UDC 519.853.6+519.856

# Algorithms of Inertial Mirror Descent in Stochastic Convex Optimization Problems

**A. V. Nazin***

* *V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Profsoyuznaya str. 65, Moscow, 117997, Russia*

**Abstract.** The goal is to modify the known method of mirror descent (MD) in convex optimization, which having been proposed by A.S. Nemirovsky and D.B. Yudin in 1979 and generalized the standard gradient method. To start, the paper shows the idea of a new, so-called inertial MD method with the example of a deterministic optimization problem in continuous time. In particular, in the Euclidean case, the heavy ball method by B.T. Polyak is realized. It is noted that the new method does not use additional averaging of points. Then, a discrete algorithm of inertial MD is described. The proved theorem of the upper bound on error in objective function is formulated.

**Keywords:** stochastic optimization problem, convex optimization, mirror descent, heavy ball method, inertial MD method.

## 1. Introduction

Many problems of an applied nature can formally be reduced to minimization problem $f(x) \to \min_{x\in X}$, where a priory unknown function $f : X \to \mathbb{R}$ is convex, set $X$ is convex compact in $\mathbb{R}^N$; see e.g. [1, 2] where both problem statements and optimization methods are described. In such problems, in order to sequentially estimate the minimum point $x^* \in \operatorname{Argmin}_{x\in X} f(x)$ it is assumed that, at each time $t = 1, 2, \ldots$, there is an ability to get subgradient $g_t = g_t(x_{t-1}) \in \partial f(x_{t-1})$ or its stochastic version $u_t(x_{t-1}) = g_t + \xi_t$ at current point $x_{t-1} \in X$ where $\partial f(x)$ denotes subdifferential of function $f$ at point $x$, and $\xi_t$ represents a disturbance of the subgradient.[1] The foregoing assumes that the minimized function is known up to its membership in a given class $\mathcal{F}$ of convex functions (probably, under additional smooth properties); in addition, it is assumed that at each current time $t \geq 1$ it is possible to access the oracle at current input point $x_{t-1}$ and get a stochastic subgradient as the output $u_t(x_{t-1})$.

In [3] it is shown that in convex problems with the "correct" choice of the MDM parameters, the latter is an effective method in the sense that

[1] We are talking about the concept of an oracle of the first order in the optimization problem under consideration (either deterministic problem, when $\xi_t \equiv 0$, or stochastic one, under $\mathbb{E}\{\xi_t\} \equiv 0$) [3].

for each $t > 1$ the upper and lower bounds of the error (by the objective function)

$$f(\hat{x}_t) - \min_{x \in X} f(x)$$

coincide up to an absolute constant; here $\hat{x}_t$ represents "final" estimate of the minimum point by the time $t$, based on previous observations of subgradients at the obtained points $x_k$, $k = 0, 1, \ldots, t-1$.

Often, as estimate $\hat{x}_t$, the arithmetic mean of the preceding points is taken, $\hat{x}_t = (x_0 + x_1 + \cdots + x_{t-1})/t$. We note that the fundamentally new in the structure of the MD method in comparison with the classical methods of gradient type is the (explicit or implicit) presence of two spaces, primal space $E \supset X$ with an initial norm $\|\cdot\|$ and conjugate one $E^*$ with dual norm $\|\cdot\|_*$; for details, see [4], section 3.1. In the particular, "Euclidean" case $E = E^*$ when both norms are Euclidean and the set $X = \mathbb{R}^N$ is the whole initial space, the MD method is transformed into subgradient method $x_t = x_{t-1} - \gamma_t u_t(x_{t-1}), \quad t = 1, 2, \ldots$ Recall that the introduction of an additional inertia term into the gradient method can improve the convergence properties of the algorithm. This refers to the heavy-ball method proposed by B.T. Polyak in 1964 [5] (see also [6]). Hence, it is reasonable to generalize the MDM by adding an appropriate inertia term [7]. Sections 3 and 4 are devoted to the realization and study of this idea.

