Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Gap-free Bounds for Stochastic Multi-Armed Bandit*

A. Juditsky * **A. V. Nazin** ** **A.B. Tsybakov** *** **N. Vayatis** ****

* *LJK – Université Grenoble I, France (e-mail: anatoli.iouditski@imag.fr)*
** *Laboratory for Adaptive and Robust Control Systems – Institute of Control Sciences of RAS, Profsoyuznaya str., 65, 117997 Moscow, Russia (Tel: +7-495-334-7641; e-mail: nazine@ipu.rssi.ru).*
*** *Laboratoire de Statistique – CREST, Malakoff, France (e-mail: alexandre.tsybakov@ensae.fr)*
**** *CMLA – ENS Cachan, UniverSud and CNRS, France (e-mail: vayatis@cmla.ens-cachan.fr)*

Abstract: We consider the stochastic multi-armed bandit problem with unknown horizon. We present a randomized decision strategy which is based on updating a probability distribution through a stochastic mirror descent/exponentiated gradient type algorithm. We consider separately two assumptions: nonnegative losses or arbitrary losses with an exponential moment condition. We prove optimal (up to logarithmic factors) gap-free bounds on the excess risk of the average over time of the instantaneous losses induced by the choice of a specific action.

1. INTRODUCTION

The multi-armed bandit is a celebrated problem in the field of sequential prediction. In this problem, the forecaster has to choose, over some time sequence, among a finite set of available actions and the only information he gets at each step is the instantaneous loss he suffers for the selected action. A standard criterion to assess the performance of a given strategy for action selection is the difference between the average over time of the instantaneous losses and the minimal (over the set of actions) average loss. The statistical theory is then devoted to estimating how does this risk function can be controlled in terms of the number N of possible actions and the length T of the time sequence. Several variants of this problem have been studied and we refer to Cesa-Bianchi and Lugosi (2006) and the PhD thesis of Stoltz (2005) for a detailed and modern account on this topic. Early book-length studies in Nazin and Poznyak (1986) and Najim and Poznyak (1994) can be also of interest.

In the present paper, we consider the stochastic setup which was first introduced by Robbins (1952), Lai and Robbins (1985) and further studied by Auer et al. (2002a). The horizon is not assumed to be known in advance. We then propose a stochastic optimization algorithm which can be viewed as a modification of the exponentiated gradient algorithm of Kivinen and Warmuth (1997) or, more generally, of the mirror descent algorithm given in Juditsky et al. (2005). Previous works on stochastic multi-armed bandit, (cf. Auer et al. (2002a) and more recent developments inspired by that paper) provide bounds on the excess risk with fast rates but involving (unknown) gaps between the expected loss of each non-optimal arm and the minimal expected loss. Such bounds can be called gap-dependent. They are similar in the spirit to Lai and Robbins (1985) whose bounds used Kullback divergences between distributions of arms rather than gaps. For instance, recent gap-dependent bounds by Audibert et al. (2007) cannot be really used because they involve unknown parameters of the problem. Note also that these bounds go to infinity over the parameter class (see Corollary 1, Theorem 3, and Theorem 9 in Audibert et al. (2007)).

A different approach has been suggested by Auer et al. (2002b) and further developed by Cesa-Bianchi and Lugosi (2006). They established gap-free bounds for a non-stochastic multi-armed bandit setting where the losses were assumed to take values in $[0, 1]$. In the present paper we derive gap-free bounds for the expected excess risk, with tight constants, under two different assumptions on the stochastic process of the instantaneous losses. In the case of nonnegative losses with finite variance we obtain an explicit expected excess bound, for any horizon $T \geq 1$; in particularly, it implies the convergence rate of the order $\sqrt{(N/T)(\ln N)}$. In the case of signed losses under an exponential moment assumption we get the similar explicit bound with an additional logarithmic factor $(\ln T)\sqrt{(N/T)(\ln N)}$.

The rest of the paper is organized as follows. We first introduce the setup and state the main convergence results (Sections 2 and 3). Then we describe the algorithm (Section 4) and provide the technical details for the proof of the upper bounds in the Appendix A.

2. STATEMENT OF THE PROBLEM

Let $X = \{x(1), \ldots, x(N)\}$ be a set of N available actions. At each time $t = 1, 2, \ldots$, we have to choose sequentially an action $x_t \in X$. We denote by η_t the observable

* The work of the second author was supported in part by Russian Foundation for Basic Research through the grant RFBR 06-08-01474.

