Optimization Letters (2023) 17:1215–1228
https://doi.org/10.1007/s11590-022-01931-4

ORIGINAL PAPER

# Regularized quasi-monotone method for stochastic optimization

V. Kungurtsev[1] · V. Shikhman[2]

Received: 8 July 2021 / Accepted: 6 September 2022 / Published online: 21 September 2022

**Abstract**
We adapt the quasi-monotone method, an algorithm characterized by uniquely having convergence quality guarantees for the last iterate, for composite convex minimization in the stochastic setting. For the proposed numerical scheme we derive the optimal convergence rate of $\mathrm{O}\left(\frac{1}{\sqrt{k+1}}\right)$ in terms of the last iterate, rather than on average as it is standard for subgradient methods. The theoretical guarantee for individual convergence of the regularized quasi-monotone method is confirmed by numerical experiments on $\ell_1$-regularized robust linear regression.

**Keywords** Composite minimization · Quasi-monotone method · Individual convergence · Regularization · Stochastic optimization

## 1 Introduction

In the minimization of nonsmooth convex functions, typically, algorithms generate a sequence of iterates using subgradients or estimates thereof. The convergence rates are then derived for some linear combination of the iterates, rather than for the last estimate computed. Obtaining guarantees on the last iterate per se is often a challenging task. A significant contribution in that direction – sometimes also refered to as individual convergence – was given in [5] with the quasi-monotone subgradient method. The corresponding analysis was simplified and extended to solving minimization problems on decentralized networks in [4]. In this paper we extend the work

✉ V. Kungurtsev
kunguvya@fel.cvut.cz

V. Shikhman
vladimir.shikhman@mathematik.tu-chemnitz.de

[1] Department of Computer Science, Faculty of Electrical Engineering, Czech Technical University in Prague, 121 35 Praha 2, Karlovo náměsí-13, Prague, Czech Republic

[2] Department of Mathematics, Chemnitz University of Technology, Reichenhainer Str. 41, 09126 Chemnitz, Germany

of [5] in two important directions, first we consider a composite minimization problem with a simple additive function (usually a regularizer), and second we consider the stochastic case. We develop the Lyapunov-like analysis from [4] to handle the new elements and present numerical experiments confirming the performance guarantees. We obtain the convergence rate of order $\mathrm{O}\left(\frac{1}{\sqrt{k+1}}\right)$ in expectation of function evaluations, which is optimal for nonsmooth convex optimization.

Let us briefly comment on the related literature. In [6] the authors introduce an adaptation of mirror descent in order to attain the optimal individual convergence. They successively apply the latter for regularized nonsmooth learning problems in the stochastic setting. As shown in [7], the Nesterov's acceleration alternatively provides the individual convergence of projected subgradient methods as applied to nonsmooth convex optimization. Especially, the suggested methodology guarantees the regularization structure while keeping an optimal rate of convergence. Our contribution to individual convergence consists in theoretically justifying that also the initially proposed quasi-monotone subgradient method from [5] can be successively adjusted for composite minimization in the stochastic setting. We note that the setting we consider is distinct from the specialized algorithms that also adapt mirror descent for the important case wherein there are separable linear constraints, e.g., the classical [3] or more recent alternating minimization [2] and proximal point based method [1], but extending to this setting could be an interesting topic to pursue for future work.

## 2 Regularized quasi-monotone method

We consider the composite minimization problem

$$\min_{x} F(x) = \bar{f}(x) + g(x), \tag{1}$$

where $\bar{f}, g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are closed convex functions. Moreover,

$$\bar{f}(x) = \mathbb{E}\left[f(x,\xi)\right]$$

for some $f$ closed and convex in the first argument and $\xi$ is a sample from some random space $\Xi$. We assume that $\mathrm{dom}\,(f(\cdot,\xi)) \subset \mathrm{dom}\,(g)$ for a.e. $\xi$, and $\mathrm{dom}\,(g)$ is closed. Usually, $\bar{f}$ plays the role of a loss function, whereas $g$ is used for regularization. In our setting, $f$ need not to be differentiable, but unbiased finite variance estimates of its subgradients, i.e. $w(x,\xi) \sim \nabla f(x,\cdot)$ with $\mathbb{E}[w(x,\xi)] \in \partial \bar{f}(x)$, should be available. Here, we use $\nabla \bar{f}(x)$ to denote an element of the convex subdifferential $\partial \bar{f}(x) = \partial \mathbb{E}[f(x,\xi)]$, i.e.

$$\bar{f}(y) \geq \bar{f}(x) + \langle \nabla \bar{f}(x), y - x \rangle, \quad y \in \mathrm{dom}\,(g). \tag{2}$$

In addition, $g$ has to be simple. The latter means that we are able to find a closed-form solution for minimizing the sum of $g$ with some simple auxiliary functions. For that, we assume that for the effective domain of $g$ there exists a prox-function

$\Psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ w.r.t. an arbitrary but fixed norm $\|\cdot\|$. The prox-function $\Psi$ has to fulfil:

(i) $\Psi(x) \geq 0$ for all $x \in$ dom $(g)$.