## 2. Stochastic optimization problem

Consider well-known minimization problem

$$f(x) \triangleq \mathbb{E}\, Q(x, Z) \to \min_{x \in X}, \tag{1}$$

where loss function $Q : X \times \mathcal{Z} \to \mathbb{R}_+$ contains random variable $Z$ with unknown distribution on space $\mathcal{Z}$, $\mathbb{E}$ — mathematical expectation, set $X \subset \mathbb{R}^N$ — given convex compact in $N$-dimension space, random function $Q(\cdot, Z) : X \to \mathbb{R}_+$ is convex a.s. on $X$.

Let i.i.d sample $(Z_1, \ldots, Z_{t-1})$ be given where all $Z_i$ have the same distribution on $\mathcal{Z}$ as $Z$. Introduce notation for stochastic subgradients

$$u_k(x) = \nabla_x Q(x, Z_k), \quad k = 1, 2, \ldots, \tag{2}$$

such[2] that $\forall x \in X$, $\mathbb{E}\, u_k(x) \in \partial f(x)$. The goal is in constructing and proving novel recursive MD algorithms meant for minimization (1) and using stochastic subgradients (2) at current points $x = x_{t-1} \in X$, $t \geq 1$.

[2]Below we mean $\nabla_x Q(x, Z_k)$ be the subgradient which are measurable functions defined on $X \times \mathcal{Z}$ such that, for any $x \in X$, the expectation $\mathbb{E} u_k(x)$ belongs to $\partial f(x)$.

## 3. The idea of method of inertial mirror descent

In this section, let $f : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ be convex, continuously differentiable function having a unique minimum point $x^* \in \mathrm{Argmin} f(x)$ and its minimal value $f^* = f(x^*)$. Consider continuous algorithm which extends MDM

$$\dot{\zeta}(t) = -\nabla f(x(t)), \quad t \geq 0,\ \zeta(0) = 0, \tag{3}$$

$$\mu_t \dot{x}(t) + x(t) = \nabla W(\zeta(t)), \quad x(0) = \nabla W(\zeta(0))\,. \tag{4}$$

Functional parameter in (4) is a convex, continuously differentiable function $W : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}_+$ having conjugate function

$$V(x) = \sup_{\zeta \in \mathbb{R}^{\mathbb{N}}} \{\langle \zeta, x\rangle - W(\zeta)\}. \tag{5}$$

Let $W(0) = 0$, $V(0) = 0$, and $\nabla W(0) = 0$ for simplicity.

**Remark.** *Under parameter $\mu_t \equiv 0$ in (4), algorithm (3)–(4) represents MDM (in continuous time) [3]; in particular, the identical mapping $\nabla W(\zeta) \equiv \zeta$ and $\mu_t \equiv 0$ lead to a continuous standard gradient method. Under $\mu_t \equiv \mu > 0$ and $\nabla W(\zeta) \equiv \zeta$, algorithm (3)–(4) leads to continuous method of heavy ball (MHB) [6].* □

Further, we assume that parameter $\mu_t \geq 0$ is differentiable, and method (3)–(4) we call continuous Method of Inertial Mirror Descent (MIDM).

Assume a solution $\{x(t)\}_{t\geq 0}$ to system equations (3)–(4) exists. Consider function $W_*(\zeta) = W(\zeta) - \langle \zeta, x^*\rangle, \zeta \in \mathbb{R}^{\mathbb{N}}$, attempting to find a candidate Lyapunov function. Trajectory derivative to system (3)–(4) be

$$\begin{aligned} dW_*(\zeta(t))/dt &= \langle \dot{\zeta}, \nabla W - x^*\rangle = -\langle \nabla f(x), \mu_t \dot{x} + x - x^*\rangle \leq & (6)\\ &\leq f(x^*) - f(x(t)) - \mu_t d[f(x(t)) - f^*]/dt & (7)\end{aligned}$$