(instantaneous) loss for the choice of x_t, and introduce the average loss up to horizon T which is to be minimized:

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

A strategy $\mathcal{U}$ is a sequence of rules for the choice x_t at times $t = 1, \dots, T$. 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 $(x_1, \dots, x_{t-1}; \eta_1, \dots, \eta_{t-1})$. Any strategy $\mathcal{U}$ generates a flow of σ-algebras $\mathcal{F}_t = \sigma\{x_1, \dots, x_t; \eta_1, \dots, \eta_t\}$, $t \geq 1$ (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 $z \in \mathbb{R}^N$.

Introduce the following two basic assumptions:

A1. With probability 1, the conditional expectations satisfy

$$\mathbb{E}\{\eta_t \,|\, \mathcal{F}_{t-1}\,,\, x_t = x(k)\} = a_k,\ k = 1, \dots, N, \tag{2}$$

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:

$$\mathbb{E}\{\eta_t^2 \,|\, \mathcal{F}_{t-1}\,,\, x_t\} \leq \sigma^2 < \infty\,. \tag{3}$$

It is easy to prove (see, e.g., Nazin and Poznyak (1986)) 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

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

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

$$\Phi_T \to a_{\min} \quad \text{as} \quad T \to \infty\,, \tag{4}$$

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 $N = 2$ by Vogel (1960) and for arbitrary N by Auer et al. (2002b) (see also Theorem 6.11 in Cesa-Bianchi and Lugosi (2006)).

We will suppose that one of the following two assumptions on the loss sequence $(\eta_t)_{t\geq 1}$ holds.

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

A4. The random variables η_t have finite exponential moments: there exist constants $C_\eta, \kappa > 0$ such that

$$\mathbb{E}\left\{ |\eta_t| e^{-\kappa\eta_t/\sigma} \,\middle|\, \mathcal{F}_{t-1}\,,\, x_t \right\} \leq C_\eta\sigma < \infty,\ \forall\, t \geq 1. \tag{5}$$

3. RESULTS

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 $p_t \triangleq \left(p_t^{(1)}, \dots, p_t^{(N)}\right)^\top$ over X where:

$$p_t^{(k)} \triangleq \mathbb{P}(x_{t+1} = x(k) \,|\, \mathcal{F}_t)\,, \quad k = 1, \dots, N\,. \tag{6}$$

The update of the distribution p_t over time is given by the algorithm described in Section 4.

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

$$\Theta \triangleq \left\{ p \in \mathbb{R}_+^N \,\middle|\, \sum\nolimits_{k=1}^{N} p^{(k)} = 1 \right\}. \tag{7}$$

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

$$A(p) = \sum_{k=1}^{N} a_k p^{(k)} = a^\top p\,, \quad p \in \Theta\,, \tag{8}$$

where $a = (a_1, \dots, a_N)^\top$. 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

$$\mathbb{E}(\Phi_T) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(\mathbb{E}(\eta_t \,|\, x_t\,, \mathcal{F}_{t-1})) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(A(p_{t-1}))\,. \tag{9}$$

We are now in a position to state our results.

Theorem 1. Let assumptions A1-A2-A3 be satisfied and let the conditional distributions $(p_t)_{t\geq 0}$ be defined by the algorithm of Section 4 with parameters (15), (17) and $c_0 = 1$. Then, for any horizon $T \geq 1$,

$$\mathbb{E}\,(\Phi_T) - a_{\min} \leq 2\,\sigma\,\frac{\sqrt{(T+1)N\ln N}}{T}\,. \tag{10}$$

Theorem 2. Let assumptions A1-A2-A4 be satisfied and let the conditional distributions $(p_t)_{t\geq 0}$ be defined by the algorithm of Section 4 with parameters (16), (17). Then, for any horizon $T \geq 2$,

$$\mathbb{E}\,(\Phi_T) - a_{\min} \leq \sigma\,\frac{\sqrt{(T+1)}}{T} \times \left\{ 2C_\eta + \sqrt{N\ln N}\left(\left[c_0 + \frac{1}{2\kappa^2 c_0} \right] \ln(T+1) + \frac{2}{c_0} \right) \right\}. \tag{11}$$

Remark 1. Minimizing the right hand side of (11) in $c_0 > 0$ one can find both the optimal parameter c_0 and the corresponding optimal upper bound. Optimal c_0 depends on the horizon T but this dependence becomes negligible for large T. If T is unknown, it looks reasonable to set c_0 by minimizing the main term, i.e., the expression in square brackets in (11): $c_0 = (\sqrt{2}\kappa)^{-1}$.