(ii) $\Psi$ is strongly convex on dom $(g)$ with convexity parameter $\beta > 0$, i.e. for all $x, y \in$ dom $(g)$ and $\alpha \in [0, 1]$ it holds:

$$\Psi\big(\alpha x + (1-\alpha)y\big) \leq \alpha\Psi(x) + (1-\alpha)\Psi(y) - \frac{\beta}{2}\alpha(1-\alpha)\|x-y\|^2.$$

(iii) The auxiliary minimization problem

$$\min_x \{\langle s, x\rangle + g(x) + \gamma\Psi(x)\}$$

is easily solvable for $s \in \mathbb{R}^n$ and $\gamma > 0$.

In our analysis, we consider that $g$ is strongly convex with convexity parameter $\sigma \geq 0$ w.r.t. the norm $\|\cdot\|$. Note that $\sigma = 0$ corresponds to the mere convexity of $g$.

For stating our method, we choose a sequence of positive parameters $(a_k)_{k\geq 0}$, which is used to average the subdifferential information of $f$. We set:

$$A_k = \sum_{\ell=0}^{k} a_\ell.$$

Equivalently, it holds:

$$A_{k+1} = A_k + a_{k+1}. \tag{3}$$

Another sequence of positive parameters $(\gamma_k)_{k\geq 0}$ controls the impact of the prox-function $\Psi$. We assume:

$$\gamma_{k+1} \geq \gamma_k, \quad k \geq 0. \tag{4}$$

Now, we are ready to formulate the regularized quasi-monotone method for solving the composite minimization problem (1):

**Regularized Quasi-Monotone Method (RQM)**

```
0. Initialize $x_0 = \arg\min_x \left\{A_0 g(x) + \gamma_0\Psi(x)\right\}$, $s_{-1} = 0$.
1. Sample $\xi_k \sim \Xi$.
2. Compute $w\big(x_k, \xi_k\big)$ and set $s_k = s_{k-1} + a_k w\big(x_k, \xi_k\big)$.
3. Forecast $x_k^+ = \arg\min_x \left\{\langle s_k, x\rangle + A_{k+1}g(x) + \gamma_{k+1}\Psi(x)\right\}$.
4. Update $x_{k+1} = \frac{A_k}{A_{k+1}}x_k + \frac{a_{k+1}}{A_{k+1}}x_k^+$.
```

It is clear that iterates of (RQM) are convex combinations of forecasts:

$$x_k = \frac{1}{A_k}\left( a_0 x_0 + \sum_{\ell=1}^{k} a_\ell x^+_{\ell-1} \right). \tag{5}$$

In order to achieve convergence rates for RQM, the control parameters $\left(a_k\right)_{k\geq 0}$ and $\left(\gamma_k\right)_{k\geq 0}$ should be properly specified. How to do this, will be clear from our convergence analysis in the next section, cf. possible choices in (18), (21) and (22) below.

## 3 Convergence analysis

Before performing the convergence analysis of (RQM), let us deduce some useful properties of the following auxiliary function:

$$\varphi_k(s) = \max_{x\in\mathbb{R}^n} \left\{ \langle s, x\rangle - A_k g(x) - \gamma_k \Psi(x) \right\}, \quad s \in \mathbb{R}^n. \tag{6}$$

Since $A_k g + \gamma_k \Psi$ is strongly convex with convexity parameter

$$\mu_k = A_k \sigma + \gamma_k \beta, \tag{7}$$

the convex function $\varphi_k$ is differentiable and its gradient $\nabla\varphi_k$ is $\frac{1}{\mu_k}$-Lipschitz continuous. The latter property means:

$$\varphi_k(s') \leq \varphi_k(s) + \langle \nabla\varphi_k(s), s' - s\rangle + \frac{1}{2\mu_k}\|s' - s\|^2_*, \quad s, s' \in \mathbb{R}^n. \tag{8}$$

Moreover, it holds:

$$\nabla\varphi_k(-s_{k-1}) = x^+_{k-1}. \tag{9}$$

Let us derive the convergence rate of (RQM). For that, we set:

$$B_k = \frac{1}{2}\sum_{\ell=0}^{k} \frac{a_\ell^2}{\mu_\ell}\mathbb{E}\left\|w(x_\ell, \xi_\ell)\right\|^2_*, \quad k \geq 0, \tag{10}$$

where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$. We shall denote, as standard, the filtration $\sigma$-algebra corresponding to the sequence of iterates as $\{\mathcal{F}_k\}$.

