с числом состояний $K = 7$ и числом рук $N = 5$ представлены в двойном логарифмическом масштабе на интервале времени $t = 100, \ldots, 10000$.

Список литературы

[1] Nemirovskii, A.S. and Yudin, D.B., *Slozhnost' zadach i effektivnost' metodov optimizatsii*, Moscow: Nauka, 1979 (in Russian). Translated under the title *Problem Complexity and Method Efficiency in Optimization*, Chichester: Wiley, 1983.

[2] A. Ben-Tal, T. Margalit, A. Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIOPT* 12(1), 79–108, 2001.

[3] A. Beck, M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.* 31(3), 167–175, 2003.

[4] 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.

[5] Yu. Nesterov. Primal-dual subgradient methods for convex problems // *Mathematical Programming*, 2007. DOI: 10.1007/s10107-007-0149-x.

[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.

[7] A. Juditsky, G. Lan, A. Nemirovski, and A. Shapiro. Stochastic Approximation approach to Stochastic Programming. *Optimization Online*, 10/03/2007 http://www.optimization-online.org/DB_HTML/2007/09/1787.html

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

[9] K. Najim and A.S. Poznyak. *Learning automata: theory and applications*. Pergamon Press, Inc., Elmsford, NY, USA, 1994. ISBN 0-08-042024-9.

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

[11] Nazin, A.V., B.M. Miller. Robust Mirror Decent Algorithm for a Multi-Armed Bandit Governed by a Stationary Finite Markov Chain. *IFAC MIM'2013 Conference on Manufacturing Modelling, Management, and Control, June 19–21, 2013, Saint Petersburg, Russia.*

[12] Nazin, A.V., B.M. Miller. Mirror Decent Algorithm for a Multi-Armed Bandit Governed by a Stationary Finite State Markov Chain. *The 12th European Control Conference, ECC13, July 17–19, 2013, Zurich, Switzerland.*

THANK YOU FOR YOUR ATTENTION !!!