The previously known gap-free results (Auer et al. (2002b) and Theorem 6.10 in Cesa-Bianchi and Lugosi (2006)) assume that $\eta_t \in [0, 1]$ and prove that $\Phi_T - a_{\min}$ is $O\big(\sqrt{(N/T)\ln(NT/\delta)}\big)$ or $O\big(\sqrt{(N/T)\ln(N/\delta)}\big)$, respectively, with probability at least $1-\delta$, where $0 < \delta < 1$. The bound of Theorem 1 is given for the expectation $\mathbb{E}(\Phi_T) - a_{\min}$ and therefore it is not directly comparable to these bounds in probability. Note however that those papers assume $\eta_t \in [0, 1]$ whereas Theorem 1 assures the result for unbounded nonnegative losses in the stochastic context. Bounds in expectation are obtained in the same form as in Theorem 1 with $\sigma = 1$ by Stoltz (2005), again

under the assumption that $\eta_t \in [0,1]$. Finally, note that Theorem 1 is proved for the pure exponentiated gradient/mirror descent algorithm (with time-dependent choice of tuning parameters), whereas Auer et al. (2002b) and Cesa-Bianchi and Lugosi (2006) obtain their bounds for a more sophisticated procedure.

4. DEFINITION OF THE STRATEGY

In this section we introduce our algorithm. It is related to the exponentiated gradient method of Kivinen and Warmuth (1997) and to the mirror descent algorithm given in Juditsky et al. (2005). We refer to Nemirovski and Yudin (1983) 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

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

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

The algorithm is defined as follows.

(1) Fix $p_0 \in \Theta$ and $\zeta_0 = 0 \in \mathbb{R}^N$.
(2) For $t = 1, \dots, T$:
- *(a) draw an action $x_t = x(k_t)$ with random k_t distributed according to p_{t-1};*
- *(b) compute the thresholded stochastic gradient*

$$u_t(p_{t-1}) = \frac{\max\{\eta_t + \Delta_t, 0\}}{p_{t-1}^{(k_t)}} \, e_N(k_t) \,; \quad (12)$$

- *(c) update the dual and probability vectors*

$$\zeta_t = \zeta_{t-1} + \gamma_t u_t(p_{t-1}) \,, \quad (13)$$

$$p_t = G_{\beta_t}(\zeta_t) \,. \quad (14)$$

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

The tuning parameters Δ_t, γ_t and β_t involved in the algorithm are defined differently for Theorems 1 and 2. For all $t \geq 1$, in Theorem 1 we set

$$\gamma_t \equiv 1, \quad \Delta_t \equiv 0, \quad \beta_{t-1} = \beta_0 \sqrt{t}, \quad (15)$$

whereas in Theorem 2 we set

$$\gamma_t \equiv 1, \quad \Delta_t = \frac{\sigma}{\kappa} \ln \sqrt{t}, \quad \beta_{t-1} = \beta_0 \sqrt{t} \ln (t \vee e) \quad (16)$$

with the constant β_0 given, for both theorems, by

$$\beta_0 = c_0 \sigma \sqrt{N/(\ln N)} \quad (17)$$

with some $c_0 > 0$; $t \vee e \triangleq \max\{t, e\}$. It is important to note that these choices do not involve the horizon T which is not necessarily known in advance.

Remark that the vector $(\eta_t / p_{t-1}^{(k_t)}) \, e_N(k_t)$ is in fact a stochastic gradient: its conditional expectation given $\mathcal{F}_{t-1}$ equals to the gradient of the mean loss $A(p)$:

$$\mathbb{E} \left\{ \frac{\eta_t}{p_{t-1}^{(k_t)}} \, e_N(k_t) \,\middle|\, \mathcal{F}_{t-1} \right\} = a = \nabla A(p_{t-1}) \,. \quad (18)$$

The threshold parameter $\Delta_t > 0$ in the definition of $u_t(p_{t-1})$ modifies the stochastic gradient by lower bounding the loss η_t. In Theorem 1 we have no thresholding: there $u_t(p_{t-1})$ is just a stochastic gradient and our algorithm is a special case of the mirror descent/exponentiated gradient method. However, in Theorem 2 thresholding plays a crucial role.