**Theorem 1** *Let* $x_* \in$ dom $(g)$ *solve the composite optimization problem* (1), *and the sequence* $\left(x_k\right)_{k\geq 0}$ *be generated by* (*RQM*). *Then*, *it holds for* $k \geq 0$ *that*:

$$\mathbb{E}\left[F(x_k)\right] - F(x_*) \leq \frac{\gamma_k}{A_k}\Psi(x_*) + \frac{B_k}{A_k}. \tag{11}$$

***Proof*** Let us define the stochastic Lyapunov function:

$$V_k = A_k\big(F(x_k) - F(x_*)\big) + \varphi_k(-s_k) + \langle s_k, x_*\rangle + A_k g(x_*) - B_k.$$

We consider the expected difference:

$$\begin{aligned}\mathbb{E}\left[V_{k+1}|\mathcal{F}_k\right] - V_k \;=\; & \underbrace{A_{k+1}\left(\mathbb{E}\left[F(x_{k+1})|\mathcal{F}_k\right] - F(x_*)\right) - A_k\left(F(x_k) - F(x_*)\right)}_{=\mathrm{I}} \\ & + \underbrace{\mathbb{E}\left[\varphi_{k+1}(-s_{k+1})|\mathcal{F}_k\right] - \varphi_k(-s_k)}_{=\mathrm{II}} \\ & + \underbrace{\mathbb{E}\left[\langle s_{k+1}, x_*\rangle|\mathcal{F}_k\right] + A_{k+1}g(x_*) - \langle s_k, x_*\rangle - A_k g(x_*)}_{=\mathrm{III}} \\ & \underbrace{-\mathbb{E}\left[B_{k+1}|\mathcal{F}_k\right] + B_k}_{=\mathrm{IV}}.\end{aligned}$$

Let us estimate the expressions I-IV from above.

*Estimation of I* We split:

$$\begin{aligned}\mathrm{I} \;=\; & \underbrace{A_{k+1}\left(\mathbb{E}\left[\mathbb{E}\left[f(x_{k+1},\xi)\right]|\mathcal{F}_k\right] - \mathbb{E}\left[f(x_*,\xi)\right]\right) - A_k\left(\mathbb{E}\left[f(x_k,\xi)\right] - \mathbb{E}\left[f(x_*,\xi)\right]\right)}_{=\,\mathrm{I}_f} \\ & + \underbrace{A_{k+1}\left(\mathbb{E}\left[g(x_{k+1})|\mathcal{F}_k\right] - g(x_*)\right) - A_k\left(g(x_k) - g(x_*)\right)}_{=\,\mathrm{I}_g}.\end{aligned}$$

Due to convexity of $f$, the definitions of $A_k$ and $x_k$, we obtain:

$$\begin{aligned}\mathrm{I}_f \;& \overset{(3)}{=} a_{k+1}\left(\mathbb{E}\left[\mathbb{E}\left[f(x_{k+1},\xi)\right]|\mathcal{F}_k\right] - \mathbb{E}\left[f(x_*,\xi)\right]\right) \\ & \quad + A_k\left(\mathbb{E}\left[\mathbb{E}\left[f(x_{k+1},\xi)\right]|\mathcal{F}_k\right] - \mathbb{E}\left[f(x_k,\xi)\right]\right) \\ & \overset{(2)}{\leq} a_{k+1}\mathbb{E}\left[\left\langle \nabla\bar{f}(x_{k+1}), x_{k+1} - x_*\right\rangle|\mathcal{F}_k\right] + A_k\mathbb{E}\left[\left\langle \nabla\bar{f}(x_{k+1}), x_{k+1} - x_k\right\rangle|\mathcal{F}_k\right] \\ & \overset{\mathbf{4.}}{=} \mathbb{E}\left[\left\langle a_{k+1}\nabla\bar{f}(x_{k+1}), x_k^+ - x_*\right\rangle|\mathcal{F}_k\right].\end{aligned}$$

By using convexity of $g$, it also follows:

$$\begin{aligned}\mathrm{I}_g \;& \overset{\mathbf{4.}}{\leq} A_{k+1}\left(\frac{A_k}{A_{k+1}}g(x_k) + \frac{a_{k+1}}{A_{k+1}}\mathbb{E}\left[g(x_k^+)|\mathcal{F}_k\right] - g(x_*)\right) - A_k\left(g(x_k) - g(x_*)\right) \\ & \overset{(3)}{=} a_{k+1}\left(\mathbb{E}\left[g(x_k^+)|\mathcal{F}_k\right] - g(x_*)\right).\end{aligned}$$

Overall, we deduce:

$$\mathrm{I} \;\leq \mathbb{E}\left[\left\langle a_{k+1}\nabla\bar{f}(x_{k+1}), x_k^+ - x_*\right\rangle|\mathcal{F}_k\right] + a_{k+1}\left(\mathbb{E}\left[g(x_k^+)|\mathcal{F}_k\right] - g(x_*)\right).$$