where last inequality results from convexity $f(\cdot)$. Now, integrating on interval $[0, t]$ with $W_*(0) = 0$, we obtain
$\int_0^t [f(x(t)) - f^*]dt \leq -W_*(\zeta(t)) - \mu_t [f(x(t)) - f^*]|_0^t + \int_0^t [f(x(t)) - f^*]\dot{\mu}_t dt$, where two last terms in RHS got by integrating in parts. Taking (5) into account, we continue $\left(1 - \sup_{s\in[0,t]} \dot{\mu}_s\right) \int_0^t [f(x(t)) - f^*]dt \leq V(x^*) - \mu_t [f(x(t)) - f^*]|_0^t$. Therefore, it is reasonable to introduce the following constraints on parameter $\mu_t \geq 0 :\ \mu_0 = 0,\ \dot{\mu}_t \leq 1\ \forall t > 0$, leading to inequality $f(x(t)) - f^* \leq V(x^*)/\mu_t$. Maximizing $\mu_t$ under constraints above we get $\mu_t = t, \quad t \geq 0$. The related (continuous) IMD algorithm

$$\dot{\zeta}(t) = -\nabla f(x(t)), \quad t \geq 0, \quad \zeta(0) = 0, \tag{8}$$

$$t\,\dot{x}(t) + x(t) = \nabla W(\zeta(t)), \tag{9}$$

proves upper bound $f(x(t)) - f^* \leq V(x^*)\, t^{-1}, \quad \forall t > 0\,.$

## 4. Algorithm IMD. Main results

Now return back to optimization problem of section 2. Let $\|\cdot\|$ be a norm in primal space $E=\mathbb{R}^N$, and $\|\cdot\|_*$ be the related norm in dual space $E^*=\mathbb{R}^N$; set $X\subset E$ is convex compact, $\beta>0$ is a scale parameter.

**Assumption (L).** *Convex function $V: X\to\mathbb{R}_+$ is such that its $\beta$-conjugate $W_\beta(\zeta)\triangleq\sup_{x\in X}\left\{-\zeta^T x-\beta V(x)\right\}$ is continuously differential on $E^*$ with Lipschitz condition $\|\nabla W_\beta(\zeta)-\nabla W_\beta(\tilde{\zeta})\|\le(\alpha\beta)^{-1}\|\zeta-\tilde{\zeta}\|_*$, $\forall\ \zeta,\tilde{\zeta}\in E^*$, $\beta>0$, where $\alpha$ is positive constant being independent of $\beta$.*

Consider now the discrete time $t=1,2,\dots$ and introduce sequence $\gamma_t>0$. Write a discrete version of algorithm IMD (8)–(9) using stochastic subgradients (2) instead of the gradients $\nabla f(\cdot)$:

$$\tau_t = \tau_{t-1}+\gamma_t\,,\quad t\ge 1,\ \tau_0=0, \tag{10}$$

$$\zeta_t = \zeta_{t-1}+\gamma_t u_t(x_{t-1}),\quad \zeta_0=0, \tag{11}$$

$$\tau_t\,\frac{x_t-x_{t-1}}{\gamma_{t+1}}+x_t = -\nabla W_{\beta_t}(\zeta_t),\quad x_0=-\nabla W_{\beta_0}(\zeta_0); \tag{12}$$

parameters $\gamma_t>0$ and $\beta_t\ge\beta_{t-1}>0$ are specified in (13), function $W_\beta$ is defined by proxy-function $V:X\to\mathbb{R}_+$ via Legendre–Fenchel type transform [4], $W_\beta(\zeta)=\sup_{x\in X}\left\{-\zeta^T x-\beta V(x)\right\}$, $\zeta\in E^*$.

**Remark.** *Equation (12) may be written as*

$$x_t=\frac{\tau_t}{\tau_t+\gamma_{t+1}}x_{t-1}-\frac{\gamma_{t+1}}{\tau_t+\gamma_{t+1}}\nabla W_{\beta_t}(\zeta_t).$$

*Since the vectors $[-\nabla W_{\beta_t}(\zeta_t)]\in X$ under each $t\ge 0$, equations (10)–(11) show that $x_t\in X$ by induction.* □

Further, let sequences $(\gamma_i)_{i\ge1}$ and $(\beta_i)_{i\ge1}$ are of view

$$\gamma_i\equiv 1\,,\quad \beta_i=\beta_0\sqrt{i+1}\,,\quad i=1,2,\dots,\quad \beta_0>0. \tag{13}$$