REFERENCES

J.Y. Audibert, Remi M., and Cs. Szepesvari. Tuning bandit algorithms in stochastic environments. In *18th International Conference on Algorithmic Learning Theory*, pages 150–165, Sendai, 1–4 October 2007.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002a.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b.

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.

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

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.

J. Kivinen and M. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

T.L. Lai and H. Robbins. Asymptotic efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6: 4–22, 1985.

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

A.V. Nazin and A.S. Poznyak. *Adaptive Choice of Variants*. Nauka, Moscow, 1986.

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

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.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 55:527–535, 1952.

G. Stoltz. *Incomplete information and internal regret in prediction of individual sequences*. PhD thesis, Université Paris-Sud, 2005.

W. Vogel. An asymptotic minimax theorem for the two armed bandit problem. *The Annals of Mathematical Statistics*, 31:444–451, 1960.

Appendix A. PROOFS

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

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

Furthermore, W_β and the entropy type function

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

are related to each other via convex duality formula:

$$W_\beta(z) = \sup_{\theta\in\Theta}\left\{-z^T\theta - \beta V(\theta)\right\}, \quad z \in \mathbb{R}^N. \tag{A.1}$$

A.1 ***Proof*** *of Theorem 2*

Note that

$$W_{\beta_{t-1}}(\zeta_t) - W_{\beta_{t-1}}(\zeta_{t-1}) = \beta_{t-1}\ln\left(\frac{\sum_{k=1}^N e^{-\zeta_t^{(k)}/\beta_{t-1}}}{\sum_{k=1}^N e^{-\zeta_{t-1}^{(k)}/\beta_{t-1}}}\right)$$
$$= \beta_{t-1}\ln(p_{t-1}^\top v_t)$$

where the k-th entry of vector v_t equals $v_t^{(k)} = e^{-u_t^{(k)}/\beta_{t-1}}$ and $u_t^{(k)}$ is the k-th entry of $u_t(p_{t-1})$. Since $e^x \leq 1 + x + x^2/2$ for $x \leq 0$, we get

$$v_t^{(k_t)} \leq 1 - \frac{u_t^{(k_t)}}{\beta_{t-1}} + \frac{(u_t^{(k_t)})^2}{2\beta_{t-1}^2}$$

and clearly $v_t^{(k)} = 1$ for all $k \neq k_t$. Introduce the vectors

$$\widetilde{\eta}_t \triangleq \max\{\eta_t + \Delta_t\,, 0\}\, e_N(k_t)$$

having the a.s. nonnegative entries $\widetilde{\eta}_t^{(k)}$. Then we have

$$\beta_{t-1}\ln(p_{t-1}^\top v_t) \leq \beta_{t-1}\ln\left(1 - \frac{\widetilde{\eta}_t^{(k_t)}}{\beta_{t-1}} + \frac{(\widetilde{\eta}_t^{(k_t)})^2}{2p_{t-1}^{(k_t)}\beta_{t-1}^2}\right)$$
$$\leq -\widetilde{\eta}_t^{(k_t)} + \frac{(\widetilde{\eta}_t^{(k_t)})^2}{2p_{t-1}^{(k_t)}\beta_{t-1}}\,. \tag{A.2}$$

Note that W_β is monotone decreasing in β, as follows from (A.1). Using this, taking expectation of both sides of (A.2) (first over k_t, conditional on p_{t-1}, then over p_{t-1}) and applying assumption A2 we obtain

$$\mathbb{E}\left(W_{\beta_t}(\zeta_t) - W_{\beta_{t-1}}(\zeta_{t-1})\right)$$
$$\leq -\mathbb{E}(\widetilde{\eta}_t^\top p_{t-1}) + \frac{1}{2\beta_{t-1}}\mathbb{E}\left(\sum_{k=1}^N(\widetilde{\eta}_t^{(k)})^2\right) \tag{A.3}$$
$$\leq -\mathbb{E}(a^\top p_{t-1}) - \Delta_t + \frac{(\sigma^2+\Delta_t^2)N}{\beta_{t-1}}\,.$$

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

$$\sum_{t=0}^{T-1}\mathbb{E}(a^\top p_t) \leq -\mathbb{E}W_{\beta_T}(\zeta_T) + N\sum_{t=0}^{T-1}\frac{\sigma^2+\Delta_{t+1}^2}{\beta_t} - \sum_{t=1}^T \Delta_t\,.$$