*Estimation of II* First, in view of the definitions of $\varphi_k$, $A_k$, and $x_k^+$, we obtain:

$$\begin{aligned}\mathbb{E}\left[\varphi_k(-s_k)|\mathcal{F}_k\right] &\overset{(6)}{\geq} \mathbb{E}\left[\left\langle -s_k, x_k^+\right\rangle\right] - A_k\mathbb{E}\left[g(x_k^+) - \gamma_k\Psi(x_k^+)|\mathcal{F}_k\right] \\ &\overset{(3)}{=} \mathbb{E}\left[\left\langle -s_k, x_k^+\right\rangle - A_{k+1}g(x_k^+) - \gamma_{k+1}\Psi(x_k^+)\right] \\ &\quad + \mathbb{E}\left[a_{k+1}g(x_k^+) + \left(\gamma_{k+1} - \gamma_k\right)\Psi(x_k^+)|\mathcal{F}_k\right] \\ &\overset{\mathbf{3.}}{=} \mathbb{E}\left[\varphi_{k+1}(-s_k) + a_{k+1}g(x_k^+) + \left(\gamma_{k+1} - \gamma_k\right)\Psi(x_k^+)|\mathcal{F}_k\right].\end{aligned} \tag{12}$$

Second, due to Lipschitz continuity of $\nabla\varphi_k$ and definitions of $s_k$ and $x_k^+$, we have:

$$\begin{aligned}\mathbb{E}\left[\varphi_{k+1}(-s_{k+1})\right] &\overset{(8)}{\leq} \mathbb{E}\left[\varphi_{k+1}(-s_k) + \langle\nabla\varphi_{k+1}(-s_k), -s_{k+1} + s_k\rangle|\mathcal{F}_k\right] \\ &\quad + \frac{1}{2\mu_{k+1}}\mathbb{E}\left[\|-s_{k+1} + s_k\|_*^2|\mathcal{F}_k\right] \\ &\overset{\mathbf{2.},(9)}{=} \mathbb{E}\left[\varphi_{k+1}(-s_k) - \langle x_k^+, a_{k+1}w(x_{k+1},\xi_{k+1})\rangle|\mathcal{F}_k\right] \\ &\quad + \frac{a_{k+1}^2}{2\mu_{k+1}}\mathbb{E}\left[\|w(x_{k+1},\xi_{k+1})\|_*^2|\mathcal{F}_k\right].\end{aligned} \tag{13}$$

By using these two auxiliary inequalities, we are ready to estimate:

$$\begin{aligned}\mathrm{II} &= \mathbb{E}\left[\varphi_{k+1}(-s_{k+1}) - \varphi_k(-s_k)|\mathcal{F}_k\right] \\ &\overset{(12)}{\leq} \mathbb{E}\left[\varphi_{k+1}(-s_{k+1}) - \varphi_{k+1}(-s_k) - a_{k+1}g(x_k^+) - \left(\gamma_{k+1} - \gamma_k\right)\Psi(x_k^+)|\mathcal{F}_k\right] \\ &\overset{(13)}{\leq} -\mathbb{E}\left[\langle a_{k+1}w(x_{k+1},\xi_{k+1}), x_k^+\rangle + \frac{a_{k+1}^2}{2\mu_{k+1}}\|w(x_{k+1},\xi_{k+1})\|_*^2|\mathcal{F}_k\right] \\ &\quad - \mathbb{E}\left[a_{k+1}g(x_k^+) - \left(\gamma_{k+1} - \gamma_k\right)\Psi(x_k^+)|\mathcal{F}_k\right].\end{aligned}$$

*Estimation of III*

The definitions of $s_k$ and $A_k$ provide:

$$\mathrm{III} \overset{\mathbf{2.}}{=} \mathbb{E}\left[\langle a_{k+1}w(x_{k+1},\xi_{k+1}), x_*\rangle|\mathcal{F}_k\right] + a_{k+1}g(x_*).$$

*Estimation of IV*

Here, we have:

$$\mathrm{IV} \overset{(10)}{=} -\frac{a_{k+1}^2}{2\mu_{k+1}}\mathbb{E}\Big[\|w(x_{k+1},\xi_{k+1})\|_*^2|\mathcal{F}_k\Big].$$

Altogether, we can see that

$$\begin{aligned}\mathbb{E}\big[V_{k+1}|\mathcal{F}_k\big] - V_k \;&\leq \mathbb{E}\big[\langle a_{k+1}w(x_{k+1},\xi_{k+1}), x_* - x_k^+\rangle|\mathcal{F}_k\big]\\ &\quad - \mathbb{E}\big[\big\langle a_{k+1}\nabla f(x_{k+1}), x_k^+ - x_*\big\rangle|\mathcal{F}_k\big]\\ &\quad - \big(\gamma_{k+1}-\gamma_k\big)\Psi(x_k^+).\end{aligned}$$

Since $x_k^+$ is defined given $\mathcal{F}_k$, we have:

$$\mathbb{E}\big[\langle a_{k+1}w(x_{k+1},\xi_{k+1}), x_* - x_k^+\rangle|\mathcal{F}_k\big] = \mathbb{E}\big[\langle a_{k+1}\nabla\bar{f}(x_{k+1}), x_* - x_k^+\rangle|\mathcal{F}_k\big].$$