Then system equations (10)–(12) leads to the IMD algorithm (c.f. [8]):

$$\zeta_t = \zeta_{t-1}+u_t(x_{t-1}),\quad \zeta_0=0,\quad x_0=-\nabla W_{\beta_0}(\zeta_0), \tag{14}$$

$$x_t = x_{t-1}-(t+1)^{-1}\left(x_{t-1}+\nabla W_{\beta_t}(\zeta_t)\right),\quad t\ge 1. \tag{15}$$

**Theorem 1.** *Let $X$ be convex closed set in $\mathbb{R}^N$, and loss function $Q(\cdot,\cdot)$ satisfies the conditions of section 2, and, moreover, $\sup_{x\in X}\mathbb{E}\|\nabla_x Q(x,Z)\|_*^2\le L_{X,Q}^2$, where constant $L_{X,Q}\in(0,\infty)$. Let $V$ be proxy-function on $X$ with parameter $\alpha>0$ from Assumption (L), and let exists minimum point $x^*\in\operatorname*{Argmin}_{x\in X} f(x)$, perhaps non unique. Then for any $t\ge1$ estimate $x_t$, defined by algorithm (14), (15) with stochastic subgradients (2) and sequence $(\beta_i)_{i\ge1}$ from (13) with arbitrary $\beta_0>0$, satisfies inequality*

$$\mathbb{E}\,f(x_t)-\min_{x\in X}f(x)\le\left(\beta_0 V(x^*)+L_{X,Q}^2/(\alpha\beta_0)\right)\sqrt{t+2}/(t+1)\,.$$

*If constant* $\overline{V}$ *is such that* $\max\limits_{x\in X} V(x) \le \overline{V}$, *and* $\beta_0 = L_{X,Q}\,(\alpha\,\overline{V}\,)^{-1/2}$ *then*

$\mathbb{E}\,f(x_t) - \min_{x\in X} f(x) \le 2\,L_{X,Q}\left(\alpha^{-1}\overline{V}\right)^{1/2}\sqrt{t+2}/(t+1)\,.$ □

## 5. Conclusion

We considered the well-known convex problem of stochastic optimization with the goal of constructing and investigating the novel recursive algorithms of mirror descent type which generalize both heavy ball method and MDM. It turned out that the new method does not require additional averaging of the input points to the oracle and it ensures the same upper bound on the objective function, as the previous, effective method of MD (on the class of considered problems) [3, 4]. It seems interesting the further research for another classes of objective functions and requirements to oracle.


## Acknowledgments

The work is partially supported by the Russian Science Foundation grant No 16–11–10015. The author considers it his duty to thank B.T. Polyak for his attention to this work and A. Juditsky for important discussions and sending the reference [8].


## References


1. *Boyd S., Vandenberghe L.* Convex Optimization. — Cambridge University Press, 2004.
2. *Nesterov Yu.* Introductory Lectures on Convex Optimization. — Boston: Kluwer, 2004.
3. *Nemirovskii A.S., Yudin D.B.* Problem Complexity and Method Efficiency in Optimization. — Chichester: Wiley, 1983.
4. *Juditsky A.B., Nazin A.V., Tsybakov A.B., Vayatis N.* Recursive aggregation of estimators by the mirror descent algorithm with averaging // Problems of Information Transmission. — 2005. — Vol. 41, no. 4. — P. 368–384.
5. *Polyak B.T.* Some methods of speeding up the convergence of iteration methods // USSR Comp. Math. and Math. Phys. — 1964. — Vol. 4, no. 5. — P. 1–17.
6. *Polyak B.T.* Introduction to optimization. — New York: Optimization Software Inc., 1987.
7. *Nazin A.* Algorithms of Inertial Mirror Descent in Convex Problems of Stochastic Optimization // ArXiv: 1705.01073. — 2017.
8. *Nesterov Yu., Shikhman V.* Quasi-monotone Subgradient Methods for Nonsmooth Convex Minimization // J. Optim. Theory Appl. — 2015. — No. 165. — P. 917–940. — DOI: 10.1007/s10957-014-0677-5.