The minimizer $p^* \triangleq \arg\min_{p\in\Theta} A(p)$ of the linear form $A(p) = a^\top p$ on the simplex Θ satisfies $A(p^*) = a_{\min}$. Therefore

$$\sum_{k=1}^N \mathbb{E}\left(p^{*(k)}\mathbb{E}\{\eta_t \,|\, k_t = k, \mathcal{F}_{t-1}\}\right) = \sum_{k=1}^N a_k p^{*(k)} = a_{\min}.$$

Using (A.1), the fact that $\sup_{\theta\in\Theta} V(\theta) = \ln N$, and the last display we get

$$\mathbb{E}W_{\beta_T}(\zeta_T) \geq -\mathbb{E}(\zeta_T^\top p^*) - \beta_T \ln N$$
$$= -\sum_{t=1}^T\sum_{k=1}^N \mathbb{E}\left(\widetilde{\eta}_t^{(k)} p^{*(k)}\right) - \beta_T \ln N$$
$$= -Ta_{\min} - \beta_T \ln N + \sum_{t=1}^T\left(-\Delta_t + \mathbb{E}\sum_{k=1}^N p^{*(k)}\nu_t^{(k)}\right)$$

where

$$\nu_t^{(k)} = \mathbb{E}\{(\eta_t + \Delta_t)1\{\eta_t < -\Delta_t\} \,|\, k_t = k, \mathcal{F}_{t-1}\} \tag{A.4}$$

and $1\{\cdot\}$ denotes the indicator function. Thus,

$$\sum_{t=0}^{T-1}\mathbb{E}(a^\top p_t) \leq \beta_T \ln N + Ta_{\min} + N\sum_{t=0}^{T-1}\frac{\sigma^2+\Delta_{t+1}^2}{\beta_t}$$
$$+ \sum_{t=1}^T\sum_{k=1}^N p^{*(k)}\mathbb{E}|\nu_t^{(k)}|$$
$$= \beta_T \ln N + Ta_{\min} + N\sum_{t=0}^{T-1}\frac{\sigma^2+\Delta_{t+1}^2}{\beta_t} + \sum_{t=1}^T \max_k \mathbb{E}|\nu_t^{(k)}|.$$

We now use the exponential Markov inequality and Assumption A3. This yields

$$\sum_{t=0}^{T-1}\left[\,\mathbb{E}A(p_t) - a_{\min}\right] \leq \beta_T \ln N \tag{A.5}$$
$$+ N\sum_{t=0}^{T-1}\frac{\sigma^2+\Delta_{t+1}^2}{\beta_t} + \sigma C_\eta \sum_{t=1}^T e^{-\kappa\Delta_t/\sigma}\,. \tag{A.6}$$

Recall that $\Delta_t = (\sigma/\kappa)\ln\sqrt{t}$ and $\beta_{t-1} = \beta_0\sqrt{t}\,\ln(t\vee e)$ for $t \geq 1$. Therefore,

$$\sum_{t=0}^{T-1}\frac{\Delta_{t+1}^2}{\beta_t} \leq \frac{\sigma^2\sqrt{T}\ln T}{2\beta_0\kappa^2}$$

and the result of the theorem easily follows from (A.5)–(A.6) and Eq.(9). □

A.2 ***Proof*** *of Theorem 1*

Here $\Delta_t \equiv 0$, $\widetilde{\eta}_t = \eta_t e_N(k_t)$, and η_t are nonnegative (implying $\nu_t^{(k)} = 0$). Therefore, from (A.3) we get

$$\mathbb{E}\left(W_{\beta_t}(\zeta_t) - W_{\beta_{t-1}}(\zeta_{t-1})\right) \leq -\mathbb{E}(a^\top p_{t-1}) + \frac{N\sigma^2}{2\beta_{t-1}}\,.$$

Using this inequality and acting as in the proof of Theorem 2, we arrive to a simplified analog of (A.5)–(A.6):

$$\sum_{t=0}^{T-1}\left[\,\mathbb{E}A(p_t) - a_{\min}\right] \leq \beta_T \ln N + N\sum_{t=0}^{T-1}\frac{\sigma^2}{2\beta_t}\,.$$

The choice of β_t as in (15), (17) with $c_0 = 1$ as well as Eq.(9) finish the proof. □