By additionally using that the sequence $(\gamma_k)_{k\geq 0}$ is by assumption nondecreasing, and $\Psi(x)\geq 0$ for all $x\in$ dom $(g)$, we obtain:

$$\mathbb{E}\big[V_{k+1}|\mathcal{F}_k\big] - V_k \leq 0.$$

Hence, we get by induction and taking total expectations:

$$\mathbb{E}[V_k] \leq \mathbb{E}[V_0]. \tag{14}$$

It turns out that the expectation of $V_0$ is nonnegative. For that, we first estimate due to the choice of $x_0$:

$$\begin{aligned}\varphi_0(-s_0) \;&\overset{(8)}{\leq}\; \varphi_0(0) + \langle\nabla\varphi_0(0), -s_0\rangle + \frac{1}{2\mu_0}\|s_0\|_*^2\\ &\overset{\mathbf{0}.}{=}\; -a_0 g(x_0) - \gamma_0\Psi(x_0) - \langle x_0, a_0 w(x_0,\xi_0)\rangle\\ &\quad + \frac{a_0^2}{2\mu_0}\|w(x_0,\xi_0)\|_*^2.\end{aligned} \tag{15}$$

This gives:

$$\begin{aligned}\mathbb{E}[V_0] \quad &= \quad A_0\mathbb{E}\big[F(x_0)-F(x_*)\big] + \mathbb{E}\big[\varphi_0(-s_0) + \langle s_0, x_*\rangle\big]\\ &\qquad + A_0 g(x_*) - B_0\\ &\overset{(2)}{\leq} \quad a_0\big\langle\nabla\bar{f}(x_0), x_0\big\rangle + \mathbb{E}[\varphi_0(-s_0)] + a_0 g(x_0) - B_0\\ &\overset{(15),(10)}{\leq} \quad -\gamma_0\Psi(x_0) \leq 0,\end{aligned} \tag{16}$$

where again the last inequality is due to the assumptions on $\gamma_0$ and $\Psi$. Additionally, it holds by definition of $\varphi_k$:

$$\varphi_k(-s_k) \geq \langle -s_k, x_* \rangle - A_k g(x_*) - \gamma_k \Psi(x_*). \tag{17}$$

Hence, we obtain:

$$A_k \mathbb{E}\left[\left(F(x_k) - F(x_*)\right)\right] \underset{(14)}{=} \mathbb{E}\left[V_k - \varphi_k(-s_k) - \langle s_k, x_* \rangle\right] - A_k g(x_*) + B_k \overset{(16),(17)}{\leq} \gamma_k \Psi(x_*) + B_k.$$

The assertion (11) then follows.

Now, let us show that the convergence rate of (RQM) derived in Theorem 1 is optimal for nonsmooth optimization, i.e. it is of order $\mathrm{O}\left(\frac{1}{\sqrt{k+1}}\right)$. For that, we exemplarily consider the following choice of control parameters:

$$a_k = 1, \quad \gamma_k = \sqrt{k+1}, \quad k \geq 0. \tag{18}$$

We also assume that the subgradients' estimates of *f* have uniformly bounded second moments, i.e. there exists $G > 0$ such that

$$\mathbb{E}\left[\|w(x,\xi)\|_*\right] \leq G, \quad x \in \operatorname{dom}(f). \tag{19}$$

**Corollary 1** *Let* $x_* \in \operatorname{dom}(g)$ *solve the composite optimization problem* (1), *and the sequence* $\left(x_k\right)_{k\geq 0}$ *be generated by* (*RQM*) *with control parameters from* (18). *Then, it holds for all* $k \geq 0$:

$$\mathbb{E}\left[F(x_k)\right] - F(x_*) \leq \left(\Psi(x_*) + \frac{G^2}{\beta}\right)\frac{1}{\sqrt{k+1}}. \tag{20}$$

***Proof*** In order to obtain (20), we estimate the terms in (11) which involve control parameters:

$$\frac{\gamma_k}{A_k} = \frac{\sqrt{k+1}}{k+1} = \frac{1}{\sqrt{k+1}},$$

$$\frac{B_k}{A_k} = \frac{1}{2(k+1)} \sum_{\ell=0}^{k} \frac{1}{\mu_\ell} \mathbb{E}\left[\left\|w(x_\ell, \xi_\ell)\right\|_*^2\right] \overset{(7),(19)}{\leq} \frac{G^2}{2\beta(k+1)} \sum_{\ell=0}^{k} \frac{1}{\sqrt{\ell+1}}$$

$$\leq \frac{G^2}{2\beta(k+1)} \int_{-\frac{1}{2}}^{k+\frac{1}{2}} \frac{\mathrm{d}\,\tau}{\sqrt{\tau+1}} = \frac{G^2}{\beta(k+1)}\left(\sqrt{k+\frac{3}{2}} - \sqrt{\frac{1}{2}}\right)$$

$$\leq \frac{G^2}{\beta\sqrt{k+1}}.$$

□

From the proof of Corollary 1 we see that also other choices of control parameters guarantee the optimal convergence rate of RQM. E.g., we may have chosen:

$$a_k = k, \quad \gamma_k = (k+1)^{\frac{3}{2}}, \quad k \geq 0. \tag{21}$$

We show that the convergence rate of (RQM) derived in Corollary 1 can be improved to $\mathrm{O}\left(\frac{\ln k}{k}\right)$ if the regularizer $g$ turns out to be strongly convex. Additionally, an estimate in terms of generated iterates can be obtained. For that, consider the control parameters as follows:

$$a_k = 1, \quad \gamma_k = \ln(2k+3), \quad k \geq 0. \tag{22}$$

**Corollary 2** *Let $x_* \in$ dom $(g)$ solve the composite optimization problem* (1), *and the sequence $\left(x_k\right)_{k\geq 0}$ be generated by* (*RQM*) *with control parameters from* (22). *Additionally, let $g$ be strongly convex with convexity parameter $\sigma > 0$. Then, it holds for all $k \geq 0$:*

$$\mathbb{E}\left[F(x_k)\right] - F(x_*) \leq \left(\Psi(x_*) + \frac{G^2}{\sigma}\right)\frac{\ln(2k+3)}{k+1}. \tag{23}$$

$$\mathbb{E}\left\|x_k - x_*\right\|^2 \leq \frac{2}{\sigma}\left(\Psi(x_*) + \frac{G^2}{\sigma}\right)\frac{\ln(2k+3)}{k+1}. \tag{24}$$

***Proof*** In order to obtain (23), we estimate the terms in (11) which involve control parameters:

$$\begin{aligned}
\frac{\gamma_k}{A_k} &= \frac{\ln(2k+3)}{k+1},\\
\frac{B_k}{A_k} &= \frac{1}{2(k+1)}\sum_{\ell=0}^{k}\frac{1}{\mu_\ell}\mathbb{E}\left[\left\|w(x_\ell,\xi_\ell)\right\|_*^2\right] \overset{(7),(19)}{\leq} \frac{G^2}{2\sigma(k+1)}\sum_{\ell=0}^{k}\frac{1}{\ell+1}\\
&\leq \frac{G^2}{2\sigma(k+1)}\int_{-\frac{1}{2}}^{k+\frac{1}{2}}\frac{\mathrm{d}\,\tau}{\tau+1} = \frac{G^2}{\sigma(k+1)}\left(\ln\left(k+\frac{3}{2}\right) - \ln\frac{1}{2}\right)\\
&= \frac{G^2}{\sigma}\cdot\frac{\ln(2k+3)}{k+1}.
\end{aligned}$$

For (24), we use the assumption that $g$ is strongly convex, hence, also $F$ is. In particular, we obtain:

$$F(x_k) \geq F(x_*) + \langle s, x_k - x_*\rangle + \frac{\sigma}{2}\left\|x_k - x_*\right\|^2, \quad s \in \partial F(x^*).$$

Since $x_*$ solves (1), we have $0 \in \partial F(x_*)$ and it follows:

$$\left\|x_k - x_*\right\|^2 \leq \frac{2}{\sigma}\left(F(x_k) - F(x_*)\right).$$

By taking expectation and recalling (23), we are done. □

Finally, we note that estimating convergence rates for the generated sequence of iterates itself is a hard task in the framework of subgradient methods. We refer e.g. to [8], where the dual averaging methods were adjusted for stochastic composite minimization. There, just the boundedness of iterates could be in general shown. If $g$ is strongly convex, an estimate was nevertheless provided. This is also the case for RQM as we have shown in Corollary 2.

## 4 Numerical experiments

We performed numerical experiments on two representative synthetic problems with various parameters for the data generation and observed two general patterns in regards to the relative performance of solvers.

For each problem instance we ran one hundred trials of (RQM) in order to investigate the robustness and spread of the performance. Note that the initial $x_0$ set by (RQM) is the zero vector. First, we compare the parameter choice Parameters A as in (18), i.e. $a_k = 1$ and, thus, $A_k = k + 1$, with $\gamma_k = \sqrt{k+1}$, to the choice of $a_k = k$ and, thus, $A_k = \frac{k(k+1)}{2}$, with the more aggressive constant step-size $\gamma_k = 10$ Parameters B.

We also compare (RQM) to the stochastic regularized subgradient (SRSG) with Nesterov's extrapolation from [7]. There, by choosing control parameters

$$\theta_k = \frac{2}{k+1}, \quad \gamma_k = (k+1)^{3/2},$$

the authors iterate:

$$\begin{aligned} y_k &= \hat{x}_k + \theta_k\left(\theta_{k-1}^{-1} - 1\right)\left(\hat{x}_k - \hat{x}_{k-1}\right), \\ \hat{x}_{k+1} &= \arg\min_x \left\{\langle w(y_k, \xi_k), x\rangle + g(x) + \gamma_k \Psi(x - y_k)\right\}. \end{aligned} \tag{25}$$

Finally, we also compare the procedure to the standard mirror descent algorithm, in this case the update becomes,

$$x_{k+1} = \arg\min_x \left\{\langle w(x_k, \xi_k), x\rangle + g(x) + \gamma_k \Psi(x - x_k)\right\}. \tag{26}$$

In the experiments we use the Euclidean dual map $\Psi(x) = \frac{1}{2}\|x\|_2^2$ and the theoretically optimal rate $\gamma_k = \sqrt{k+1}$.

### 4.1 Huber loss and $\ell_1$-regularization

Consider linear regression with a robust Huber loss and $\ell_1$-regularization, i.e.

$$\min_{\mathbf{a},b} \sum_{i=1}^{N} L_\delta\big(\mathbf{a}^T x_i + b - y_i\big) + \lambda \|(\mathbf{a}, b)\|_1,$$

where

$$L_\delta(z) = \begin{cases} \frac{1}{2} z^2 & \text{for } |z| \le \delta, \\ \delta\Big(|z| - \frac{1}{2}\delta\Big) & \text{otherwise.} \end{cases}$$

Here, we expect the number $N$ of data samples to be large. The $\ell_1$-regularization on the parameters encourages sparsity, i.e. most of the parameters to become zero. The Huber loss is a means of mitigating the impact of outliers on the stability of the regression estimate, i.e. by enforcing linear as opposed to quadratic growth of the loss beyond the influence boundary $\delta$. We take the subgradients

$$\partial L_\delta(z) \ni \begin{cases} z & \text{for } |z| \le \delta, \\ \delta \cdot \operatorname{sign}(z) & \text{otherwise.} \end{cases}$$

Denoting $x = (\mathbf{a}, b)$ and choosing as prox-function $\Psi(x) = \frac{1}{2}\|x\|_2^2$, the subproblem in (RQM) admits an explicit solution:

$$\begin{aligned} x^+ &= \arg\min_x \left\{ \langle s, x\rangle + A\lambda \|x\|_1 + \frac{\gamma}{2}\|x\|_2^2 \right\} \\ &= \operatorname{sgn}\left(-\frac{s}{\gamma}\right) \max\left\{ \left|-\frac{s}{\gamma}\right| - \frac{A\lambda}{\gamma}, 0 \right\}. \end{aligned}$$

The explicit solution of the subproblem in (25) is

$$\begin{aligned} \hat{x} &= \arg\min_x \left\{ \langle w, x\rangle + \lambda \|x\|_1 + \frac{\gamma}{2}\|x - y\|_2^2 \right\} \\ &= \operatorname{sgn}\left(y - \frac{w}{\gamma}\right) \max\left\{ \left|y - \frac{w}{\gamma}\right| - \frac{\lambda}{\gamma}, 0 \right\}. \end{aligned}$$

Now we describe the generation of the synthetic data used to generate the problem for comparison. We let $\mathbf{a} \in \mathbb{R}^{10}$, $\delta = 2$, and $N = 10000$ and conduct the following procedure:

1. Choose 4 components of **a** to be nonzero. Randomly sample these components and $b$.
2. Choose 10000 input samples $\{x_i\}$ uniformly in $[-5, 5]$.

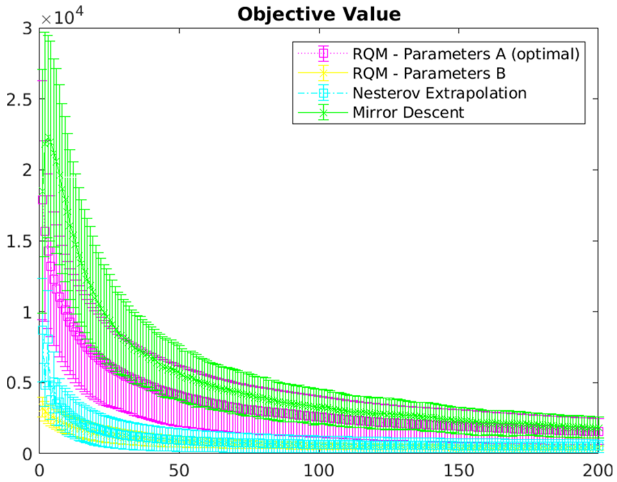

**Fig. 1** Evolution of the objective error with the iterations for Huber Loss and $\ell_1$-Regularization

3. With probability 0.95 generate $y_i \sim \mathcal{N}\left(\mathbf{a}^T x_i + b, 1\right)$, and otherwise $y_i \sim \mathcal{N}\left(\mathbf{a}^T x_i + b, 5\right)$.
4. Run (RQM).

We set $\lambda = 0.1$. The trajectory of the objective value with the associated one standard deviation confidence interval is shown in Fig. 1. Note that for the Nesterov's extrapolation algorithm we report the objective value on $\hat{x}_k$, as this is what the theoretical convergence guarantees in [7] are derived for. We see that SRSG with Nesterov's extrapolation performs better that our method with the theoretically optimal choice of Parameters A. However, it is worth mentioning that alternative parameter choices (Parameter B) may work better in practice. In any case, all the methods produce convergent sequences w.r.t. the function value.

### 4.2 $\ell_1$-regression and box constraint

Now, we present a problem and data generation parameters for which the theoretically optimal algorithm RQM exhibit the better comparative performance. The optimization problem is defined to be $\ell_1$-regression with an indicator of membership to a box constraint

$$\min_{\mathbf{a},b} \sum_{i=1}^{N} \left|\mathbf{a}^T x_i + b - y_i\right| + \mathbf{1}_C(\mathbf{a}, b),$$

where

$$C = \{x : x_l \leq x \leq x_u\}.$$

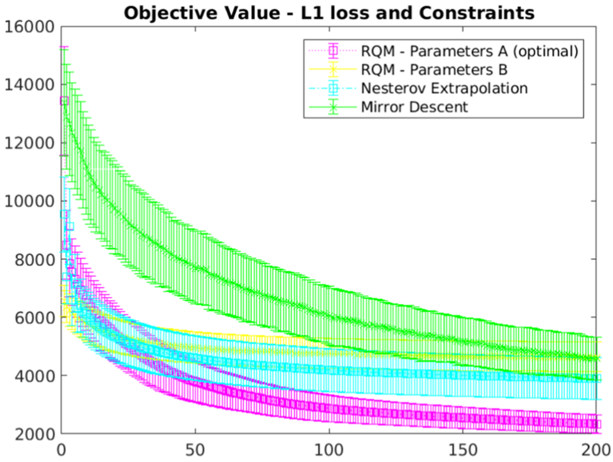

**Fig. 2** Evolution of the objective error with the iterations for $\ell_1$-Regression and Box Constraint

In this case, the subproblem uses a stochastic subgradient of the loss term, i.e., for sample $i$, we have that $s$ or $w$, respectively for the two algorithms is given by

$$\operatorname{sign}\left(\mathbf{a}^T x_i + b - y_i\right) \begin{pmatrix} x_i \\ 1 \end{pmatrix}.$$

Additionally, we use the same prox-function $\Psi(x) = \frac{1}{2}\|x\|_2^2$. Finally, incorporating the indicator in the subproblem is simply projection onto the box $C$. We let $\mathbf{a} \in \mathbb{R}^5$ and $N = 1000$ and generate the data as follows:

1. Randomly sample $\mathbf{a} \sim \mathcal{N}(0, 7\mathbf{I})$ and $b \sim \mathcal{N}(0, 8)$.
2. Choose 1000 input samples $\{x_i\}$ uniformly in $[-15, 15]$.
3. Generate $y_i \sim \mathcal{N}\left(\mathbf{a}^T x_i + b, 1\right)$.
4. Finally $\mathbf{x}_l = x_l \mathbf{1}$ was chosen randomly uniformly in $[-20, 0]$ and $\mathbf{x}_u = x_u \mathbf{1}$ in [0, 20].

We can see now in Fig. 2 that in this case the theoretically optimal choice of Parameters A exhibit the best performance.

**Acknowledgements** The authors would like to thank the anonymous referees for their suggestions in improving the presentation and content of this work. Research Supported by the OP VVV project CZ. 02.1.01/0.0/0.0/16_019/0000765 “Research Center for Informatics”.

## References

1. Bai, J., Zhang, H., Li, J.: A parameterized proximal point algorithm for separable convex optimization. Optim. Lett. **12**(7), 1589–1608 (2018)

2. Bitterlich, S., Boţ, R.I., Csetnek, E.R., Wanka, G.: The proximal alternating minimization algorithm for two-block separable convex optimization problems with linear constraints. J. Optim. Theory Appl. **182**(1), 110–132 (2019)
3. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
4. Liang, S., Wang, L., Yin, G.: Distributed quasi-monotone subgradient algorithm for nonsmooth convex optimization over directed graphs. Automatica **101**, 175–181 (2019)
5. Nesterov, Y., Shikhman, V.: Quasi-monotone subgradient methods for nonsmooth convex minimization. J. Optim. Theory Appl. **165**(3), 917–940 (2015)
6. Tao, W., Pan, Z., Wu, G., Tao, Q.: Primal averaging: a new gradient evaluation step to attain the optimal individual convergence. IEEE Trans. Cybern. **50**, 835–845 (2020)
7. Tao, W., Pan, Z., Wu, G., Tao, Q.: Strength of Nesterov's extrapolation in the individual convergence of nonsmooth optimization. IEEE Trans. Neural Netw. Learn. Syst. **31**, 1–12 (2020)
8. Xiao, L.: Dual averaging methods for regularized stochastic learning and online optimization. J. Mach. Learn. Res. **11**(88), 2543–2596 (2010)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.